# Методи аналізу тексту

## Methods of Text Analysis

Чернівці
Чернівецький національний університет
2009

**МЕТОДИ АНАЛІЗУ ТЕКСТУ:**
**Збірник наукових праць. – Чернівці: ЧНУ, 2009. – 350 с.**
**METHODS OF TEXT ANALYSIS:**
**Omnibus volume. – Chernivtsi: ČNU, 2009. – 350 p.**

У збірнику представлено результати досліджень, здійснених у країнах Західної та Східної Європи, США, а також у Китаї. Статті відображають різноманітність сучасних методів аналізу тексту і його окремих рівнів. Особлива увага приділяється квантитативним методам дослідження тексту.

В сборнике представлены результаты исследований, выполненных в странах Западной и Восточной Европы, США, а также в Китае. Статьи отображают разнообразие современных методов анализа текста и его отдельных уровней. Особое внимание уделяется квантитативным методам исследования текста.

The omnibus volume provides results of various approaches to text analysis and related issues in Western and Eastern Europe, USA and China. Special attention is paid to quantitative methods in text analysis.

**Наукові редактори:**
Еммеріх Келіх (Грац, Австрія)
Віктор Левицький (Чернівці, Україна)
Габріель Альтманн (Люденшайд, Німеччина)

**Editors:**
Emmerich Kelih (Graz, Austria)
Viktor Levickij (Černivci, Ukraine)
Gabriel Altmann (Lüdenscheid, Germany)

Друкується за ухвалою Редакційно-видавничої Ради
Чернівецького національного університету імені Юрія Федьковича.

Gedruckt mit der Unterstützung der Karl-Franzens-Universität Graz.

ISBN 978-966-423-043-5

# Художественное слово и поэтическая картина мира

*Николай Алефиренко (Россия, Белгород)*

Категория «картина мира» (КМ), впервые используемое в философских трактатах Л. Витгенштейном и перенесенное в лингвосемиотику Лео Вайсгербером, порождает в науке множество разных субкатегорий. В их ряду появилось понятие «поэтическая картина мира» (ПКМ), которое становится предметом спора и в лингвокультурологии, и в лингвистике (см.: Чумак-Жунь 2009). К сожалению, не обретя строго терминологического значения, словосочетание *поэтическая картина мира* нередко употребляется в качестве научной метафоры с размытым и завуалированным содержанием. Для получения статуса истинно терминологического значения в его содержании необходимо отразить, по крайней мере, три момента: а) когнитивную составляющую, б) интерпретационный характер и в) семиотическую природу.

*Когнитивная* сущность ПКМ заключается в том, что она является авторским преломлением коллективного отражения мира в этнокультурном сознании того или иного языкового сообщества. Такое отражение действительности представляет собой результат двуединого процесса – рационального и чувственного познания (Schwarz 1992), что, собственно, и определяет творческий, преобразующий и интерпретирующий характер ПКМ. Преобладание второго аспекта познания обусловливает семиотическую природу средств репрезентации ПКМ. Этим, собственно, ПКМ отличается от языковой (ЯКМ), в содержании которой рациональное и чувственное сосуществуют на паритетных началах. ЯКМ определяется как языковые образы реальных предметов и отношений, периферийные участки вербальных представлений, которые становятся источником дополнительных сведений об окружающей нас действительности. Причем они часто производят стойкие отложения в сознании познающего субъекта в силу образного характера их информации. ПКМ – индивидуально-авторская *интерпретация* смыслового содержания ЯКМ, направленная не столько на репрезентацию реальной действительности, сколько на моделирование возможных миров.

*Интерпретационный потенциал* понятия ПКМ обусловливается тем, что в отличие от логико-предметного содержания ЯКМ, экспрессивно-образное и эмотивно-оценочное содержание ПКМ генетически связано с образными *представлениями*. Первая вербализуется прямономинативной, производной и профессиональной лексикой и терминами, а вторая – преимущественно знаками вторичной и косвенно-производной номинации (метафорами, фразеологизмами, паремиями). Первые представляют собой элементы объективно сложившегося коллективного сознания, вторые – элементы художественного (поэтического) сознания, отфильтрованные в идиоэтническом десигнате соответствующего поэтического знака. Особую культурологическую значимость имеют те поэтические знаки, в основе которых лежат когнитивные категории, совмещающие в себе универсальные и идиоэтнические обобщения действительности, реальные и ментальные (возможные) миры. Знания об идиоэтнических, по своей сути ментальных, мирах образуют ПКМ – своеобразную сферу существования культуры, формой существования которой служат концепты, которые формируются в результате своеобразного членения ПКМ на некие микромиры (Скаб 2008: 51), соответствующие всем возможным ситуациям, известным человеку и поэтому называемым возможными мирами.

Исследования последних лет дают основание разграничивать в ПКМ макроконцепты (константы национальной культуры) и субконцепты (идиоэтнические варианты последних). Сегодня особенно продуктивной разработке подвергается проблема константных концептов. Одной из причин такого положения дел служит количественный фактор (константы культуры могут быть представлены ограниченным списком типа «мир», «вечность», «сущность», «время», «огонь», «вода», «язык» и др.»). К тому же их представляют слова весьма сложной семантической структуры, которые, несмотря на свою рациональную универсальность, обладают, как показала в своих работах А. Вежбицкая, имплицитными национально-культурными различиями. Это служит основанием для выделения универсальной, языковой (ЯКМ) и поэтической картины мира. Наиболее дискуссионной является проблема ЯКМ и ПКМ, поскольку она связана с критикуемой многими исследователями гипотезой Сепира-Уорфа. С такой критикой следовало бы, разумеется, согласиться, если бы теория ЯКМ и ПКМ действительно исходила из постулатов пресловутой гипотезы. На самом же деле

в ее основу кладутся постулаты об *инвариантности* логических и *вариативности* языковых категорий, участвующих в формировании и ЯКМ, и ПКМ. Остовом любой ЯКМ является, конечно же, универсальная (логическая, инвариантная) модель действительности, структурированная при помощи таких невербальных средств, как универсальный предметный код или «язык» когнитивных примитивов. Этнические языки лишь переводят инвариантный код в соответствующие ему этнокультурные коды. Язык в подобных научных доктринах не принимает активного участия в познавательных процессах и поэтому связан с культурой только как средство ее репрезентации. Согласиться с таким пониманием взаимоотношения языка культуры – значит, оставить без внимания интереснейшие наработки отечественных и зарубежных ученых, раскрывающие имплицитные механизмы взаимодействия языковой и культурной семиотики (М. Коул). Достаточно вспомнить гипотезу «культурных знаков» Л.С. Выготского, согласно которой между человеком как субъектом познания и окружающим его миром существует влиятельнейший посредник и интерпретатор язык культуры, важнейшим из которых является естественный язык. Поскольку же язык вообще существует в конкретных этносемиотических системах, то язык каждого народа преломляет отражаемый в сознании мир в соответствии с его семиотическим устройством, грамматической структурой и накопленной в семантике языковых единиц социокультурной информацией (Левицкий 2006).

Подобная лингвокультурологическая доктрина при всей ее привлекательности и непротиворечивости способна, однако, вызвать иллюзорное представление о господстве языка над культурой, возвратив нас тем самым в прокрустово ложе гипотезы лингвистической относительности Сепира-Уорфа. Язык в рамках этой доктрины выступает неким таинственным демиургом культурной реальности, ее творцом, созидающим началом, креативной силой. И все же (в который раз!) и сегодня приходится задумываться над тем, все ли уж так иллюзорно в этой гипотезе. Известны на этот счет два типа диаметрально противоположных суждения (оба представлены в работах последнего десятилетия XX века). Одно из них принадлежит С. Пинкеру, безоговорочно отрицающему какой-либо разумный смысл нашумевшей гипотезы (Пинкер 2004). Такие столь безапелляционые оценки вызывают вполне аргументированные возражения А. Вежбицкой, которая, обращая внимание на явные преувеличения роли родного языка в восприятии и понимании

мира, все же принимает основной тезис Уорфа о том, что мы расчленяем природу в направлении, подсказанном нашим родным языком, что мы расчленяем мир, [как это] закреплено в системе моделей нашего языка. На наш взгляд, споры о том, насколько язык определяет стиль и образ мышления, подпитываются досадным исключением из обсуждаемой проблемы такой ее важнейшей когнитивной составляющей, как *этнокультурное сознание* и способы его семиотизации. Иными словами, различного рода иллюзии об абсолютном господстве одного из базовых элементов речемыслительной деятельности – *языка* или *мышления* – порождаются неразличением когнитивной значимости языковой и культурологической семиотики в познании и отражении мира. Экспериментальные исследования показали, что различия в мышлении обусловлены различиями не между языками, а между распространенными в той или иной культуре видами деятельности. Поэтому, если вновь апеллировать к гипотезе Сепира-Уорфа, следует говорить не о лингвистической, а о «деятельностной относительности», поскольку не может быть и речи о том, будто разным языкам соответствуют разные типы познавательных процессов.

В соответствии с *семиотической* концепцией сознания А.Н. Леонтьева, в структуре сознания принято выделять три образующие его компонента: чувственную ткань образа, значение и личностный смысл. Это особенно важно учитывать при описании концептосферы языка писателя. В соответствии с когнитивным подходом к языку энциклопедические знания или личностные знания о мире дают представление о том, что понимается под моделью мира в сознании писателя и читателя, и каково соотношение концептуальной и языковой картин мира в их воображении. ПКМ как субъективный образ объективного мира зарождается и существует в глубинных слоях психики писателя, как правило, скрытых от самонаблюдения. К таким слоям относятся области подсознания и сверхсознания. Каждая индивидуально-авторская картина мира всегда представляет национально-культурное видение действительности, смысловое конструирование мира в соответствии с художественной «логикой» построения поэтического текста, отражающего ПКМ автора. Процесс же восприятия и понимания ПКМ читателем является с этой точки зрения результатом соотнесения и наложения ЯКМ автора и ЯКМ читателя.

ПКМ, разумеется, не может быть полностью адекватной ЯКМ читателя. Вместе с тем по мере осмысления текста достигается

сближение субъективных ЯКМ автора и читателя, что делает возможной *понимание* поэтического текста. При этом ПКМ должна содержать новую информацию об объектах, представленных в ЯКМ. Причём ПКМ автора художественного текста должна быть шире и богаче ЯКМ читателя. Только в таком случае поэтическое произведение приобретает этнокультурную значимость. Это, в свою очередь, предполагает осмысление характера информации, представляющей в семантике поэтического знака (языкового знака вторичной и косвенно-производной номинации) соответствующие элементы универсальной картины мира и ПКМ. Иногда утверждают, что первая продуцирует логическую семантику, а ПКМ – языковую. Поскольку обе картины мира так или иначе интерпретируются языковым сознанием человека (Алефиренко 2007: 379-395), генератором и носителем как универсальной, так и идиоэтнической информации является, на мой взгляд, всё же языковая семантика.

*Языковое сознание и семантика языка.* Сознание языковой личности, представляя субъективный образ мира, реализуется в семантике поэтического знака. По Л.С. Выготскому, «сознание отражает себя в слове как солнце в малой капле воды. Слово относится к сознанию, как малый мир к большому, как живая клетка к организму, как атом к космосу. Оно и есть малый мир сознания. Осмысленное слово есть микрокосм человеческого сознания». Более того, «мысль не воплощается, а совершается в слове» (Выготский 1982: 361). Поэтому становится понятным, почему рассмотрение проблемы сознания языковой личности невозможно без анализа одной из его главных составляющих – слова. По А.А. Леонтьеву, значение существует для субъекта в двойственном виде: с одной стороны, это объект сознания, с другой – способ и механизм осознания... Они (значения) входят в систему общественного сознания, являются социальными явлениями (и в этом качестве, прежде всего, изучаются лингвистикой); но одновременно они входят в систему личности и деятельности конкретных субъектов, являются частью индивидуального сознания (и в этом качестве изучаются психологией)» (Леонтьев 1983: 8-9). Значению в поэтической речи противостоит личностный смысл как мотивированное отношение к обозначаемому. Такое противопоставление, а точнее говоря, разграничение понятий «значение» и «смысл» было впервые введено Л.С. Выготским и сегодня является основой когнитивной поэтики (А.Н. Леонтьев, А.Р. Лурия, А.А. Юм и др.).

Если под значением слова принято понимать объективно сложившаяся система связей, одинаковая для всех носителей языка, то под смыслом – индивидуальное значение поэтического слова, выхваченное из этой устоявшейся системы связей. Вместо них в поэтическом тексте оно состоит из тех связей, которые имеют отношение к данной ситуации. Поэтому поэтический смысл – результат привнесения в слово коннотаций, соответствующих конкретному ощущению, восприятию и пониманию обозначаемого предмета. Одно и то же поэтическое слово имеет два значение: (а) сформировавшееся в этноязыковом сознании исторически, и (б) которое *потенциально* сохраняется (возможно, в разном объеме и ракурсе у поэта и читателя), отражая с различной полнотой и глубинно «возможные миры». Наряду со значением каждое поэтическое слово приобретает смысл, актуализирующий в этом значении те стороны, которые связаны с данной ситуацией и аффективным отношением к ней поэта: *Больничные молитвенные дни / И где-то близко за стеною — море / Серебряное — страшное, как смерть* (А. Ахматова. 1 декабря 1961. Больница). Однако понятие смысла поэтического слова, как нам представляется, не может быть сведено к различию потенциального (денотативного) и актуального (коннотативного) значений. Смысл поэтического слова возникает в процессе речемыслительной деятельности поэта и читателя в конкретный отрезок времени и в конкретной дискурсивной ситуации: различные типы контекстов и дискурсивная ситуация – условия обнаружения нужного смысла поэтического слова. Обычное слово в ассоциативно-семантической сети (см.: поэтического дискурса обогащается особыми экспрессивно-смысловыми свойствами. Превращение беспристрастного знака в поэтический и сотворение в процессе такой метаморфозы индивидуального смысла художественного слова осуществляется благодаря его чувственному ***переживанию.*** Чувственная ткань образа, по теории Ф.Е. Василюка, – это многомерная субстанция. Чтобы понять и описать её создается *модель образа сознания,* согласно которой (1) внешний мир являет предметное содержание, (2) внутренний мир – личностный смысл, (3) культура – значение, (4) а язык – слово. Вместе все эти составляющие (синергетические «узлы») задают объем, в котором пульсирует и переливается живой образ сознания языковой личности. Образ сознания языковой личности многомерен. Существует пять синергетических измерений, четыре из которых (значение, предмет, личностный смысл,

знак) являются своего рода магнитными полюсами образа сознания языковой личности. «В каждый момент силовые линии внутренней динамики образа могут направляться по преимуществу к одному из этих полюсов, и возникающим при этом доминированием одного из динамических измерений создается особый тип образа» (Василюк 1984: 18). Пятое измерение – чувственная ткань, особая внутренняя «составляющую» образа сознания языковой личности поэта.

Высказанные суждения позволяют определить категориальные свойства понятия «языковое сознание поэта». Прежде всего, его нельзя ни сводить к совокупности речевых умений поэта и его знания языка, ни к отрицанию их взаимосвязи. «Языковое сознание поэта» скорее сближается с пониманием «образа мира». Поэтому языковое сознание поэта является сложным феноменом. Во-первых, это вербальное средство формирования, хранения и переработки информации, получаемой поэтом извне. Во-вторых, это структура, кодирующая полученную информацию языковыми знаками косвенно-производной номинации вместе с выражаемыми ими переживаниями, субъективными значениями (смыслами), правилами их сочетания и употребления. Все это выражает отношение автора поэтического текста к действительности, своеобразие его мировосприятия и эстетические установки на речевое творчество. Языковое сознание поэта ни онтологически, ни функционально не может быть замкнутой структурой. Оно связано с языковыми сознаниями читателей. Если рассуждать психосемантическими категориями, то языковое сознание функционирует благодаря общей нейронной сети. С этой точки зрения, языковое сознание – явление кооперативное. Однако поэт как личность творческая обладает еще и уникальным сверхсознанием – творческой интуицией, или вдохновением, благодаря которому сигнал извне может вызывать взрывоподобный эффект цепных реакций и соединять вход нейронной сети буквально со всей информацией, уже хранящейся в мозге. Причем сверхсознание не контролируется сознанием. Сознание лишь осуществляет окончательный отбор и категоризацию вновь полученной информации, которая может использоваться им на уровне подсознания – набора программ поведения, усвоенных в процессе культурной социализации. Здесь они окончательно усваиваются, автоматизируются и становятся навыками. При наличии связи между нейронными сетями, находящимися в критическом состоянии, появляется возможность передавать информацию из

любой части такой нейронной сети. Таким механизмом является, прежде всего, естественный язык.

Итак, когнитивно-синергетическая энергия поэтического текста исходит от двуединого лингвокриативного процесса – его порождения и восприятия. Данное утверждение основывается на том, что, во-первых, в их основе лежит единый универсальный механизм текстовой деятельности, во-вторых, своеобразие поэтического мышления, изначально определяясь особым восприятием объекта действительности, затем испытывает потребность в его ассоциативно-образном выражении, типичным средством которого, как известно, служит поэтический текст, Без образного восприятия мира и моделирования поэтической картины мира невозможно порождение поэтического текста – живой формы существования языковой личности поэта. При этом, как нам представляется, соотношение в семантической структуре языкового знака универсального и идиоэтнического обусловливается природой той когнитивной категории (представления, концепта, гештальта, фрейма), которая лежит в основе семантики поэтического знака. Именно она определяет характер его культурной коннотации.

Большинство исследователей коннотации рассматривают в качестве возможных ее компонентов эмотивность, экспрессию, оценку, образность и стилистические характеристики языковых единиц. Полным набором перечисленных компонентов характеризуется коннотация в работах И.В. Арнольд и И.А. Стернина. При этом коннотативные признаки противопоставляются денотативным как вторичные, сопутствующие, дополнительные и поэтому факультативные. Эту точку зрения разделяют авторы многих исследователей (О.С. Ахманова, Н.А. Лукьянова, И.В. Арнольд, Э.В. Кузнецова и др.). Ей противостоит концепция В.А. Булдакова и В.И. Шаховского, согласно которой коннотация – равноправный макрокомпонент языкового значения.

Вторая проблема связана с тем, какие из обсуждаемых признаков являются собственно коннотативными, а какие сопутствующими – культурологическими. Наиболее спорным оказался стилистический признак языкового знака. В.А. Булдаков его называет в иерархии коннотативных признаков доминантными, а Н.А. Лукьянова выносит за рамки коннотации. Подобное отношение наблюдается и к образности: В.К. Харченко считает ее компонентом коннотации, а О.В. Загоровская рассматривает ее третьим макрокомпонентом значения после денотативного и коннотативного,

т.е. элементом коннотации не признает. В.И. Шаховским за пределы коннотативной структуры выводятся экспрессия и оценка, поскольку относит их к денотативному макрокомпоненту значения, а смыслообразующим элементом коннотации рассматривается эмотивность как коммуникативно-прагматическая категория. Столь противоречивое понимание коннотации обусловлено зыбкими критериями ее отграничения от денотации.

В терминологическом значении слово «коннотация» начало широко использоваться в лексикографической практике в двух основных смыслах: а) для обозначения «добавочных» (модальных, оценочных и эмотивно-экспрессивных) элементов лексических значений, фиксируемых в словарных статьях; б) для выражения оценочного отношения к предметам знакообозначения, которое элементом лексического значения не является. Однако это различие чаще всего не соблюдалось, что порождало терминологическую путаницу:

- *коннотация* интенсионал, смысловой конструкт, противопоставляемый *денотации* (логико-философская традиция, истоки которой следует искать в работах Дж.С. Милля);
- *коннотация* – синтаксическая валентность слова (психолингвистическая традиция, сформированная К. Бюлером);
- *коннотация* – переносное значение фигурального происхождения (А.В. Исаченко);
- *коннотация* – факультативный элемент лексического значения (У. Майер-Барановска).

Приведенные определения помогают уяснить сущность *культурной коннотации* как особой разновидности традиционно выделяемого макрокомпонента семантики поэтического знака. Ср.:

*Месяц* ***рогом*** *облако* ***бодает****,*
***В*** *голубой* ***купается пыли****.*
*В эту ночь никто не отгадает,*
*Отчего кричали журавли.* С. Есенин.

Переносное значение фигурального происхождения и факультативный элемент лексических значений именной метафоры *рог месяца* и глагола *бодает* в сопряжении с образной зарисовкой ночи (метафорическая семантика выражения *Месяц купается в голубой пыли)* в ритмико-мелодической тональности русской народной песни рождают народнопоэтические коннотации сквозь призму специфически есенинского мировосприятия.

Рассматриваемая категория, тем не менее, имеет и свои отличительные свойства. С одной стороны, понимание культурной коннотации сближается с этимологическим толкованием коннотации как со-значения слова. А с другой, – она все более явно приобретает собственно культурологическую значимость, становясь базовым понятием ПКМ. В ее содержание входят и когнитивные, и дискурсивные, и культурные смыслы.

Культурные смыслы отражают цель, значение и ценность слова-события. В связи с этим он креативен, способен саморазвиваться и быть средством моделирования возможных миров. Он контекстуален, и в этом смысле процессуален, но и атемпорален одновременно, поскольку может быть транслирован. В этом плане он интертекстуален и интердискурсивен. При таком подходе под культурной коннотацией следует понимать дискурсивно-когнитивная интерпретанту поэтического знака (знака этнокультурного сознания), связанную с его образно мотивированным значением. Именно из совокупности мотивированных значений поэтических знаков и моделируется поэтическая (индивидуально-авторская) картина мира.

*ЛИТЕРАТУРА*

Алефиренко Н.Ф. (2007): Когнитивные основания лингвосемиозиса. In: Deutschmann, P. (unter Mitarbeit von Grzybek, P.; Karničar, L.; Pfandl, H.): *Kritik und Phrase. Festschrift für Wolfgang Eismann*. Wien: Praesens, 379-395

Василюк Ф.Е. (1984): *Психология переживания*. – Москва.

Выготский Л.С. (1982): *Мышление и речь* // Собр. соч.: В 6 т. – Москва, Т. 2.

Левицкий В.В. (2006): *Семасиология*. – Винница: Нова книга.

Леонтьев А.А. (1983): Формы существования значения // *Психолингвистические проблемы семантики*. – Москва.

Пинкер С. (2004): *Язык как инстинкт* / Пер. с англ. Е.В. Кайдаловой. Общ. ред. В.Д. Мазо. – Москва: Едиториал УРСС.

Скаб М.В. (2008): *Закономірності концептуалізації та мовної категорізації сакральної сфери*: Монографія. – Чернівці: «Рута».

Чумак-Жунь И.И. (2009): *Поэтический текст в русском лирическом дискурсе конца XVIII – начала XXI веков*. – Белгород: Изд-во Белгородского государственного ун-та.

Schwarz M. (1992): *Einführung in die Kognitive Linguistik*. – Tübingen.

# Weight Factor Formalisms in the Study of Lexical Growth: The Case of Textually Modelled Strings of English Verbs

*Michael Bilynsky, Andriy Pereymybida (Lviv, Ukraine), Gabriel Altmann (Lüdenscheid, Germany)*

## *1. INTRODUCTORY REMARKS*

The placement of words within synonymous strings on the principle of a gradual loss of proximity of each subsequent constituent to the head-word (string's dominant) makes up the onomasiological dictionary. The latter is also known as a thesaurus. Such an arrangement of words is in a way reflexive of the mental lexicon of speakers.

A thesaurus is a metric object due to the discreteness of the lexicon that consists of a number of related words. The issue of relatedness of words may become more meaningful should we establish the varying degrees of such a relatedness existent in the lexical system at the present time and/or diachronically.

A unit of a thesaurus is a *string* (also referred to as a *series* or *set*) of synonymous words placed in a specific sequence. The constituents of such a series are characterized by their varying currency in the lexicon prior to the attestation of each subsequent series member. This can be determined from the dates of their earliest attested usage in the written medium which in the corpus format of diachronic analysis are known as the *historical textual prototypes*. Thus, the synonymous string of words also looks like a relevant unit for diachronic onomasiology.

Owing to a rich onomasiological tradition English lexicography has a number of dictionaries of synonymous strings with an arbitrary, *i.e. non-alphabetical*, placement of constituents within the lexical set. This principle of placing synonymous constituents within the respective series is taken for the intrinsic feature of the thesaurus. The available thesauri open up a possibility of multiple diachronic reconstruction of the onomasiological databases. For the present our research interest is confined to just one such dictionary (WNWT).

Relevant and unmatched for our interest in the entire diachronic expansion of the English lexicon are the earliest quotations (textual prototypes) from the *Oxford English Dictionary (OED)*, the 3rd version of the Second CD-ROM edition of which (OED) was fully processed for this study. Understandably, the absolute chronology of synonyms can also be presented as a relative sequence of words.

We will separate the head-word from the rest of the string of synonyms with the mark ⊂. Upon historical reconstruction of the string its arbitrary constituent (the present-day dominant inclusive) proves to be the one with the oldest textual prototype. It will be separated from the rest of the string with the mark *⊂. Conversely, the present-day string head-word finds itself in an arbitrary position within the historically reconstructed string. In our notation it will be followed by the mark [⊂]. Finally, the present-day string head-word may happen to be its historically earliest constituent. This will be double marked as *⊂ [⊂].

A somewhat similar view on the problem of diachronic reconstruction as ours, though taking into account the entry into the semantic spaces from Roget's thesaurus of individual *OED* attested meanings (senses) of words on the bases of their dated textual prototypes related to the appropriate OED registered (sub-)meanings and/or senses, can be found in the *Historical Thesaurus of English* (Kay 2002).

As we are interested in the lexical rather than epidigmatic growth and collective lexical memory over time we seem justified in introducing a simplification. We will regard the word's entry into the lexicon as a single event. This is in line with the practice of compiling chronological dictionaries.

Nonetheless in the current study verbs bestowed with polysemy as long as it is recognized in the chosen thesaurus are attributed respective multiple strings as opposed to verbs initiating single strings. In the former case respective meanings disambiguation always follows the dominant: e.g. ***lade*** *[To fill] ⊂ replenish, stuff, pack;* ***lade*** [*To dip] ⊂ scoop, bail, spoon.* Polysemic dominants just as those with single strings, understandably, concatenate an arbitrary number of synonyms: e.g. ***lay*** *[To knock down] ⊂ trounce, defeat, club;* ***lay*** *[To place] ⊂ put, deposit, set;* ***lay*** *[To put in order] ⊂ arrange, organize, systematize;* ***lay*** *[To bring forth] ⊂ generate, deposit, yield;* ***lay*** *[To smooth out] ⊂ steam;* ***lay*** *[To bet] ⊂ game, wager;* ***lay*** *[To work out] ⊂ devise, concoct, design.*

The length of verbal synonymous strings in the chosen thesaurus ranges from just two to ninety-nine constituents (as in the case of the head verb ***hinder*** ⊂ … ) although very long strings are quite uncommon.

## *2. THE CHRONOLOGICAL VARIANTS OF STRINGS OF SYNONYMOUS VERBS IN ENGLISH*

The bulk of the reconstructed *historical thesaurus* of verbs contains over six thousand dated textual extracts (earliest quotations) from the illustrative material in the *OED*. They repeat themselves in varying succession in total over thirty thousand times within strings of different lengths.

This extensive *textual* corpus allows for a multitude of sequentially relevant placement of dated fragments of chronological homo- (approximately same period) and/or heterogeneity (partial or/and complete long-term diachrony) containing synonymously related words. The accepted chronological homogeneity of the string is arbitrary. It may lie, for instance, in the historical period(s) or shared century/-ies or/and generation(s) affiliation of its constituting prototypes.

The conducted diachronic reconstruction of lexical sets falls back on the constituents textual prototypes attested in the *OED*. Unless adduced in full the reference to the latter may just follow the string constituent as the *OED* dating given in brackets. The present-day string of synonyms in a thesaurus originates from the chronologically sequential set modelled on textual prototypes of its constituents over time. This relationship of diachronic derivation will be marked by the sign < following the present-day string and preceding its historical rearrangement.

The reconstruction of sets of lexemes over time based on the documented textual prototypes is the main research tool of diachronic corpus onomasiology. In the accepted *OED* dating approximations of the type *circa* and *about* were ignored. In the case of a period dating of the prototype, e.g. *implore (1500-20)*, the earlier date was accepted. The Old English part of textual prototypes was incorporated into the corpus according to the *OED* dating. In this paper we proceed from the accepted chronological layering of the history of English into Old, Middle, Early New and post-Early New English periods admitting fully attested, although in part sporadic or even unique, calculus of twelve types of *textually modelled* historical strings of English verbs. The principle of the calculus lies in the fact that textual prototypes of any single, two, or three period affiliations are not attested whereas in the maximum chronological string the texts for all of the affiliations are registered:

**OE, ME, ENE and post-ENE textual prototypes,** e.g. cram (1000) [To stuff] *⊂ [⊂] crush (1398), compact (1530), jam (1706):

***c*1000** ÆLFRIC *Gram.* (Z.) 190 *Farcio*, ic crammiȝe oððe fylle.

**1398** TREVISA *Barth. De P.R.* x. vii. (1495) 379 Cole quenchyd though it greue not wyth brennynge ym that trede theron it makyth crusshynge and grete noyse.

**1530** PALSGR. 490/2, I compacte a thing shorte togyther to make it stronge, *je trousse*.

**1719** DE FOE *Crusoe* i. xiii, The Ship..stuck fast, jaum'd in between two Rocks.

It looks as though stratification of the thesaurus along the line of the age factor of constituents of synonymous strings is capable of opening up a new area of research in diachronic onomasiology. The distribution of the *OED* textual prototypes within the string may fall on any single period in the accepted sequence of synchronic layers, for instance on just Middle English or Early New English.

**ME textual prototypes,** e.g. conjure (1290) [To appeal to] *⊂ [⊂] entreat (1340), adjure (1382), implore (1500):

***c*1290** *S. Eng. Leg.* I. 172/2291 And is Abbod cam to him bifore is ende-dai And coniurede him þat he scholde after is deþe þere to him comen.

***c*1340** *Cursor M.* 24795 (Fairf.) To entrete of þe pais betwix him & þa danais.

**1382** WYCLIF *1 Kings* xviii. 10 He hath adjurid (Vulg. *adjuravit*) alle rewmes and folkis, for thi that thou art not foundun.

**1500–20** DUNBAR *Poems* lxxxv. 55 Implore, adore, thow indeflore, To mak our oddis evyne.

**ENE textual prototypes**, e.g. annihilate (1525) *⊂ [⊂] exterminate (1541), demolish (1570), obliterate (1600):

**1525** LD. BERNERS *Froiss.* II. cliii. 421 That shulde breke or adnychilate..the alyances that hath been sworne.

**1541** ELYOT *Image Gov.* (1549) 146 Oppression, xtorcion..were out of the citee of Rome..vtterly exterminate.

**1570-6** LAMBARDE *Peramb. Kent* (1826) 285 The Chapell of Hakington..was quite and cleane demolished.

**1600** W. WATSON *Decacordon* (1602) 224 To obliterate, eradicate, and vtterly extinguish the name of Bishops.

On the other hand, such a distribution may be characterized by a mixed period affiliation of texts where within a single string of synonyms constituents' prototypes fall on (non)adjacent periods, for example Early New English and post-Early New English or on Middle English and post-Early New English:

**ENE and post-ENE textual prototypes**, e.g. browbeat (1603) *⊂ [⊂] intimidate (1646), frighten (1666), bully (1710):

**1603** HOLLAND *Plutarch's Mor.* 129 We must entertaine our friends and guests, with courtesie..and not to brow-beat them.
**1646** H. LAWRENCE *Comm. Angells* 121 Nothing intimidates more than ignorance.
**1666** PEPYS *Diary* 4 Sept., Which at first did frighten people more than any thing.
**1710** PALMER *Proverbs* 69 His poor neighbour is bully'd by his big appearance.

**ME and post-ENE textual prototypes,** e.g. flatter (1225) [To praise unduely] *⊂ [⊂] glorify (1340), overpraise (1387), adulate (1777):

***a*1225** *Ancr. R.* 222 (MS. Cleop. C. vi) Men..þet flattereð hire of freolac.
***a*1340** HAMPOLE *Psalter* xiv. 5 Þaim þat dredis god he glorifys. Þat is he haldis þaim gloriouse and worthi to rest in godis hill.
**1387** TREVISA *Higden* (Rolls) V. 339 It may wel be þat Arthur is ofte overpreysed.
**1777** DALRYMPLE *Trav. Spain & Port.* xxxix, The way to preferment here is by..adulating some superior, who probably is a despicable character.

Understandably, the combining of flank period affiliation within one string in distant diachrony or single flank period affiliation of textual prototypes of the string's constituents are uncommon:

**OE textual prototypes**, e.g. spare (825) *⊂ [⊂] forbear (888), forgive (900):

***c*825** *Vesp. Psalter* lxxi. 13 Goḍ..spearað dearfan & weðla.

***c*888** K. ÆLFRED *Boeth.* xxxvi. §1 Hwa mæʒ forbæran þæt he þæt ne siofiʒe.

***c*900** tr. *Bæda's Hist.* i. xvi. [xxvii.] (1890) 84 Forþon ne bið þæt forʒifen þætte alefed bið, ac þæt bið riht .

**post-ENE textual prototypes**, e.g. over-develop (1869) *⊂ [⊂] over-extend (1937):

**1869** *Eng. Mech.* 19 Nov. 238/3 He would be likely to over-develope it.

**1937** R. ERSKINE *Stout Adventure M. Stewart* iii. 62 A culture and a civilization,..which, reckoned in the gross, outweighed and over-extended by a deal feudal, that is to say, English culture, manners and customs.

In longer strings the distribution of textual prototypes over time typically exceeds single period affiliation. Distant diachrony of textual prototypes of constituents in short strings is attested, but it is far less common than in longer strings.

The strings that contain words with the textual constituents spanning across different periods can be split into respective sub-strings. These are compatible with the strings whose constituents reveal textual prototypes of just one period in language history. The period splitting may result in the situation when no sub-strings but just individual period affiliated lexemes come out of this procedure.

The threshold measured by a period in language history is but a conventional requirement for chronological homogeneity of textual prototypes of related lexemes. Present-day possibilities of electronic modelling allow for multiple experimentation with the textual prototypes age differential values yielding variant data stratification of the evolution of lexicon over time

The adduced examples are characteristic of a part of the corpus (1573 out of 4758 strings) where the present-day and historical dominants of the string coincide. Moreover, ordinal places of other constituents in the string may remain intact over time. There is, however, a limitation of the strings' length here. Numerically, the sequential intactness of synonyms within a string is but a marginal feature (337 as opposed to 4107 sets) at the string length of up to four constituents: ***riot*** *⊂ revolt < **riot** (1375) *⊂ [⊂] revolt (1548);* ***lack*** *⊂ want, require < **lack** (1175)*⊂ [⊂] want (1200), require (1375);* ***savour*** *⊂ enjoy, relish,*

*appreciate* < ***savour*** *(1300) *⊂ [⊂] enjoy (1380), relish (1586), appreciate (1655).* The exceeding of string length with zero positional permutation over time is uniquely represented in the five-constituent series ***slay*** *⊂ murder, slaughter, butcher, assassinate* < ***slay*** *(725) *⊂ [⊂] murder (1225), slaughter (1535), butcher (1562), assassinate (1618).* This sole example stands out of the remainder of 1570 strings with more than four constituents prone to diachronic reshuffling.

When the present-day dominant is chronologically replaced its position is taken over by the oldest constituent. Yet this very lexeme may well be the dominant of (an)other string(s) with no other constituents predating its own textual prototype. The previous situation adds up to this one producing clusters of strings in an historical thesaurus started by the earliest constituent of the string. Only in some strings within such a cluster it coincides with the present-day dominant.

## *3. COMPUTATIONAL FRAMEWORKS FOR COMPARING PRESENT-DAY AND HISTORICAL SEQUENCES OF VERBS*

Proceeding from the availability of the *OED* dated textual prototype of each word the place of a lexeme in the synonymous sequence is historically fixed. For instance, we have a sequence Word: 1 2 3 4 5 6 7 … *n*. In the present-day language this order may be changed, and there are *n!* possibilities for *permutations*. Let us assume that one of them is indeed, e.g. 1 2 6 4 5 7 3. What is the *extent of change* and *how can we measure* the extent of constituents' permutation within present-day and historical strings?

The above question calls for a numerical solution. A feasible one lies in writing the ranks in two rows and computing the absolute value of differences between them:

1 2 3 4 5 6 7 historical ranking
1 3 2 5 6 7 4 present-day ranking
---
0 1 1 1 1 1 3 difference (absolute value)

yielding R = 0 + 1 + 1 + 1+1 + 1 +3 = 8 (1)

as in the present-day string *(1)* ***approach*** *[To approach personally] ⊂ (2) address, (3) propose, (4) request, (5) accost, (6) button-hole, (7) corner*

which is produced the historical series that is sequentially different from its present-day ordering, respectively:

*(1)* ***approach*** *(1305) *⊂[⊂] (2) propose (1340), (3) address (1374), (4) corner (1387), (5) request (1533), (6) accost (1578), (7) button-hole (1828) vs.*

*(1)* ***approach*** *[To approach personally] ⊂(3) propose, (2) address, (7) corner (4) request, (5) accost, (6) button-hole.*

Certainly, the same value of *R* is attainable from a somewhat different correlation of ordinal positions. Understandably, the present-day sequence can be taken for the base of calculus as well, for instance

1 2 3 4 5 6 7 present-day ranking
1 2 6 5 3 4 7 historical ranking
---
0 0 3 1 2 2 0 difference (absolute value): R = 8

as in (1) **slide** [To move with a sliding motion] ⊂(2) glide, (3) skate, (4) skim, (5)slip, (6) coast, (7) toboggan < (1) **slide** (950) *⊂[⊂] (2) glide (1000), (6) skate (1696), (5) skim (1420), (3) slip (1300), (4) coast (1340), (7) toboggan (1846).

In order to get a relative measure we will divide *R* by the maximum it can attain. Consider the maximum difference which arises when the ranks are in the reverse order

1 2 3 4 5 6 7 historical ranking
7 6 5 4 3 2 1 present-day ranking
---
6 4 2 0 2 4 6 Hence here $R_{Max} = 24$ (2)

In general, we get the maximum differences by adding the individual maximum differences which are given as

$$D = 2(n - (2k + 1)),\ k = 0,1,\ldots [n/2 - 1], \quad (3)$$

where [C] is the floor function (the greatest integer, not greater than C).

Taking the sum of D-s we get

$$R_{Max} = \sum_{k=0}^{[n/2 - 1]} 2(n - (2k+1)) = 2\left[\frac{n^2}{4}\right] \qquad (4)$$

Hence the relative difference of permutations is given as

$R/R_{Max}$. (5)

The value of $R_{Max}$ is easy to compute, we have for

| | | |
|---|---|---|
| n | 2 3 4 5 6 7 8 9 10 … | (6) |
| $R_{max}$ | 2 4 8 12 18 24 32 40 50… | |

Some of the string's permutations may be unattested in the historical thesaurus as the number of strings of the respective length is insufficient numerically. Strings with five to nine constituents reach three-digit numbers but they are fewer, and sometimes considerably, than two hundred cases each.

Strings with the length of two, three and four constituents claim almost two thirds of the entire corpus of slightly over six thousand. The present-day dominant may retain or lose its historically precedent position in the string, respectively *e.g.* ***recoup*** *(1430) *⊂ [⊂] regain (1548)* [204 strings] vs. ***bedew*** *(1340) *⊂ dampen [⊂] (1630)* [207 strings]; ***pitch*** *(1205) )*⊂ [⊂] incline (1300), slant (1521)* [395 strings] *vs.* ***shun*** *(950) *⊂* ***eschew*** *[⊂] (1340) abstain (1380)* [754 strings]*;* ***poach*** *[To steal] (1528) )*⊂ [⊂] pilfer (1548), filch (1561), smuggle (1687*) [719 strings] vs. ***seethe*** *(1000) *⊂ scowl* (1340), frown (1386), miff [To frown] [⊂] (1797) [2295 strings]. However, the historical reshuffle of the dominant falls on an arbitrary consecutive position in the string.

In the strings containing from five to nine constituents the present-day dominant retains its chronological precedence in 134 out of 510 strings. In the remainder of the strings containing over nine synonyms this quota drops almost twofold and equals 107 out of 990 strings.

The value of $R_{max}$ can also be computed by means of the recurrence formula

$$a(n) = a(n-1) + a(n-2) – a(n-3) + 2 \qquad (7)$$

It is identical with the maximum sum of absolute values of differences of neighbours in a cyclic permutation of 1…n. If one computes (5) for *all* synonymous sequences in the lexicon, it would be a valuable historical result, the relative measure $R/R_{Max}$. being probably beta-distributed.

Synonymy is partial meaning coverage and hence it can be made measurable. One possibility is to take a word in all its senses and collect its synonyms. Then take the first synonym (second, third…) and compare its synonyms with that of the first. We get three classes: A – synonyms common to both; B – synonyms present with the first but not with the second; C – synonyms present with the second but not with the first. A similarity measure can be set up, e.g.

$$A/(A+B+C) \qquad (8)$$

which is a proportion and can be treated as such. At last the "commonality" of a word with its synonyms can be computed as a function of expression (4), e.g. weighted mean. The above formula is identical with

$$\frac{W \cap S_i}{W \cup S_i} \qquad (9)$$

where *i* is the given synonym of the word W. One can take also the Dice index or Tversky theory. In any case it is measurable but very lengthy.

Formula (1) takes into account relative chronology of constituents in the historically rearranged sequences. Certainly, the same position in relative chronology may be variedly distanced from the neighbouring ones in terms of absolute textual prototypes dating. For instance the fourth and the fifth constituents in the sequences ***break*** *(851) [To fall apart] *⊂ [⊂] burst (1000), shiver (1200), shatter (1330), dilapidate (1570), splinter (1582), collapse (1732), disintegrate (1796)* and ***dream*** *(1250) [To entertain] *⊂ [⊂] conceive (1300), imagine (1340), create (1386), picture (1489), fancy (1545), sublimate (1566), idealize (1786)* differ by one position, but the age difference between their restective textual prototypes is 240 and only 56 years. Hence the relative calculus of the permutation effect does not take into account the diachronic density of textual prototypes within the string.

The logic of building up an historical sequence out of the present-day synonymous string constituents holds when the datings of the respective textual prototypes differ at least by one year. This condition,

however, is not met in 1097 strings. Within them two or even more textual prototypes sometimes including that of the historical dominant (then placed in the string alphabetically) are dated in the same year: ***exchange*** *(1300) *⊂ interchange (1374), relieve (1374), substitute (1532), alternate [⊂] (1595);* ***shimmer*** *(1100) *⊂ sparkle (1200), blink [⊂] (1300), glimmer (1399), glitter (1399);* ***handle*** *(1000) *⊂ settle (1000), receive (1300), manage (1561), collect *⊂ (1573);* ***ferry*** *(1000) *⊂ pull (1000), tow [⊂] (1000), tug (1225), lug (1375), drag (1440), haul (1557), yank (1822);* ***lean*** *(950) *⊂ dip (975), shift (1000), turn (1000), sway (1399), tilt(cause to fall) [⊂] (1399), tip(overthrow) (1399), slant (1521), slope (1591), rake(ships:have a rake.), (1627), slouch (1754).*

The likelihood of the dating overlap in textual prototypes tends to increase with the growth of the strings' length. In the corpus of strings with seven to fourteen constituents (the total of 725 strings) zero difference in the textual prototypes dating occurs in 321 sequences. In the strings exceeding fourteen members (the total of 574 strings) this peculiarity is characteristic of 478 sequences. The record here seems to be held by the 71-constituents long series of the verb *destroy [To bring to nothing]* in the position of the present-day dominant. The sequential placement of twenty-four pairs within this string is impossible owing to their textual prototypes dating overlap.

It may well be that identical dating as well as very narrow chronological difference between textual prototypes of synonyms is related to some peculiarities of coverage and period affiliation of sources in the textual corpus of the *OED*.

The ordinal placement of chronologically coincident constituents within historically rearranged strings is impossible without the counter-intuitive attribution of the present-day sequence to the identically dated prototypes. Hence it is expedient to look for an alternative model of computing constituent's permutation in the present-day and historically rearranged strings.

It seems promising for this purpose to extend over the sequential characteristics of the constituents age the application of *Levickij's weight factor* ($w_i$) (Levickij, Sternin 1989) describing the weight of an arbitrary constituent of the string as dependent upon its ordinal number ($i$) and string's length ($n$)

$$w_i = \frac{n - i + 1}{n} \qquad (10)$$

The value of the difference in the constituents' weight factors in the present-day string equals that inherent of the concluding constituent of the sequence. Each constituent adds up to the tightness of the string.

We suggest visualising the synonymous string as a vector $\{i\}_{i=1}^{n}$ whose length is the aggregate value of the weight factor of all the constituents $\{w_i\}_{i=1}^{n}$.

In the historical thesaurus we have the historical sequence of the dated first *OED* citations of the constituents within the string $\{j_i\}_{i=1}^{n}$ or their ordinal positions within the historical string $\{y_i\}_{i=1}^{n}$.

The weight formalism for the relative chronological scale is

$$\left\{\overline{w}_j = \frac{n - j_i + 1}{n}\right\}_{i=1}^{n} \qquad (11)$$

whereas that for the absolute one is

$$w_i^{(y)} = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}}\left(\frac{1}{n+1} - 1\right) + 1 \qquad (12)$$

The diachronic relevance of facts starts at certain arbitrary values and a difference in the dating of just several or even a couple decades is artificial from the view-point of chronological heterogeneity. In both cases identical ordering will imply zero difference. Formalism (12) gives more precise values in an uneven distribution.

In both (11) and (12) the oldest lexeme $v_{j_n}$ will have the historical weight $w_{j_n}^{(y)}$ =1. Those words that have this characteristic close to the oldest word will have the weight value nearing 1.Younger words will have this value closer to 0.

Then both present-day and historical versions of the strings can be represented in terms of lengths of the respective vectors. The difference between the two vectors is suggested to be taken as *a measure of permutation* of the string's constituents over time.

$$\left\| \overrightarrow{w} - \overrightarrow{w^{(y)}} \right\|^2 = \sum_{i=1}^{n} \left( w_i - w_i^{(y)} \right)^2 \qquad (13)$$

Under the conditions of identical ordinal placement of the constituents within historical and contemporary strings the contemporary weight factor values and the relative historical weight factor values will coincide. To meet this condition for the absolute historical weight factor value the chronological distances between the constituents should be placed on the chronological scale evenly. This, however, should seldom hold true.

Identical ordering within the historical string basing on the actual dates but not that on the ordinal chronological positions of constituents will not imply the zero distance between the respective vectors.

## *4. DISCUSSION*

### *4.1. Applying the permutation factor formalisms and some conventions*

The simplest permutation situation occurs when the string contains just two synonyms. When the sequence is unchanged the sought differential of vectors amounts to zero. If it is changed, its value is 0.71 (Table 1).

Table 1

The difference in vectors' lengths in present-day and historical relative (upper part of the table) and absolute sequences of synonymous pairs (here and further on the number before the verb originates from the internal taggings in our corpus)

| | | | | |
|---|---|---|---|---|
| 3. ACCOMPANY , to supplement, [2] Different: 0,0000 | | | | |
| ACCOMPANY | 1 | 1,00 ( 1460 ) | 1 | 1,00 |
| COMPLETE | 2 | 0,50 ( 1530 ) | 2 | 0,50 |
| 4. ADJOIN , to be close to, [2] Different: 0,7100 | | | | |
| ADJOIN | 1 | 1,00 ( 1325 ) | 2 | 0,50 |
| ABUT | 2 | 0,50 ( 1230 ) | 1 | 1,00 |

However, there could be a situation when two consecutive members of the present-day string are dated identically as regards their textual prototypes. In such a case the relative weight factor values are approximated between w =1.00 and w = 0.5 and the respective permutation factor value gets halved (Table 2).

Table 2
The difference in vectors lengths at the coincident dating of textual prototypes

| | | | | | |
|---|---|---|---|---|---|
| 7. LOCK , lock, [2] Different: 0,3500 | | | | | |
| LOCK | 1 | 1,00 ( 1300 ) | 1 | 0,75 | |
| BAR | 2 | 0,50 ( 1300 ) | 1 | 0,75 | |

In three-member strings, identical sequencing at the relative placement of constituents over time with their present-day placement offers no differentiation. This is not the case should we take into account absolute datings of the textual prototypes as the lapses of time between the appearance of lexemes are not the same (Table 3)

Table 3
The difference in vectors lengths in present-day three-member strings and their historical sequences at relative and absolute chronology of textual prototypes

| | | | | | |
|---|---|---|---|---|---|
| 3. ABIDE , to remain, [3] Different: 0,0000 | | | | | |
| ABIDE | 1 | 1,00 ( 1000 ) | 1 | 1,00 | |
| CONTINUE | 2 | 0,67 ( 1340 ) | 2 | 0,67 | |
| PERSEVERE | 3 | 0,33 ( 1374 ) | 3 | 0,33 | |
| 11. ACHE , ache, [3] Different: 0,0000 | | | | | |
| ACHE | 1 | 1,00 ( 1000 ) | 1 | 1,00 | |
| PAIN | 2 | 0,67 ( 1300 ) | 2 | 0,67 | |
| THROB | 3 | 0,33 ( 1362 ) | 3 | 0,33 | |
| 3. ABIDE , to remain, [3] Different: 0,2800 | | | | | |
| ABIDE | 1 | 1,00 ( 1000 ) | 1 | 1,00 | |
| CONTINUE | 2 | 0,67 ( 1340 ) | 2 | 0,39 | |
| PERSEVERE | 3 | 0,33 ( 1374 ) | 3 | 0,33 | |
| 11. ACHE , ache, [3] Different: 0,2200 | | | | | |
| ACHE | 1 | 1,00 ( 1000 ) | 1 | 1,00 | |
| PAIN | 2 | 0,67 ( 1300 ) | 2 | 0,45 | |
| THROB | 3 | 0,33 ( 1362 ) | 3 | 0,33 | |

The number of actual values of difference in the vectors lengths at relative chronology is smaller than at absolute chronology. The distribution of corpus segments falling at these values in the case of absolute chronology, correspondingly, is expected to be smoother.

At the relative chronological placement of constituents in three-member strings the distribution of the corpus falls on five different permutations. However, these produce only three values of difference in the respective vectors length.

The permutation value for the constituents reshuffle of the present day three-member string is small when its head-word retains this position over time (1-3-2) or when it exchanges its placement with the second consecutive string constituent (2-1-3) (Table 4).

Table 4

The smallest permutation factor values in three-member strings

| | | | | | |
|---|---|---|---|---|---|
| 230. PERVADE , pervade, [3] Different: 0,4700 | | | | | |
| PERVADE | 1 | 1,00 ( 1653 ) | 2 | 0,67 | |
| SUFFUSE | 2 | 0,67 ( 1590 ) | 1 | 1,00 | |
| PERMEATE | 3 | 0,33 ( 1656 ) | 3 | 0,33 | |
| 231. PICK , to choose, [3] Different: 0,4800 | | | | | |
| PICK | 1 | 1,00 ( 1300 ) | 1 | 1,00 | |
| SELECT | 2 | 0,67 ( 1567 ) | 3 | 0,33 | |
| SEPARATE | 3 | 0,33 ( 1432 ) | 2 | 0,67 | |

When the contemporary head-word falls on the penultimate position in the string with the final present-day constituent fitting in the place of the historical dominant (3-1-2) or when it occupies the ultimate sequential position diachronically (2-3-1), with no complete reverse ordering though, the measure of sequential distance between the present-day and historical strings is farther than in the previous case and amounts to 0.8200 (Table 5).

Table 5

Medium permutation factor values at the constituents relative chronological placement

| | | | | | |
|---|---|---|---|---|---|
| 15. AWARD , award, [3] Different: 0,8200 | | | | | |
| AWARD | 1 | 1,00 ( 1386 ) | 3 | 0,33 | |
| GRANT | 2 | 0,67 ( 1225 ) | 1 | 1,00 | |
| BESTOW | 3 | 0,33 ( 1315 ) | 2 | 0,67 | |
| 35. CHAIN , to bind, [3] Different: 0,8200 | | | | | |
| CHAIN | 1 | 1,00 ( 1377 ) | 2 | 0,67 | |
| SHACKLE | 2 | 0,67 ( 1440 ) | 3 | 0,33 | |
| FETTER | 3 | 0,33 ( 1300 ) | 1 | 1,00 | |

When the historical constituents sequence is opposite to their present-day positioning (3-2-1) the permutation factor attains its maximal value (Table 6).

Identical dating of two textual prototypes in a three-member string occurs when the oldest or the second as regards its age textual prototype is dated by the same year as that of one more counterpart.

Table 6
Permutation factor calculus at the reverse historical ordering of constituents in three-member strings

```
1. ABDICATE , abdicate, [3] Different: 0,9500
ABDICATE                    1  1,00 ( 1541 )    3  0,33
RELINQUISH                  2  0,67 ( 1472 )    2  0,67
WITHDRAW                    3  0,33 ( 1225 )    1  1,00
```

When the datings of the last two constituents in a three-member string coincide by the year they are attributed the value of the second counterpart in the consecutively placed constituents of a three-member string. This situation though may repeat itself variedly in the permutation process not only as (1-2-2) but also as (2-1-2) or even (1-1-2). In the latter case the string was historically initiated by two lexemes with identically dated textual prototypes. Thus it has no definite lexeme pre-emptive of all other string members eligible to be imputed the weight factor value of 1.00. Nor does it have the second counterpart with the value 0.67. Instead they are equally attributed the averaged values between 1.00 and 0.67, *i.e.* 0.83 (Table 7).

Table 7
Computing conventions with the dating overlap in the textual prototypes of three-member strings

```
260. PUNCH , to hit, [3] Different: 0,8500
PUNCH                       1  1,00 ( 1382 )    3  0,33
STRIKE                      2  0,67 ( 1000 )    1  0,83
KNOCK                       3  0,33 ( 1000 )    1  0,83

20. UPHOLD , to maintain, [3] Different: 0,2400
UPHOLD                      1  1,00 ( 1225 )    1  1,00
CONFIRM                     2  0,67 ( 1290 )    2  0,50
SUSTAIN                     3  0,33 ( 1290 )    2  0,50

10. HARROW , to torment, [3] Different: 0,6200
HARROW                      1  1,00 ( 1300 )    2  0,50
TORMENT                     2  0,67 ( 1290 )    1  1,00
TRY                         3  0,33 ( 1300 )    2  0,50
```

These two solutions will be extended over identical dating of textual prototypes in the penultimate and previous consecutive positions, respectively, within longer strings at the relative chronological placement of constituents.

Three-member strings with two identically dated textual prototypes are nonproductive (see points 2 and 3 on the upper curve on figure 1).

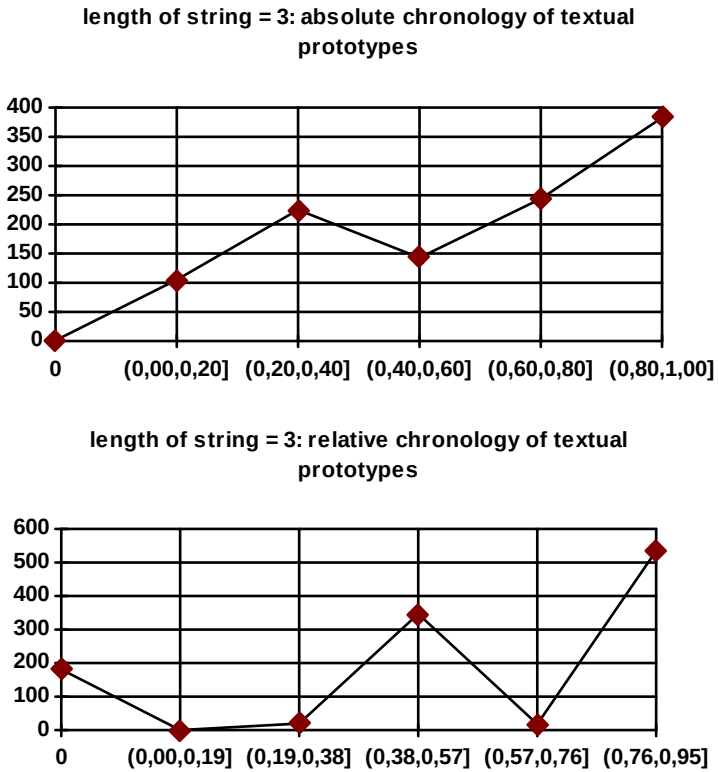



Figure 1. Distribution of permutation factor values within three-member strings: axis x – computed values as a differential in the lengths of respective vectors: axis y – the number of strings (to be repeated in subsequent graphs)

When the weight factor values of synonyms on a historical scale are computed proceeding from their absolute chronology (the exact date of the *OED* supported textual prototype) their distribution within the scale is non-linear as the respective lapses of time between the adjacent positions are arbitrary. This computational parameter provides for a more precise distancing of the respective weight vectors at the string's diachronic permutation. The values that are identical at the constituents' relative placement within the diachronic string become different and more precise at their absolute chronological placement (cf. the curves on figure 1).

The distribution of the permutation factor values is determined by constituents reshuffling as well as their age density. That is why at

the same length of the string there are more meaningful points on the curve of permutation factor values at absolute chronology as compared with the relative one.

### *4.2 Permutation factor formalisms applied to four-member strings: a case study*

It seems plausible that in the numerically most powerful group of four synonyms in a string (see section 3 above) the lexicon provides us with an intuitively clear opportunity to distinguish between close, medium and more distant proximity of a synonymous verb to the string's dominant.

The diachronic rearrangement of four-member strings admit of twenty-four permutations. The ordinal constituent filling in the position of the diachronic string's dominant determines the number of recurrent values and eventually the extent of similarity. When the present-day string's dominant retains its placement at the latter's diachronic rearrangement there are four gauges of the distance including that of zero between the weight factor grounded contemporary and historical vectors of the string. Two of these values repeat themselves (Table 8).

Table 8

Calculus of the permutation factor values in four-member strings with the overlap of the present-day and historical dominants

| | | | | | |
|---|---|---|---|---|---|
| 1393. HIT , to astound, [4] Different: 0,6100 | | | | | |
| HIT | 1 | 1,00 ( 1075 ) | | 1 | 1,00 |
| ASTONISH | 2 | 0,75 ( 1530 ) | | 3 | 0,50 |
| BEWILDER | 3 | 0,50 ( 1684 ) | | 4 | 0,25 |
| AMAZE | 4 | 0,25 ( 1230 ) | | 2 | 0,75 |
| 2283. REVILE , revile, [4] Different: 0,6100 | | | | | |
| REVILE | 1 | 1,00 ( 1303 ) | | 1 | 1,00 |
| BELITTLE | 2 | 0,75 ( 1782 ) | | 4 | 0,25 |
| MALIGN | 3 | 0,50 ( 1426 ) | | 2 | 0,75 |
| REPROACH | 4 | 0,25 ( 1489 ) | | 3 | 0,50 |
| 28. ACCORD , to grant, [4] Different: 0,0000 | | | | | |
| ACCORD | 1 | 1,00 ( 1123 ) | | 1 | 1,00 |
| ALLOW | 2 | 0,75 ( 1300 ) | | 2 | 0,75 |
| ACCEDE | 3 | 0,50 ( 1432 ) | | 3 | 0,50 |
| ACQUIESCE | 4 | 0,25 ( 1620 ) | | 4 | 0,25 |
| 2236. REPLENISH , replenish, [4] Different: 0,3500 | | | | | |
| REPLENISH | 1 | 1,00 ( 1340 ) | | 1 | 1,00 |
| SUPPLY | 2 | 0,75 ( 1375 ) | | 3 | 0,50 |
| REFRESH | 3 | 0,50 ( 1374 ) | | 2 | 0,75 |
| PROVISION | 4 | 0,25 ( 1809 ) | | 4 | 0,25 |

```
1389. HERALD , herald, [4] Different: 0,3500
HERALD                          1  1,00 ( 1384 )    1  1,00
PROCLAIM                        2  0,75 ( 1390 )    2  0,75
PUBLICIZE                       3  0,50 ( 1928 )    4  0,25
ANNOUNCE                        4  0,25 ( 1483 )    3  0,50

120. MEND , to improve, [4] Different: 0,7100
MEND                            1  1,00 ( 1200 )    1  1,00
AID                             2  0,75 ( 1483 )    4  0,25
REMEDY                          3  0,50 ( 1412 )    3  0,50
CURE                            4  0,25 ( 1377 )    2  0,75
```

When, however, the second or third ordinal constituent proves to be the oldest in the string there are already five such measurements in each case with the tighter weight factor values at the penultimate constituent in the position of the historical dominant (cf. Tables 9-10).

At the reverse succession of constituents in the chronological rearrangement of the textual prototypes the permutation weight factor values are larger but also tighter than in the case of the penultimate constituent filling in the place of the historical dominant (cf. Tables 10-11). Two of the values repeat themselves as in the case of the intact present-day and historical positioning of the dominant (cf. Tables 8 and 11).

Table 9

Calculus of the permutation factor values in four-member strings with the second ordinal constituent in the place of the historical dominant

```
2926. WRIGGLE , wriggle, [4] Different: 0,9400
WRIGGLE                         1  1,00 ( 1495 )    2  0,75
SQUIRM                          2  0,75 ( 1691 )    4  0,25
CONVULSE                        3  0,50 ( 1643 )    3  0,50
WIGGLE                          4  0,25 ( 1225 )    1  1,00

1392. HISS , to make a hissing sound, [4] Different: 0,8700
HISS                            1  1,00 ( 1388 )    2  0,75
SIBILATE                        2  0,75 ( 1656 )    3  0,50
FIZZ                            3  0,50 ( 1665 )    4  0,25
SEETHE                          4  0,25 ( 1000 )    1  1,00

1451. IMPUTE , to attribute, [4] Different: 0,6100
IMPUTE                          1  1,00 ( 1375 )    2  0,75
ASCRIBE                         2  0,75 ( 1382 )    3  0,50
ASSIGN                          3  0,50 ( 1297 )    1  1,00
CREDIT                          4  0,25 ( 1541 )    4  0,25

2560. SQUIRT , squirt, [4] Different: 0,7900
SQUIRT                          1  1,00 ( 1460 )    2  0,75
SPURT                           2  0,75 ( 1570 )    4  0,25
SPIT                            3  0,50 ( 950 )     1  1,00
EJECT                           4  0,25 ( 1555 )    3  0,50
```

| | | | | | |
|---|---|---|---|---|---|
| 2582. STIFLE , stifle, [4] Different: 0,3500 | | | | | |
| STIFLE | 1 | 1,00 ( 1387 ) | 2 | 0,75 | |
| SMOTHER | 2 | 0,75 ( 1200 ) | 1 | 1,00 | |
| SUFFOCATE | 3 | 0,50 ( 1526 ) | 3 | 0,50 | |
| EXTINGUISH | 4 | 0,25 ( 1545 ) | 4 | 0,25 | |
| 2566. STAIN , to colour, [4] Different: 0,5000 | | | | | |
| STAIN | 1 | 1,00 ( 1382 ) | 2 | 0,75 | |
| DYE | 2 | 0,75 ( 1000 ) | 1 | 1,00 | |
| TINT | 3 | 0,50 ( 1791 ) | 4 | 0,25 | |
| LACQUER | 4 | 0,25 ( 1688 ) | 3 | 0,50 | |

Table 10

Calculus of the permutation factor values in four-member strings with the penultimate ordinal constituent in the place of the historical dominant

| | | | | | |
|---|---|---|---|---|---|
| 2507. SOLDER , solder, [4] Different: 0,7900 | | | | | |
| SOLDER | 1 | 1,00 ( 1420 ) | 3 | 0,50 | |
| MEND | 2 | 0,75 ( 1200 ) | 1 | 1,00 | |
| PATCH | 3 | 0,50 ( 1500 ) | 4 | 0,25 | |
| CEMENT | 4 | 0,25 ( 1340 ) | 2 | 0,75 | |
| 2448. SIMPER , simper, [4] Different: 0,7100 | | | | | |
| SIMPER | 1 | 1,00 ( 1563 ) | 3 | 0,50 | |
| GIGGLE | 2 | 0,75 ( 1509 ) | 2 | 0,75 | |
| GRIN | 3 | 0,50 ( 1000 ) | 1 | 1,00 | |
| SNICKER | 4 | 0,25 ( 1694 ) | 4 | 0,25 | |
| 2531. SPLICE , splice, [4] Different: 0,6100 | | | | | |
| SPLICE | 1 | 1,00 ( 1524 ) | 3 | 0,50 | |
| KNIT | 2 | 0,75 ( 1000 ) | 1 | 1,00 | |
| GRAFT | 3 | 0,50 ( 1483 ) | 2 | 0,75 | |
| MESH | 4 | 0,25 ( 1532 ) | 4 | 0,25 | |
| 2837. UNDERWRITE , to subscribe, [4] Different: 0,9400 | | | | | |
| UNDERWRITE | 1 | 1,00 ( 1430 ) | 3 | 0,50 | |
| SIGN | 2 | 0,75 ( 1305 ) | 2 | 0,75 | |
| INITIAL | 3 | 0,50 ( 1864 ) | 4 | 0,25 | |
| SEAL | 4 | 0,25 ( 1225 ) | 1 | 1,00 | |
| 1438. IMPERIL , imperil, [4] Different: 1,0000 | | | | | |
| IMPERIL | 1 | 1,00 ( 1596 ) | 3 | 0,50 | |
| JEOPARDIZE | 2 | 0,75 ( 1646 ) | 4 | 0,25 | |
| EXPOSE | 3 | 0,50 ( 1474 ) | 1 | 1,00 | |
| HAZARD | 4 | 0,25 ( 1530 ) | 2 | 0,75 | |
| 2526. SPECULATE , to gamble in business, [4] Different: 1,0600 | | | | | |
| SPECULATE | 1 | 1,00 ( 1599 ) | 3 | 0,50 | |
| RISK | 2 | 0,75 ( 1687 ) | 4 | 0,25 | |
| HAZARD | 3 | 0,50 ( 1530 ) | 2 | 0,75 | |
| VENTURE | 4 | 0,25 ( 1430 ) | 1 | 1,00 | |

Table 11

Calculus of the permutation factor values in four-member strings with the concluding ordinal constituent in the place of the historical dominant

| | | | | | |
|---|---|---|---|---|---|
| 2557. SQUELCH , squelch, [4] Different: 1,1200 | | | | | |
| SQUELCH | 1 | 1,00 ( 1624 ) | | 4 | 0,25 |
| CRUSH | 2 | 0,75 ( 1398 ) | | 3 | 0,50 |
| OPPRESS | 3 | 0,50 ( 1340 ) | | 2 | 0,75 |
| THWART | 4 | 0,25 ( 1250 ) | | 1 | 1,00 |
| 2551. SQUABBLE , squabble, [4] Different: 1,0600 | | | | | |
| SQUABBLE | 1 | 1,00 ( 1604 ) | | 4 | 0,25 |
| ARGUE | 2 | 0,75 ( 1303 ) | | 2 | 0,75 |
| DISAGREE | 3 | 0,50 ( 1494 ) | | 3 | 0,50 |
| FIGHT | 4 | 0,25 ( 900 ) | | 1 | 1,00 |
| 2644. SUCCUMB , to die, [4] Different: 1,0600 | | | | | |
| SUCCUMB | 1 | 1,00 ( 1489 ) | | 4 | 0,25 |
| EXPIRE | 2 | 0,75 ( 1400 ) | | 3 | 0,50 |
| DROP | 3 | 0,50 ( 1000 ) | | 1 | 1,00 |
| CEASE | 4 | 0,25 ( 1300 ) | | 2 | 0,75 |
| 2549. SPUTTER , sputter, [4] Different: 0,9400 | | | | | |
| SPUTTER | 1 | 1,00 ( 1598 ) | | 4 | 0,25 |
| STUMBLE | 2 | 0,75 ( 1303 ) | | 1 | 1,00 |
| STUTTER | 3 | 0,50 ( 1570 ) | | 3 | 0,50 |
| FALTER | 4 | 0.25 ( 1340 ) | | 2 | 0.75 |
| 1437. IMPEND , to threaten, [4] Different: 0,8700 | | | | | |
| IMPEND | 1 | 1,00 ( 1599 ) | | 4 | 0,25 |
| MENACE | 2 | 0,75 ( 1303 ) | | 1 | 1,00 |
| APPROACH | 3 | 0,50 ( 1305 ) | | 2 | 0,75 |
| HOVER | 4 | 0,25 ( 1400 ) | | 3 | 0,50 |
| 2. ACCOST , to greet, [4] Different: 0,9400 | | | | | |
| ACCOST | 1 | 1,00 ( 1578 ) | | 4 | 0,25 |
| ADDRESS | 2 | 0,75 ( 1374 ) | | 2 | 0,75 |
| WELCOME | 3 | 0,50 ( 1000 ) | | 1 | 1,00 |
| SALUTE | 4 | 0,25 ( 1380 ) | | 3 | 0,50 |

The individual permutation factor values are variedly coherent in the corpus. They repeat themselves in the constituents succession patterns over time (cf. the difference in absolute numbers on the upper curve of figure 4). That is why the corpus of the strings itself can be subjected to stratification based of the reshuffle computational effects alongside of the ordinal permutations of constituents. The ranges of the permutation factor values as well as their representation in the corpus segments at relative and absolute chronology of textual prototypes of verbal synonyms in the string prove divergent (cf. the values of relevant points on the axes of figure 2).

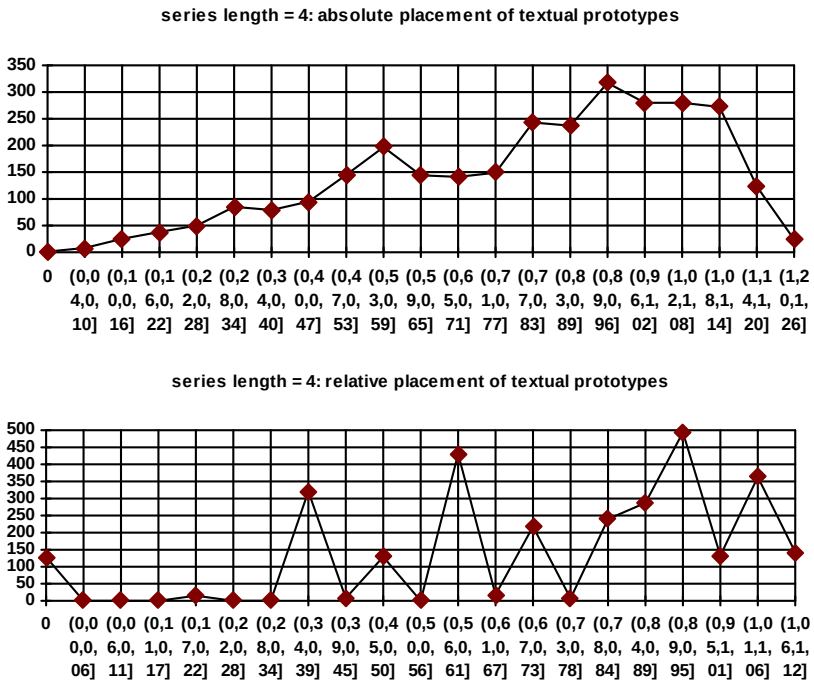



Figure 2. Distribution of the weight permutation factor values within four-member strings

### *4.3. A statistical overview*

The permutation weight factor as a difference in the length of the vectors of the present-day and historical sequences of synonyms is a possible quantification of the semantic medium of a string.

This characteristic is attributable to each synonymous string that is subjected to a diachronic reconstruction on the basis of the *textual prototypes* of its constituents (cf. a random exemplification in Table 12). Conversely, strings tend to group on the strength of value of the difference of their present-day and historical vectors lengths.

Table 12. A numeric distribution of the permutation coefficient values in a sample string. Columns succession in the table: (1) lexical sequence itself; (2) present-day ordinal numbers of constituents; (3) present-day weight factor values on an evenly descending scale (as in formalism (10) above); (4) dates of the constituents earliest OED quotations;

(5) ordinal number of constituents on a relative chronology scale with the ordinal equating of identically dated prototypes; (6) historical weight factor values falling back on the quotations relative chronology on an evenly descending scale with the ordinal equating of identically dated prototypes (as in formalism (11) above); (7) same as column 5; (8) historical weight factor values falling back on the quotations absolute chronology on an unevenly descending scale with the ordinal equating of identically dated prototypes (as in formalism (12) above). Last line: computed permutation factor values $\left\| \vec{w} - \overrightarrow{w^{(y)}} \right\|$ between the present-day and historical (relative and absolute constituents placement, respectively) sequences.

1. ABANDON , to relinquish, [25]

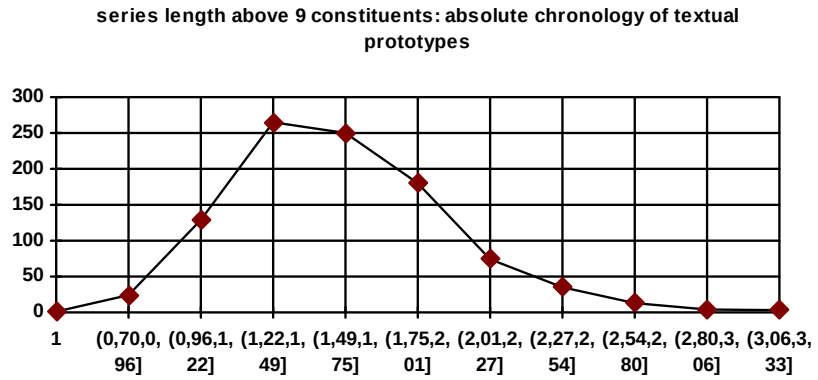
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ABANDON | 1 | 1,00 | ( 1375 ) | 11 | 0,58 | 11 | 0,45 |
| LEAVE | 2 | 0,96 | ( 1000 ) | 4 | 0,86 | 4 | 0,82 |
| WITHDRAW | 3 | 0,92 | ( 1225 ) | 6 | 0,78 | 6 | 0,60 |
| DISCONTINUE | 4 | 0,88 | ( 1479 ) | 14 | 0,48 | 14 | 0,34 |
| CEASE | 5 | 0,85 | ( 1300 ) | 9 | 0,68 | 9 | 0,52 |
| DELIVER | 6 | 0,81 | ( 1225 ) | 6 | 0,78 | 6 | 0,60 |
| DISCARD | 7 | 0,77 | ( 1586 ) | 20 | 0,24 | 20 | 0,23 |
| VACATE | 8 | 0,73 | ( 1643 ) | 23 | 0,12 | 23 | 0,18 |
| EVACUATE | 9 | 0,69 | ( 1526 ) | 16 | 0,40 | 16 | 0,29 |
| SURRENDER | 10 | 0,65 | ( 1466 ) | 13 | 0,52 | 13 | 0,35 |
| YIELD | 11 | 0,62 | ( 825 ) | 1 | 1,00 | 1 | 1,00 |
| DESIST | 12 | 0,04 | ( 1509 ) | 15 | 0,44 | 15 | 0,31 |
| CONCEDE | 13 | 0,54 | ( 1632 ) | 21 | 0,20 | 21 | 0,19 |
| DISCLAIM | 14 | 0,50 | ( 1560 ) | 19 | 0,28 | 19 | 0,26 |
| RENOUNCE | 15 | 0,46 | ( 1375 ) | 11 | 0,58 | 11 | 0,45 |
| BREAK | 16 | 0,42 | ( 851 ) | 2 | 0,96 | 2 | 0,97 |
| EMIGRATE | 17 | 0,38 | ( 1778 ) | 25 | 0,04 | 25 | 0,04 |
| APOSTATIZE | 18 | 0,35 | ( 1552 ) | 18 | 0,32 | 18 | 0,27 |
| ABDICATE | 19 | 0,31 | ( 1541 ) | 17 | 0,36 | 17 | 0,28 |
| SECEDE | 20 | 0,27 | ( 1702 ) | 24 | 0,08 | 24 | 0,12 |
| CEDE | 21 | 0,23 | ( 1633 ) | 22 | 0,16 | 22 | 0,19 |
| WAIVE | 22 | 0,19 | ( 1297 ) | 8 | 0,72 | 8 | 0,52 |
| QUITCLAIM | 23 | 0,15 | ( 1314 ) | 10 | 0,64 | 10 | 0,51 |
| FORGO | 24 | 0,12 | ( 950 ) | 3 | 0,92 | 3 | 0,87 |
| FORSWEAR | 25 | 0,08 | ( 1000 ) | 4 | 0,86 | 4 | 0,82 |
| | | | | Different: 1,9600 | | : 1,9800 | |

A string of synonyms is an object of lexical memory. Its nature as a stored group of words depends on the number of its constituents. Shorter strings are more likely to be stored as a sequence than longer ones.

The threshold between short and long strings may run at the numeric value of the so-called depth hypothesis which is also known as Ingve's hypothesis. It postulates the maximal length of the optimal stored group at seven plus or minus two words. There are 600 such strings in the analyzed thesaurus. They form a border line between the numerically predominant strings of two-three and especially *four* lexemes (the total of 4431 strings) and lengthier strings exceeding nine constituents (the total of 978 series).

The strings whose length exceeds nine constituents reveal only large values of dissimilarity in the present-day and historical sequencing of constituents. The difference in the distribution of the respective permutation weight factor values within such strings obtained proceeding from the absolute and relative chronology of the respective textual prototypes is rather negligible (cf. the curves and respective values on axis x on fig. 3).

Conversely, in the synonymous strings of verbs containing nine and fewer synonyms there are not only large but also quite minimal values of sequential dissimilarity. At large values of dissimilarity their distribution in the case of absolute and relative chronology of constituents is fairly convergent (fig. 4). At smaller values the differences between the permutation weight factor values obtained falling back on textual prototypes datings and their relative chronological positions are more apparent.



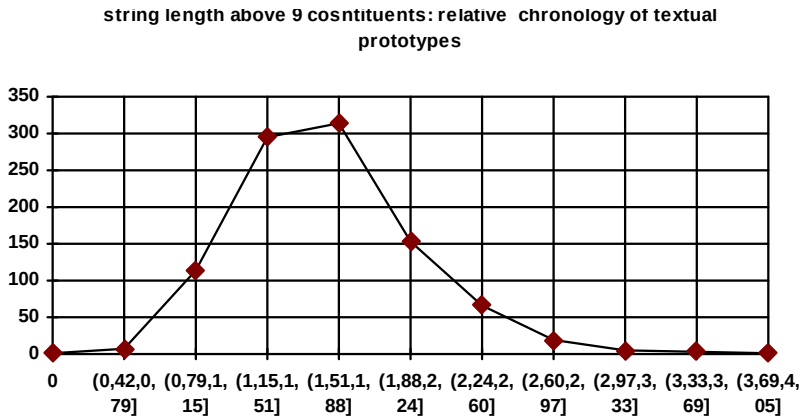



Figure 3. Distribution of the permutation weight factor values in synonymous strings exceeding nine constituents

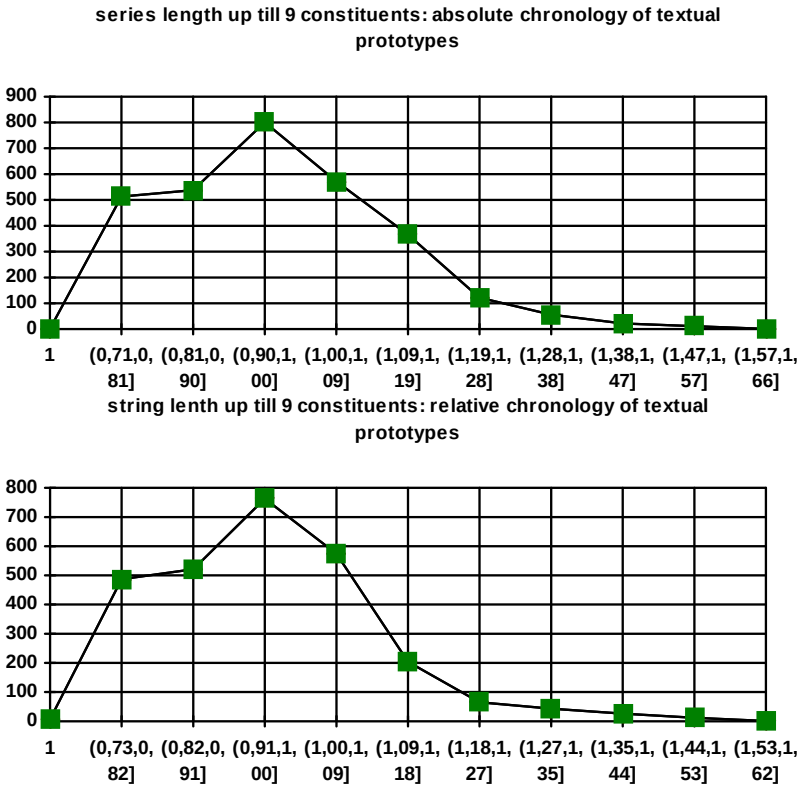



Figure 4. Distribution of higher values of the permutation weight factor in shorter synonymous strings

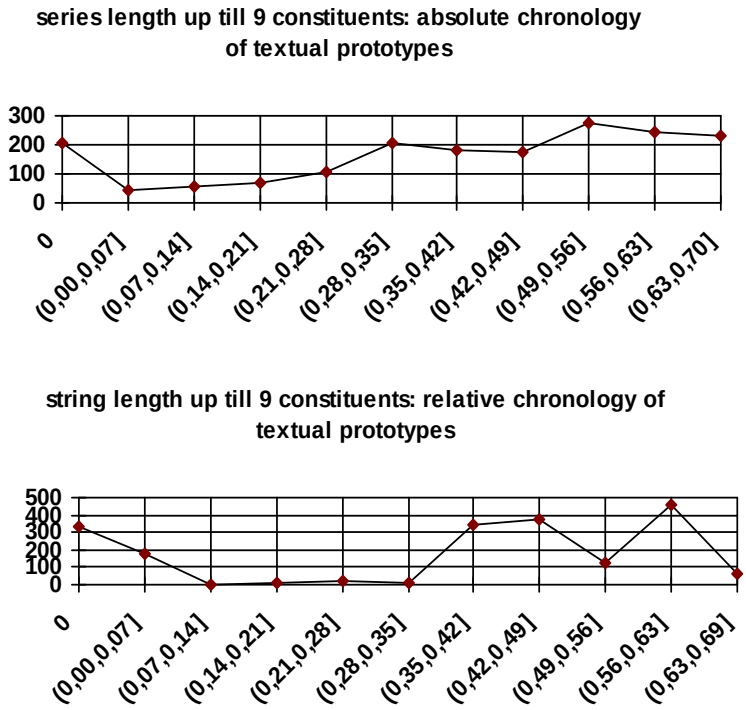



Figure 5. Distribution of lower values of the permutation weight factor in shorter synonymous strings

## *5. CONCLUDING REMARKS*

Apart from the entire corpus of strings subjected to an historical rearrangement could be also its variedly determined partitions from the present-day thesaurus. The chronological homogeneity of the constituents within the string and the diachronic width inside and/or between the periods of its diachronic formation could be related to its length and in part chronological as well as etymological or/and thematic make-up.

The developed framework is based on the application of mathematical formalisms to problems of diachronic onomasiology. Formulas 1-9 in section 3 were suggested by Professor Gabriel Altmann and formulas 11-13 in the same section as well as all the software elaboration by Andriy Pereymybida. The whole approach seems an extension of

the heuristic potential of Levickij's formalism contained in formula (10), which is a tool capable of providing new answers to some old questions. The foundation for this extension can be seen in self-compiled corpora attained from the specifically designed and implemented digitalizations of available dictionary data-bases (cf. Markus 2007; Bilynsky 2007).

## *REFERENCES*

Bilynsky M. (2007): At the crossroads of synonymy and word-formation: thesauri of English deverbatives as software construable lexicographic corpora // *35. Österreichische Linguistiktagung*. Innsbruck, 26-28 Oktober 2007. http://www.onomastik.at-index.pkp?article_id=72.

Kay, Ch J.; Wotherspoon I.A.W. (2002): Turning the dictionary inside out: Some issues in the compilation of a historical thesaurus. In: Diaz Vera, Javier E. (ed.). *A changing world of words. Studies in English historical lexicography, lexicology and semantics.* Amsterdam: Rodopi.

Levickij, V.V.; Sternin, I.Ja. (1989): Ėksperimental'nye metody v semasiologii. Voronež.

Markus, M. (2007): "Joseph's Wrights *English Dialect Dictionary* computerized: architecture and retrieval routine". *Dagstuhl Seminar Proceedings 06491. Digital Historical Corpora. Architecture, Annotation and Retrieval.* [htp///drops.dagstuhl.de/opus/volltexte/2007/1052].

OED = Oxford English Dictionary (Second Edition) on CD-ROM Version 3.0. 2002. Oxford: Oxford University Press

WNWT = Webster's New World Thesaurus. Prepared by W.D. Lutz. New York: Prentice Hall, 1985.

# Дискурсивные и текстовые категории: к разграничению понятий

*Лилия Безуглая (Харьков, Украина)*

## *1. ПОСТАНОВКА ПРОБЛЕМЫ*

Философское и общенаучное понимание категории предполагает предельно общее понятие, отражающее наиболее существенные связи и отношения реальной действительности и познания (Горский 1991: 77). Это понятие связано с категоризацией – построением таксономических классов, которое обеспечивает «отнесение слова (или объекта) к более общему классу (группе) на основе определенных представлений о мире» (Фрумкина 1991: 5). Категоризовать явления действительности и её образы в сознании означает построить упорядоченную картину, классификацию, типологию. Понятие категории позволяет, во-первых, отграничить одно множество объектов, фактов, явлений, признаков от другого множества, а во-вторых, показать для каждого из таких множеств категориальный признак, объединяющий все элементы этого множества.

В лингвистике категория понимается как:

- парадигматическая группировка языковых элементов (категории гласных и согласных, частей речи и т.д.) (Реформатский 2004: 317);
- парадигматическая группировка фактов строения слов и предложений на основе формальных и содержательных категориальных признаков (грамматические категории – морфологические и синтаксические) (Адмони 1986: 11ff; Бондарко 1976);
- парадигматические группировки, в которых в качестве носителей тождественных или близких значений выступают разнообразные языковые средства (морфологические, словообразовательные, синтаксические, лексические, интонационные) во взаимодействии с контекстом (понятийные (Есперсен 2002; Звегинцев 2001; Мещанинов 1978), функционально-семантические (Бондарко 1984), семантические (Сусов 1985), речемыслительные (Кацнельсон 2004), коммуникативные

(Стернин 2002), когнитивно-коммуникативные (Недобух 2002) категории);

- инвариантные, различительные свойства (признаки) множества текстов как знаковых посредников дискурса (текстовые (Богданов 1993), текстово-дискурсивные (Селиванова 2002: 191ff), дискурсивные (Карасик 2004: 240ff) категории).

Очевидно, что первые три трактовки соответствуют широкому пониманию термина ‘категория’, последняя – узкому, что восходит к противопоставлению А.В. Бондарко (1976: 7ff) категории как группы, разряда и как системы признаков определенной категории. Проблемным представляется статус дискурсивных категорий и их соотношение с текстовыми.

Цель статьи – дифференциация понятий дискурсивной и текстовой категории с позиций когнитивно-коммуникативной парадигмы лингвистики на основе разграничения понятий ‘дискурс’ и ‘текст’.

## *2. ДИСКУРС VS. ТЕКСТ*

Рассмотрение сущности дискурса неизменно предполагает соотнесение его с понятием текста. Лингвистические подходы к пониманию дискурса можно разделить на процессуальные, результативные и комплексные.

Результативные подходы характеризуются отождествлением дискурса с текстом с привлечением условий его производства и восприятия (Звегинцев 2001: 170; Карасик 2004: 238). Дискурс, как и текст, остается при этом продуктом, результатом деятельности говорящих индивидов. Определяющим понятием в дефинициях дискурса этого подхода неизменно является понятие ‘текст’: “Дискурс – це зв'язний текст у контексті багатьох конституюючих і фонових чинників – соціокультурних, психологічних і т.д. Дискурс називають зануреним у життя текстом” (Штерн 1998: 87).

Процессуальное понимание дискурса предусматривает динамический процесс, выливающейся в знаковый продукт – текст (Haberland 1999). Такое понимание, на наш взгляд, граничит с противопоставлением дискурса и текста как процесса и продукта, что, несомненно, чревато сложностями методологического плана. Дискурс и текст неразрывны, как неразрывны процесс и продукт:

продукт не мыслим без процесса, как и процесс неизбежно предполагает продукт. Поэтому мы присоединяемся к третьей группе подходов к дискурсу, рассматривающей его в комплексе, как «мисленнєво-комунікативну діяльність, яка є сукупністю процесу й результату і включає як позалінгвальний, так і власне лінгвальний аспект» (Шевченко 2005: 17).

Лингвальный аспект дискурса охватывает речевую деятельность коммуникантов (процесс) и текст (продукт). Экстралингвальный аспект включает когнитивную, коммуникативную деятельность и дискурсивный контекст.

Дискурсивный контекст является экстралингвальной базой функционирования дискурса, ментальной моделью, репрезентирующей знания о возможной или актуальной ситуации. В дискурсивном контексте мы выделяем такие составляющие: онтологический контекст (время, место, физическая среда коммуникации, присутствующие/наблюдатели), коммуникативный (коммуниканты, их коммуникативная компетенция, цели, стратегии и тактики, код и канал связи), социальный (биосоциальные роли, статусы коммуникантов, институциональные аспекты коммуникации), социокультурный (культурологические и социально-исторические аспекты коммуникации), психофизиологический (психическое и физическое состояние коммуникантов), когнитивный (знания коммуникантов, включая знания друг о друге, метазнания об этих знаниях, когнитивные операции коммуникантов), психолингвистический контекст (лингвистическая компетенция коммуникантов).

Дискурсивный контекст не следует путать с контекстом в узком смысле слова – языковым, речевым контекстом, ко-текстом или со-текстом, т.е. окружающим текстом (Лайонз 2003: 287; Макаров 2003: 147).

Деятельность понимаем вслед за А.Н. Леонтьевым как совокупность действий и операций, имеющую общественную природу, кооперативный характер, целенаправленность, структурированность. «Иными словами, деятельность – это не реакция и не совокупность реакций, а система, имеющая строение, свои внутренние переходы и превращения, свое развитие» (Леонтьев 1983: 141). В системе типов деятельности релевантными видятся когнитивная, коммуникативная и речевая деятельность.

Когнитивная деятельность коммуникантов представляет собой процесс обработки ментальных репрезентаций – концептов, сценариев, пропозиций и т.п., который не всегда производится

при помощи языковых знаков. На основе когнитивной деятельности индивидов происходит их коммуникативная деятельность (коммуникация, общение), которая тоже может обходиться и без языковых знаков. Поэтому мы причисляем когнитивную и коммуникативную деятельности к экстралингвальным компонентам дискурса. В отличие от когнитивной деятельности, коммуникация включена в социальные отношения общающихся, поскольку коммуниканты оказываются включенными в систему социальных отношений и вынуждены оформлять свою деятельность в соответствии с определенными социальными конвенциями.

Если коммуникативная деятельность осуществляется при помощи языкового кода, речь идет о речевой деятельности (речевой коммуникации, вербальном общении), которая представляет собой обмен речевыми актами – минимальными единицами дискурса. Подчеркнем, что понятие ‘речевая деятельность’ не приравнивается к понятию ‘речь’. Речь шире, чем речевая деятельность: представляя собой «способ формирования и формулирования мысли посредством языка» (Зимняя 2001: 41), речь присуща всем формам человеческой деятельности, специфика которой «во всех ее проявлениях, обусловленная социально-историческими законами развития, тем самым опосредуется специфически человеческой формой отражения действительности – вербальным мышлением, языком – речью» (Зимняя 2001: 44). Иными словами, речевая деятельность – это процесс внешнего выражения речи.

Тезис А.А. Леонтьева о том, что, «строго говоря, речевой деятельности как таковой не существует», «есть лишь система речевых действий, входящих в какую-то деятельность – целиком теоретическую, интеллектуальную или частично практическую» (Леонтьев 2003: 27), выводит к пониманию взаимосвязи всех видов деятельности, прежде всего, когнитивной, коммуникативной и речевой. Речевая деятельность не может существовать в отрыве от других видов деятельности. Как и деятельность вообще, она определяется мотивом, а значит, является оптимальным средством для достижения определенных целей.

Речевая деятельность невозможна не только без коммуникативной и когнитивной деятельности, но и без активации всех составляющих дискурсивного контекста. Поэтому в каждом конкретном дискурсе сосуществуют все компоненты. С другой стороны, когнитивная и коммуникативная деятельности могут проходить и без речевой – тогда о дискурсе говорить нельзя. В качестве

примера можно привести партию игры в шахматы: осуществляется когнитивная и коммуникативная деятельность, активировано большинство составляющих дискурсивного контекста (кроме психолингвистического), однако дискурс не реализуется. Следовательно, дискурс имеет лингвальную основу, основными его компонентами являются лингвальные – речевая деятельность, речевые акты и текст.

Понятие речевого акта уже прочно утвердилось в лингвистическом обиходе и, на первый взгляд, не вызывает сомнений относительно своей трактовки. Тем не менее, в отдельных работах по когнитивной прагмалингвистике и дискурсивному анализу в это понятие вкладывается новый, более широкий, смысл. Так, М.Л. Макаров отмечает, что «категория "речевой акт" вышла за пределы теории речевых актов *per se*» и составила наряду с дискурсом объект анализа в дискурсивной онтологии (Макаров 2003: 162). Речевой акт представляет собой основанное на коллективной интенции речевое взаимодействие коммуникантов, в процессе которого ими конструируются смыслы – пропозициональные, иллокутивные и перлокутивные. Т.е. речевой акт понимается, во-первых, как взаимодействие, во-вторых, как процесс (ср. у Дж. Лайонза: «действие – это процесс, контролируемый агентом; акт – это единица действия или активности» (Лайонз 2003: 252)). Продукт данного действия – высказывание – фиксируется текстом. Иначе говоря, высказывание является продуктом речевого акта. Отличие же высказывания от предложения состоит в деятельностной природе первого и в структурной природе второго: «высказывание есть высказанное предложение» (Todorov 1987: 32).

Текст представляет собой результативную часть дискурса, вербализованный продукт мыслекоммуникативной деятельности субъектов коммуникации, это «языковой материал, фиксированный на том или ином материальном носителе с помощью начертательного письма (обычно фонографического или идеографического)» (Богданов 1993: 5f). Результативный характер текста подчеркивает и метафора К. Бюлера: «Создатели слова „Text" имели в виду ткань, хотя мне точно неизвестно, какую именно» (Бюлер 2000: 352).

Как образно выразился Г. Хаберланд, текст – это «замерзший дискурс», «если текст может быть в разных местах в разное время, то дискурс является событием, которое совершается здесь и сейчас» (Haberland 1999: 914).

Результативный характер текста не противоречит такому его качеству, как динамичность: текст, отражая дискурс, непременно

отражает и его динамику, речемыслительные процессы его продуцентов. В тексте проявляются все характеристики когнитивной, коммуникативной и речевой деятельности и дискурсивного контекста.

Представленная модель дискурса позволяет выделять различные его типы. На уровне речевой деятельности выделяются речевые (прагматические, стратегические) разновидности дискурса (аргументативный, конфликтный, юмористический и т.п.), на уровне коммуникативной деятельности – социально-коммуникативные разновидности (педагогический, юридический, политический, медицинский и т.п.). На уровне текста релевантными типологиями дискурса являются жанровая (разговорный, публицистический, научный, художественный, деловой), кодовая (немецко-, англо-, украино-, русскоязычный дискурс и т.д.). Диалогический и монологический дискурс выделяются на основании критерия наличия мены коммуникативных ролей (адресант – адресат).

Важно подчеркнуть, что разграничение монологического и диалогического дискурса относится к формальной стороне диалогичности. С функциональной стороны диалогичность понимается как исконное, первичное свойство языка. Принцип диалогической природы языка, сформулированный В. фон Гумбольдтом (Humboldt 1963: 113), проявляется в том, что применение языка всегда направлено на партнера по коммуникации, реально присутствующего, подразумеваемого или представляющего собой самого говорящего. Следовательно, дискурс любого типа является функционально *диалогичным* по своей природе (Карасик 2004: 228). *Диалогический* дискурс имеет своим продуктом диалог в узком смысле слова – диалогический текст.

Таким образом, дискурс определяем как мыслекоммуникативную речевую деятельность коммуникантов в широком (ситуативно-коммуникативном, социо-культурном, когнитивно-психологическом) контексте, зафиксированную текстом. С этих позиций представляется оправданным рассмотрение дискурсивных категорий как когнитивно-прагматических процессов коммуникантов, а текстовых категорий – как свойств текста, формирующихся под воздействием этих процессов.

## *3. ДИСКУРСИВНЫЕ VS. ТЕКСТОВЫЕ КАТЕГОРИИ*

Понятие категории восходит к идеям древнегреческого мыслителя Аристотеля, который первым установил десять «высших родов» или Категорий, указав на их тесную связь с языком. В его систему категорий входили: Сущность, Количество, Качество, Отношение, Место, Время, Положение, Обладание, Действие, Претерпевание (Степанов 1981: 116). В дальнейшем изучение понятия категории лингвистами связывалось с противопоставлением грамматических и понятийных категорий как языковых и внеязыковых (Есперсен 2002: 57; Звегинцев 2001: 261) или как имеющих грамматическое оформление и не имеющих его (Бондарко 1984; Мещанинов 1978), с одной стороны, и с противопоставлением закономерностей категоризации в структуралистской и когнитивистской методологии (детерминизм – прототипический подход) (Фрумкина 1991: 45ff), с другой.

Представляется, что, прежде всего, необходимо различать природные и понятийные категории (иначе, онтологические и гносеологические (Кацнельсон 2004: 171)). Природные категории имеют в качестве денотата определенную реалию, данную человеку в непосредственном опыте и отображенную в человеческом разуме (животные, растения, люди, здания и т.п.). Понятийные категории, напротив, реального денотата не имеют, а являются результатом мыслительной деятельности человека (абстракции, научные понятия и т.п.). Среди понятийных можно выделить языковые или лингвистические категории – такие, которые относятся к системе языка и изучаются лингвистикой. Мы предпочитаем термин ‘лингвистические категории’, поскольку считаем, что все категории являются языковыми в том смысле, что они выражаются и описываются средствами определенного языка.

И дискурсивные, и текстовые категории (как и сами категории дискурса и текста) относятся к понятийным лингвистическим категориям – обобщенным и вариативно репрезентированным в языке и речи при помощи разноуровневых средств (Бондарко 1984; Есперсен 2002: 58; Карасик 2002: 167ff; Мещанинов 1978). Они являются универсальными, однако способы их репрезентации «обусловлены конкретным языком как инструментом познания мира, приоритетными для данного языка грамматическими способами, их комбинаторикой, их взаимодействием со всеми остальными способами выражения содержания в языке» (Карасик 2002: 168).

Свойства понятийных лингвистических категорий выделены И.И. Мещаниновым (Мещанинов 1978). Эти категории:

- являются категориями сознания (т.е. их денотаты не существуют в реальном мире);
- обладают признаком системности;
- вариативно выявляются в языке, в частности, в семантике лексики, синтаксическом строе предложения, в морфологическом оформлении слова.

Дискурсивные категории имеют определенную специфику третьего из данных свойств: они проявляются, прежде всего, в речи – в процессе реализации речевых актов.

Поскольку дискурс есть мыслекоммуникативная речевая деятельность, дискурсивные категории видятся такими, что отражают эту деятельность, обозначая мыслекоммуникативные процессы и состояния: оценка, импликация, эмоции, вежливость, несерьезность и т.п. Они связаны с речемыслительной деятельностью говорящих во время реализации дискурса, которая отображается в единицах языка и речи.

Если дискурсивные категории относятся к процессу реализации дискурса, который основывается на мыслительном и коммуникативно-речевом взаимодействии говорящих, то текстовые категории относятся к тесту как продукту дискурса. Они характеризуют текст, представляя собой определенное его свойство: оценочность, имплицитность, экспрессивность, этикетность, юмористичность и т.п.

Свойства текста являются результатом «работы» дискурсивных категорий, отражением речемыслительных операций коммуникантов, продуцирующих дискурс. Поэтому определенной дискурсивной категории соответствует текстовая категория, с одной стороны, и тип дискурса, выделяемый на основе этой категории, с другой. Примеры таких соответствий представлены в таблице 1.

Основное различие дискурсивных и текстовых категорий состоит в том, что первые характеризуют коммуникантов, продуцирующих дискурс, вторые – образующийся в результате реализации соответствующего дискурса текст. Показательно в этом отношении отсутствие текстовой категории, соотносящейся с дискурсивной категорией неискренности. Текст, возникающий в результате реализации неискреннего дискурса, ничем не отличается от текста, возникающего в результате реализации искреннего дискурса, поскольку перлокутивной целью неискреннего говорящего

является сокрытие от адресата своих действительных пропозициональных установок. Он стремится приблизить языковое оформление своего высказывания к «норме», чтобы не выдать своей неискренности. Поэтому, не зная дискурсивного контекста, исследователь не может достоверно установить степень искренности автора текста.

Таблица 1
Соотношение когнитивно-прагматических типов дискурса, дискурсивных и текстовых категорий

| *тип дискурса* | *дискурсивная категория* | *текстовая категория* |
|---|---|---|
| оценочный | оценка | оценочность |
| непрямой | импликация | имплицитность |
| иронический | ирония | ироничность |
| неискренний | неискренность | – |
| эмотивный | эмоция | экспрессивность |
| этикетный | вежливость | этикетность |
| юмористический | несерьезность | юмор/ юмористичность |
| побудительный | побуждение | побудительность |

Системный, вариативный характер дискурсивной категории отображает сущность речемыслительного процесса, происходящего при реализации соответствующих речевых актов. Поэтому такие категории можно назвать и когнитивно-прагматическими – такими, которые репрезентируются в дискурсе при помощи речевых актов: оценочных, имплицитных, иронических, неискренних, эмотивных, вежливых, несерьезных и т.п. Речевые акты квалифицируются в данном случае не как иллокутивные типы, а на основании репрезентируемой ими категории. При этом выделенные таким образом типы демонстрируют диффузные эффекты: оценочные речевые акты могут быть одновременно и имплицитными, неискренними, несерьезными, имплицитные речевые акты могут быть ироническими, несерьезными, побудительными, любому речевому акту могут сопутствовать эмотивность, вежливость, несерьезность и т.п.

Систематизация таких речевых актов основана на принципе когнитивно-прагматического поля. Система дискурсивной категории предстает как «множество элементов с отношениями и связями между ними, которые образуют определенную целостность» (Бондарко 1984: 47); элементами этой системы являются

соответствующие речевые акты – своеобразные «носители» данной категории, которые объединяются этой категорией в когнитивно-прагматическое поле. Например, категория дискурсивной импликации как речемыслительный процесс конструирования коммуникантами имплицитного смысла репрезентируется в дискурсе при помощи имплицитных речевых актов, которые объединяются в когнитивно-прагматическое поле на основании признаков данной категории и образуют непрямой дискурс. Такой подход позволяет систематизировать имплицтные речевые акты, реализующиеся в немецкоязычном диалогическом дискурсе в виде поля, имеющего доминанту, ядро и периферию (соответственно, полиимпликативные, моноимпликативные и конвенционализированные речевые акты) (Безугла 2007).

И дискурсивные, и текстовые категории образуют соответствующую категориальную сетку (термин О.Н. Колосовой (1993: 6)), представляющую собой совокупность различных категорий, которые соотносятся друг с другом через подчинение системе, в данном случае, дискурсу или тексту. Категориальная сетка дискурса/текста помогает рефлексировать его целостность, выявить многообразие дискурсивных/текстовых связей.

Следует отметить, что не все выделяемые в литературе текстовые категории вписываются в представленную модель. Так, такие категории, как цельность, когезия, когерентность, завершенность, открытость, тема, интерпретируемость, информативность и др. (Богданов 1993; Карасик 2004: 241ff; Селиванова 2002: 199ff), не имеют дискурсивных соответствий. В этой связи целесообразным представляется различение коммуникативных и стилистических текстовых категорий. Первые (примеры которых представлены в таблице) демонстрируют корреляцию с соответствующим дискурсом, поскольку относятся к продуцирующим его коммуникантам, осознающим эти категории. Вторые представляют собой жанрово-стилистические параметры текста, выделяемые исследователем безотносительно к деятельности его продуцента/продуцентов. Их анализ может осуществляться без привлечения дискурсивного контекста.

Коммуникативные текстовые категории являются дискурсивно релевантными. Их анализ невозможен без анализа всех составляющих дискурсивного контекста, без учета когнитивно-прагматических характеристик текста как продукта соответствующего дискурса. Например, анализ экспрессивности текста неизбежно

предполагает обращение к эмоциям коммуникантов, их мотивам, стратегиям, интенциям, к ситуативным параметрам коммуникации и пр.

## *4. ВЫВОДЫ*

Понимание дискурса как мыслекоммуникативной речевой деятельности коммуникантов, зафиксированной текстом, предполагает включение текста в дискурс. Речемыслительные процессы и состояния коммуникантов в ходе реализации дискурса (такие, так оценка, импликация, эмоции и т.п.) представляют собой дискурсивные категории. Они определяют тип реализующегося дискурса (оценочный, непрямой, эмотивный и т.п.) и находят отражение в соответствующих текстовых категориях, характеризующих возникающий текст (оценочность, имплицтность, эмотивность и т.п.).

Изучение дискурсивных категорий представляется продуктивным путем установления когнитивно-прагматических характеристик и особенностей вербализации речеактовых смыслов в дискурсах различных типов.

Перспективными считаем исследования отдельных дискурсивных и текстовых категорий, их соотношения, когнитивно-прагматических особенностей реализации соответствующих дискурсов.

## *ЛИТЕРАТУРА*

Адмони В.Г. (1986): *Теоретическая грамматика немецкого языка: Строй современного немецкого языка* / В. Г. Адмони. – М.: Просвещение.

Безугла Л.Р. (2007): *Вербалізація імпліцитних смислів у німецькомовному діалогічному дискурсі* / Л. Р. Безугла. – Харків: ХНУ ім. В.Н. Каразіна.

Богданов В.В. (1993): *Текст и текстовое общение* / В.В. Богданов. – СПб: РИО СПбГУ.

Бондарко А.В. (1976): *Теория морфологических категорий* / А. В. Бондарко. – Л.: Наука.

Бондарко А.В. (1984): *Функциональная грамматика* / А.В. Бондарко. – Л.: Наука, Ленинггр. отд.

Бюлер К. (2000): *Теория языка. Репрезентативная функция языка* / Карл Бюлер; пер. с нем. – М.: Прогресс.

Горский Д.П. (1991): *Краткий словарь по логике* / Д. П. Горский, А. А. Ивин, А. Л. Никифоров. – М.: Просвещение.

Есперсен О. (2002): *Философия граматики* / О. Есперсен; пер. с англ.; изд. 2-е, стер. – М.: Эдиториал УРСС.

Звегинцев В.А. (2001): *Предложение и его отношение к языку и речи* / В. А. Звегинцев; изд. 2-е, стер. – М.: Эдиториал УРСС.

Зимняя И.А. (2001): *Лингвопсихология речевой деятельности* / И. А. Зимняя. – М.–Воронеж: НПО «МОДАК».

Карасик В.И. (2002): *Язык социального статуса* / В. И. Карасик. – М.: Гнозис.

Карасик В.И. (2004): *Языковой круг: личность, концепты, дискурс* / В. И. Карасик. – М.: Гнозис.

Кацнельсон С.Д. (2004): *Типология языка и речевое мышление* / С. Д. Кацнельсон; изд. 3-е, стер. – М.: Эдиториал УРСС.

Колосова О.Н. (1993): *Языковые факты в системе мыслительных категорий* / О. Н. Колосова. – Тверь: Тверской гос. ун-т.

Лайонз Дж. (2003): *Лингвистическая семантика: Введение* / Дж. Лайонз; пер. с англ. – М.: Языки славянской культуры.

Леонтьев А.А. (2003): *Язык. Речь. Речевая деятельность* / А. А. Леонтьев; изд. 2-е, стер. – М.: Эдиториал УРСС.

Леонтьев А.Н. (1983): Деятельность. Сознание. Личность / А. Н. Леонтьев // *Избранные психологические произведения*: в 2-х т. – М.: Педагогика. – Т. 2.

Макаров М.Л. (2003): *Основы теории дискурса* / М. Л. Макаров. – М.: Гнозис.

Мещанинов И.И. (1978): *Члены предложения и части речи* / И. И. Мещанинов. – Л.: Наука.

Недобух С.А. (2002): *Когнитивно-коммуникативная категория персональности*: Автореф. дисс. … канд. филол. наук: 10.02.19 / С. А. Недобух; Тверской гос. ун-т. – Тверь.

Реформатский А.А. (2004): *Введение в языковедение* / А. А. Реформатский; изд. 5-е, испр. – М.: Аспект Пресс.

Селиванова Е.А. (2002): *Основы лингвистической теории текста и коммуникации* / Е.А. Селиванова. – К.: Фитосоциоцентр.

Степанов Ю.С. (1981): *Имена. Предикаты. Предложения*. (Семиологическая грамматика) / Ю. С. Степанов. – М.: Наука.

Стернин И.А. (2002): О национальном коммуникативном сознании / И. А. Стернин //*Лингвистический вестник*: Сб. науч. тр. Вып. 4. – Ижевск: УМО «Sancta Lingua», 87-95.

Сусов И.П. (1985): Проблема семантических категорий в синтаксисе / И. П. Сусов // *Семантические категории языка и методы их изучения*: тез. докл. всесоюз. науч. конф. – Уфа: Башкирский гос. ун-т. – Ч. 1, 6-7.

Фрумкина Р.М. (1991): *Семантика и категоризация* / Р. М. Фрумкина, А. В. Михеев, А. Д. Мостовая, Н. А. Рюмина. – М.: Наука.

Шевченко І.С. (2005): Когнітивно-комунікативна парадигма і аналіз дискурсу / І. С. Шевченко // *Дискурс як когнітивно-комунікативний феномен*. − Харків: Константа, 9-20.

Штерн І.Б. (1998): *Вибрані топіки та лексикон сучасної лінгвістики* / І. Б. Штерн. – К.: АртЕк.

Haberland H. (1999): Text, Discourse, Discours: The latest report from the Terminology Vice Squad / H. Haberland // *Journal of Pragmatics*. – 1999. – Vol. 31, 911-918.

Humboldt W. v. (1963): Über den Dualis / Gelesen in der Akademie der Wissenschaften am 26. April 1827 / Wilhelm von Humboldt // *ders. Schriften zur Sprachphilosophie*. – Darmstadt: Fink, 113-143.

Todorov T. (1987): *La notion de la literature at autres essays* / T. Todorov. – Paris: Ed. du Seuil.

# Rhythmische Einheiten in Hülsen, *Natur-Betrachtungen* (1800)

*Karl-Heinz Best (Göttingen, Deutschland)*

## *1. RHYTHMISCHE EINHEITEN IM DEUTSCHEN*

Rhythmische Einheiten sind seit einiger Zeit ein interessanter Gegenstand für die Quantitative Linguistik (Strauss, Fan & Altmann 2008: 59f.). Es handelt sich bei den rhythmischen Einheiten um die Zahl der unbetonten Silben zwischen zwei betonten. Es geht vor allem um die Frage, ob die Häufigkeiten, mit denen rhythmische Einheiten verschiedener Länge in Texten auftreten, einem bestimmten Verteilungsgesetz folgen. Schon die Beobachtungen von Marbe (1904) zeigen, dass die kleinsten rhythmischen Einheiten aus nur zwei betonten Silben bestehen, die größten im Deutschen aber bis zu zehn unbetonte Silben enthalten, sodass sich elf Längenklassen ergeben.

Seit Marbes (1904) bahnbrechender Untersuchung hat es mehrere derartige Untersuchungen zu deutschen Texten gegeben (Bianchi 1922; Best 2001, 2002, 2006; Gropp 1915, 1916; Kaßel 2002), die die Datenbasis erweiterten.

In Best (2001) wurden die Untersuchungen Marbes wieder aufgegriffen und daraufhin geprüft, ob man auf sie eines der bekannten Verteilungsmodelle anwenden könnte, die Wimmer u.a. (1994) sowie Wimmer & Altmann (1996) für Wortlängen entwickelten. Die Idee war, dass es sich um ein Verteilungsgesetz handeln sollte, das sich generell für sprachliche Einheiten beliebiger Art als Modell eignen könnte. Als Ergebnis stellte sich heraus, dass sich in fast allen Fällen die Hyperpoisson-Verteilung

$$(1) \quad P_x = \frac{a^{x-1}}{b^{(x-1)}\,{}_1F_1(1;b;a)}, \quad x = 1,2,3,...$$

in 1-verschobener Form bewährte, obwohl es sich um Texte verschiedener Art handelte (Briefe, literarische Texte, Pressetexte). Nur die Daten

von Gropp ließen sich bisher nicht mit dem gleichen Modell bearbeiten, allerdings auch mit keinem anderen. Es bewahrheitet sich damit auch unter dem Aspekt der Verteilung rhythmischer Einheiten, dass es sich bei dem von Gropp bearbeiteten Text von August Ludwig Hülsen (1800. Natur-Betrachtungen auf einer Reise durch die Schweiz. *Athenaeum*, Dritten Bandes Erstes Stück, 34-57) um einen "eigentümlich scheinenden Rhythmus" handelt (Gropp 1915: 17); Gropp bezieht sich damit auf die Aufmerksamkeit, die Hülsens Text zur Zeit der Romantik bei seinen Zeitgenossen fand. Es handelt sich bei den *Natur-Betrachtungen* um einen Prosatext mit passagenweise deutlich rhythmisierter Sprache. Eine Vermutung dazu, warum die Anpassung eines Modells an die rhythmischen Einheiten dieses Textes misslang, war, dass Gropp willkürliche Text-Abschnitte bildete und dies die Ursache für das Scheitern sein könnte (Best 2008). Man geht allgemein davon aus, dass eine derartige Datenerhebung sich auf die Modellierung ungünstig auswirken kann (Altmann 1992). Allerdings gilt dieses Argument auch für die Erhebungen von Marbe (Best 2001) und Bianchi (Best 2006a), ohne dass es dadurch in diesen Fällen zu Misserfolgen bei der Anpassung eines Modells gekommen wäre. Es gibt also hinreichend Gründe, sich die Daten der Untersuchung von Gropp noch einmal genauer anzusehen.

## *2. DAS DILEMMA*

Zu den Daten: Gropp untersuchte 8 Abschnitte des angegebenen Textes von je 1000 Wörtern Länge und einen restlichen, kürzeren Abschnitt; zusätzlich gibt er die Verteilung rhythmischer Einheiten für den gesamten Text an, sodass insgesamt 10 Dateien zur Verfügung stehen.

Am Beispiel der Daten des ersten Textabschnittes soll das Problem, das sich hier stellt, erläutert werden. Die folgende Tabelle gibt Gropps Auszählungsergebnis sowie die Anpassung von Modell (1) an diese Daten wieder.

Legende zu Tabelle 1:

$x$ - Längenklasse der rhythmischen Einheit, beginnend mit $x = 1$ für rhythmische Einheiten, bei denen zwischen zwei betonten Silben keine unbetonte steht
$n_x$ - Anzahl der rhythmischen Einheiten der jeweiligen Klasse im Text
$NP_x$ - Anzahl der rhythmischen Einheiten der jeweiligen Klasse aufgrund der Anpassung der 1-verschobenen Hyperpoisson-Verteilung

*a*, *b* - Parameter der Hyperpoisson-Verteilung
$X^2$ - Chiquadrat
*FG* - Freiheitsgrade
*P* - Überschreitungswahrscheinlichkeit des Chiquadrats
*C* - Diskrepanzkoeffizient $X^2/n$
| - zusammengefasste Klassen

Diese Legende gilt sinngemäß für alle folgenden Tabellen. Statt rhythmischer Einheiten geht es im Abschnitt 5 um die Länge von Fußfolgen.

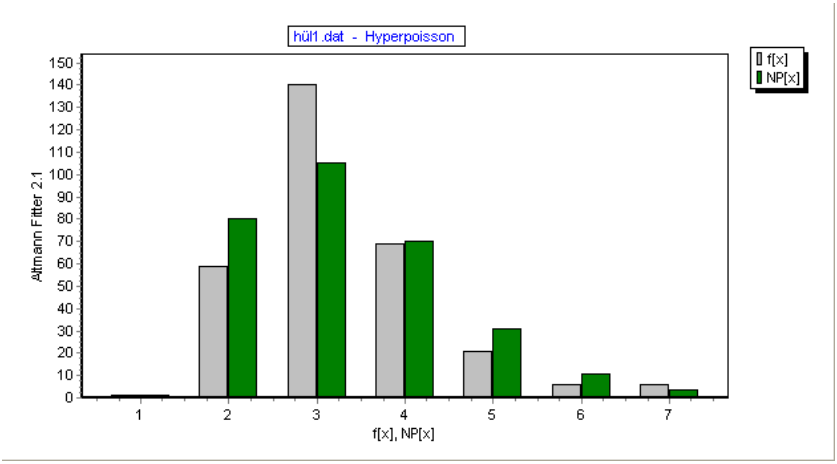
Eine Anpassung der Hyperpoisson-Verteilung mit $P \geq 0.05$ gilt als zufriedenstellend; Anpassungen mit $0.01 \leq P < 0.05$ erfüllen diese Bedingung nicht, werden aber noch toleriert. Der Diskrepanzkoeffizient *C* kommt bei umfangreicheren Dateien zu Einsatz und sollte das Kriterium $C \leq 0.01$ erfüllen, um eine gute Anpassung des Modells an die Daten anzuzeigen.

Tabelle 1
Anpassung der 1-verschobenen Hyperpoisson-Verteilung
an Textabschnitt 1 aus Hüllen (nach Gropp 1915: 25)

| $x$ | $n_x$ | $NP_x$ |
|---|---|---|
| 1 | 1 | 1.36 |
| 2 | 59 | 80.09 |
| 3 | 140 | 105.34 |
| 4 | 69 | 70.05 |
| 5 | 21 | 31.18 |
| 6 | 6 | 10.43 |
| 7 | 6 | 3.56 |
| $a$ = 1.3453 $b$ = 0.0228 $FG$ = 4 $X^2$ = 23.9437<br>$P$ = 0.0001 $C$ = 0.0793 | | |

($x$ = 1: rhythmische Einheit ohne unbetonte Silben zwischen zwei betonten; $x$ = 2: eine unbetonte Silbe zwischen zwei betonten; etc.)

Das Testergebnis ist miserabel: Weder *P* noch *C* erfüllen die genannten Kriterien. Die folgende Graphik zu Tabelle 1 veranschaulicht dies:




Die Graphik zeigt ebenso wie die Tabelle 1 ganz deutlich das Problem: Die größten Abweichungen zwischen den beobachteten (jeweils linke Säule) und den aufgrund des Modells berechneten Werten ist bei $x = 2$ und $x = 3$ auffallend groß. Das heißt: Es wurden zuwenig rhythmische Einheiten beobachtet, bei denen zwischen zwei betonten Silben nur eine unbetonte steht ($x = 2$) und zu viele, bei denen zwischen den betonten Silben zwei unbetonte vorkommen ($x = 3$). Dieser Befund ist bei allen Textabschnitten immer derselbe; nur bei Textabschnitt 4 ist er etwas weniger stark ausgeprägt, so dass die Anpassung der Hyperpoisson-Verteilung in diesem einen Fall dennoch ohne Zusammenfassung der beiden Längenklassen sofort gelingt. In allen anderen Fällen sind die Differenzen bei ihnen zu groß. Diesem Fall kann man auf unterschiedliche Weise begegnen. Eine der Möglichkeiten besteht darin, dass man die beiden fraglichen Längenklassen $x = 2$ und $x = 3$ zusammenfasst und dann die Anpassungen neu berechnet. Dies soll nun demonstriert werden.

## *3. MODELLIERUNG DER RHYTHMISCHEN EINHEITEN IN EINEM SACHTEXT*

Es folgen nun die Tabellen mit der Anpassung der 1-verschobenen Hyperpoisson-Verteilung an alle Textabschnitte einzeln und an eine Zusammenfassung aller Abschnitte in einer Datei; nur bei Textabschnitt 4 wird auf die Zusammenfassung der Klassen $x = 2$ und $x = 3$ verzichtet. Die Ergebnisse stellen sich nun wie folgt dar:

Tabelle 2
Anpassung der 1-verschobenen Hyperpoisson-Verteilung
an die Textabschnitte aus Hüllen (nach Gropp)

| | Textabschnitt 1 | | Textabschnitt 2 | | Textabschnitt 3 | |
|---|---|---|---|---|---|---|
| $x$ | $n_x$ | $NP_x$ | $n_x$ | $NP_x$ | $n_x$ | $NP_x$ |
| 1 | 1 | 1.36 | 3 | 4.05 | 8 | 9.92 |
| 2 | 59 | 80.09\| | 61 | 82.32\| | 83 | 102.89\| |
| 3 | 140 | 105.34\| | 138 | 104.75\| | 145 | 112.05\| |
| 4 | 69 | 70.05 | 61 | 68.80 | 58 | 64.39 |
| 5 | 21 | 31.18 | 35 | 30.45 | 26 | 25.14 |
| 6 | 6 | 10.43 | 4 | 10.16 | 3 | 7.43 |
| 7 | 6 | 3.56 | 1 | 2.72\| | 1 | 2.19 |
| 8 | | | 1 | 0.75\| | | |
| | $a = 1.3452$ $b = 0.0228$ $X^2 = 7.982$ $FG = 3$ $P = 0.0464$ | | $a = 1.3574$ $b = 0.0668$ $X^2 = 7.420$ $FG = 3$ $P = 0.06$ | | $a = 1.2168$ $b = 0.1173$ $X^2 = 5.117$ $FG = 3$ $P = 0.16$ | |

Zu Textabschnitt 1: Durch die Zusammenfassung der beiden Längenklassen x = 2 und x = 3 verringern sich die Freiheitsgrade um 1; die Anpassung des Modells ist wesentlich besser als beim ersten Versuch, wenn auch nicht wirklich zufriedenstellend.

Tabelle 3
Anpassung der 1-verschobenen Hyperpoisson-Verteilung
an die Textabschnitte aus Hüllen (nach Gropp)

| | Textabschnitt 4 | | Textabschnitt 5 | | Textabschnitt 6 | |
|---|---|---|---|---|---|---|
| $x$ | $n_x$ | $NP_x$ | $n_x$ | $NP_x$ | $n_x$ | $NP_x$ |
| 1 | 4 | 3.01 | 6 | 5.67 | 3 | 3.72 |
| 2 | 71 | 80.40 | 71 | 90.57\| | 97 | 120.30\| |
| 3 | 115 | 104.28 | 143 | 108.35\| | 164 | 120.72\| |
| 4 | 71 | 69.32 | 54 | 67.33 | 48 | 61.52 |
| 5 | 29 | 30.97 | 25 | 28.26 | 18 | 21.01 |
| 6 | 8 | 10.42 | 11 | 8.96 | 2 | 5.40 |
| 7 | 4 | 3.59 | 2 | 2.87 | 2 | 1.33 |
| | $a = 1.3633$ $b = 0.0510$ $X^2 = 3.301$ $FG = 4$ $P = 0.51$ | | $a = 1.2932$ $b = 0.0810$ $X^2 = 4.905$ $FG = 3$ $P = 0.18$ | | $a = 1.0356$ $b = 0.0320$ $X^2 = 7.391$ $FG = 3$ $P = 0.06$ | |

Tabelle 4
Anpassung der 1-verschobenen Hyperpoisson-Verteilung
an die Textabschnitte aus Hüllen (nach Gropp)

| | Textabschnitt 7 | | Textabschnitt 8 | | Textabschnitt 9 | |
|---|---|---|---|---|---|---|
| $x$ | $n_x$ | $NP_x$ | $n_x$ | $NP_x$ | $n_x$ | $NP_x$ |
| 1 | 9 | 13.76 | 6 | 3.57 | 11 | 10.34 |
| 2 | 64 | 97.83\| | 73 | 95.93\| | 53 | 66.64\| |
| 3 | 159 | 107.24\| | 145 | 111.81\| | 93 | 69.14\| |
| 4 | 64 | 63.68 | 56 | 66.60 | 31 | 39.00 |
| 5 | 17 | 25.93 | 25 | 26.65 | 11 | 15.11 |
| 6 | 4 | 8.04 | 8 | 8.03 | 6 | 4.46 |
| 7 | 2 | 2.52 | 1 | 1.94\| | 1 | 1.32 |
| 8 | | | 1 | 0.47\| | | |
| | $a$ = 1.2959 $b$ = 0.1822 $X^2$ = 8.429 $FG$ = 3 $P$ = 0.0379 | | $a$ = 1.2183 $b$ = 0.0453 $X^2$ = 5.024 $FG$ = 3 $P$ = 0.17 | | $a$ = 1.2364 $b$ = 0.1918 $X^2$ = 4.180 $FG$ = 3 $P$ = 0.24 | |

Führt man einmal alle Daten aus den 9 Textabschnitten zusammen, erhält man folgendes Ergebnis:

Tabelle 5
Anpassung der 1-verschobenen Hyperpoisson-Verteilung
an alle Textabschnitte zusammen (nach Gropp)

| $x$ | $n_x$ | $NP_x$ |
|---|---|---|
| 1 | 51 | 47.32 |
| 2 | 632 | 823.54\| |
| 3 | 1242 | 953.62\| |
| 4 | 512 | 571.12 |
| 5 | 207 | 230.68 |
| 6 | 52 | 70.29 |
| 7 | 20 | 17.19 |
| 8 | 2 | 4.24 |
| $a$ = 1.2405 $b$ = 0.0713 $FG$ = 4 $X^2$ = 20.515 $C$ = 0.0075 | | |

Als Ergebnis kann nun festgestellt werden: An die neun Textabschnitte kann ebenso wie an die zusammengefassten Daten die 1-verschobene Hyperpoisson-Verteilung angepasst werden. Beim ersten und beim siebenten Textabschnitt sind die Ergebnisse nicht wirklich zufriedenstellend, aber doch auch nicht so schlecht, dass man sie ganz verwerfen müsste. Die Hypothese, dass rhythmische Einheiten sich in

beliebigen Texten gesetzmäßig verhalten, wird durch dieses Ergebnis unterstützt.

## *4. MODIFIKATION*

Einen Fall mit Abweichungen bei nur zwei der Längenklassen kann man also, wie gezeigt, durch Zusammenfassung der betroffenen Klassen lösen. Eine weitere Möglichkeit besteht darin, Modelle zu nutzen, die wie die Cohen-C-Poisson-Verteilung und die Pandey-Poisson-Verteilung Verschiebeparameter enthalten. Diese beiden Verteilungen erwiesen sich im vorliegenden Fall aber nicht als geeignet.

Ein anderer Weg, die obige Idiosynkrasie zu erfassen, besteht in der Modifikation des ursprünglichen Modells. Das Modell der Hyper-poisson-Verteilung stellt einen Attraktor dar, in dessen Bereich sich rhythmische Muster bewegen und nur die Parameter der Verteilung sich nach der Art des Textes, Inhalts usw. ändern. Dies ist nur der Ausdruck der üblichen Variabilität, die allen linguistischen Daten eigen ist.

Manchmal kann man aber auch beobachten, dass der Attraktor an einer bestimmten Stelle deformiert wird. Im Fall von Hüllen ist dies ja bei $x = 2$ und $x = 3$ gegeben. Da diese Deformation sehr gezielt ist, kann man sagen, dass diese Daten vom Attraktor abweichen möchten und diese Änderung an den besagten Stellen anfing. Der Punkt $x = 2$ hat immer mehr Häufigkeiten als erwartet, der Punkt $x = 3$ in jedem Fall weniger. Die Aufgabe des Modellierens kann auch darin gesehen werden, diese Tatsache mit einem zusätzlichen Parameter zu erfassen. Bezeichnen wir diesen Parameter als $\alpha$. Wenn man einen $\alpha$-Teil der theoretischen Häufigkeiten von $x = 2$ auf $x = 3$ verschiebt, bekommt man eine adäquatere Erfassung der Daten. Man kann dieses $\alpha$ festsetzen oder von Fall zu Fall variieren, genauso wie die anderen Parameter, um eine bessere Anpassung zu erreichen. Betrachtet man hier zum Beispiel die zusammengefassten Daten in Tabelle 5, stellt man fest, welchen Teil von $NP_2$ man auf $NP_3$ verschieben muss, um einen Ausgleich zu schaffen. Die Differenz zwischen

$$\frac{823.54 - 632}{823.54} = 0.2326 = \alpha \ ,$$

d.h. 23.26% der Häufigkeit von $NP_2$ kann auf $NP_3$ übertragen werden. Dies ergibt dann $\alpha NP_2 = 0.2326\ (823.54) = 191.56$. Diesen Betrag subtrahiert man von $NP_2$: $NP_2 - 191.56 = 823.54 - 191.56 = 631.98 = NP_2^*$, und den gleichen Betrag addiert man zu $NP_3$, d.h. $NP_3^* = NP_3 + 191.56 = 953.63 + 191.56 = 1145.19$. Formal kann man diese Tatsache als

$$NP_2^* = NP_2\ (1 - \alpha)$$
$$NP_3^* = NP_3 + \alpha NP_2$$

erfassen, wobei $P^*$ die neue Wahrscheinlichkeit, $P$ die alte ist. Es ist zu bemerken, dass man mit dieser Technik nicht unbedingt eine bessere Anpassung erreichen muss als mit der Zusammenfassung von Klassen, denn man gewinnt eine Klasse, dafür aber fügt man einen Parameter hinzu, so dass die Zahl der Freiheitsgrade die gleiche bleibt. Dieses Verfahren zeigt lediglich, dass in den Daten eine Idiosynkrasie vorliegt. Falls die abweichenden Klassen nicht benachbart sind, muss man diese Technik verwenden, weil eine Zusammenfassung mehrerer gut ausgeprägter Klassen zu grob wäre.

Die modifizierte Formel würde sich als

$$(2) \qquad NP_x^* = \begin{cases} NP_x\,, & x = 1,4,5,... \\ NP_2(1-\alpha), & x = 2 \\ NP_3 + \alpha NP_2, & x = 3 \end{cases}$$

ergeben, wobei $P_x$ aus Formel (1) zu entnehmen ist.

## *5. ZUSAMMENFASSUNG UND PERSPEKTIVE*

Zunächst ist darauf zu verweisen, dass auch die Erhebungen von Gropp (1915, 1916) sich in das Bild der bisherigen Untersuchungen zu rhythmischen Einheiten im Deutschen einfügen. Nach wie vor muss die Hyperpoisson-Verteilung als das offenbar für deutsche Texte geeignetste Modell angesehen werden. Warum in den *Natur-Betrachtungen* von Hüllen anders als in allen anderen Texten eine deutliche Verlagerung von rhythmischen Einheiten mit nur einer unbetonten Silbe auf solche mit zwei unbetonten Silben erfolgte, muss vorläufig unbeantwortet bleiben. Es könnte sich dabei um ein Phänomen des persönlichen Stils des Autors oder auch des Textsortenstils handeln. Da der Text vom

Anfang des 19. Jahrhunderts stammt, muss auch an einen Zeitstil gedacht werden; dagegen spricht allerdings, dass andere Texte aus dieser Zeit bisher keinen derartigen Effekt zeigen.

Die Arbeit reiht sich mit ihren Ergebnissen aber auch in die Ergebnisse der Untersuchungen zu einigen anderen Sprachen ein. Es hat sich dabei gezeigt, dass rhythmische Einheiten teilweise anderen Modellen folgen als im Deutschen. Im Englischen hat sich bei Briefen und Pressetexten wiederum die Hyperpoisson-Verteilung bewährt (Kaßel 2002). Für das Russische weist Kelih (2008: 114-117) auf Untersuchungen Tomaševskijs aus den 20er Jahren hin; Grzybek & Kelih (2005: 46f.) nehmen dazu erfolgreiche Anpassungen der Binomialverteilung vor. Diesen Befund zum Russischen bestätigt Knaus (2008) mit weiteren Tests, während Eom (2006) und Lehfeldt (2003) eine modifizierte Form, die erweiterte positive Binomialverteilung, anwendeten. Bleibt noch Altgriechisch zu erwähnen, wo sich die geometrische Verteilung bei einigen Texten als gutes Modell erweist (Best 2006b).

Trotz aller bisherigen Bemühungen muss man noch offen lassen, welche Effekte der Sprache, der Zeit, der Textsorte oder auch dem einzelnen Autor zuzuschreiben sind. Um dies herauszufinden, werden noch wesentlich mehr Daten benötigt. Die Hypothese, dass sprachliche Einheiten unterschiedlicher Komplexität sich in Texten oder Textabschnitten gesetzmäßig verteilen, bewährt sich bisher jedoch in jedem Einzelfall.

## *LITERATUR*

Altmann, G. (1992): Das Problem der Datenhomogenität. In: Rieger, Burghard (Hrsg.), *Glottometrika 13* (S. 287-298). Bochum: Brockmeyer.

Best, K.-H. (2001): Zur Verteilung rhythmischer Einheiten in deutscher Prosa. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten* (S. 162-166). Göttingen: Peust & Gutschmidt.

Best, K.-H. (2001): Probability Distributions of Language Entities. *Journal of Quantitative Linguistics 8, 1-11.*

Best, K.-H. (2002): The Distribution of Rhythmic Units in German Short Prose. *Glottometrics 3, 136-142.*

Best, K.-H. (2005): Längen rhythmischer Einheiten. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 208-214). Berlin/ NewYork: de Gruyter.

Best, K.-H. (2006a): Lorenzo Bianchi (1889-1960). *Glottometrics 14, 72-74.*

Best, K.-H. (2006b): Rhythmische Einheiten im Altgriechischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 73-76.*

Best, K.-H. (2008): Quantitative Untersuchungen zum Rhythmus. *Göttinger Beiträge zur Sprachwissenschaft* (eingereicht).

Bianchi, L. (1922): *Untersuchungen zum Prosa-Rhythmus Johann Peter Hebels, Heinrich von Kleists und der Brüder Grimm.* Heidelberg: Weiss'sche Universitätsbuchhandlung.

Eom, J. (2006): *Rhythmus im Akzent. Zur Modellierung der Akzentverteilung als einer Grundlage des Sprachrhythmus im Russischen.* München: Verlag Otto Sagner. (Diss. phil., Göttingen, 2006)

Gropp, F. (1915): *Zur Ästhetik und statistischen Beschreibung des Prosarhythmus.* Würzburg, Diss. phil. Auch in:

Gropp, F. (1916): Zur Ästhetik und statistischen Beschreibung des Prosarhythmus. *Fortschritte der Psychologie und ihrer Anwendungen IV, 43-79.*

Grzybek, P., & Kelih, E. (2005): Zur Vorgeschichte quantitativer Ansätze in der russischen Sprach- und Literaturwissenschaft. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 23-64). Berlin/NewYork: de Gruyter.

Kaßel, A. (2002): *Zur Verteilung rhythmischer Einheiten in deutschen und englischen Texten.* Staatsexamensarbeit, Göttingen.

Kelih, E. (2008): *Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenshaft.* Hamburg: Kovač. (= Diss. Graz, 2007)

Knaus, M. (2008): Zur Verteilung rhythmischer Einheiten in russischer Prosa. *Glottometrics 16, 57-62.*

Lehfeldt, W. (2003): *Akzent und Betonung im Russischen*. München: Verlag Otto Sagner.

Marbe, K. (1904): *Über den Rhythmus der Prosa.* Giessen: J. Ricker'sche Verlagsbuchhandlung.

Strauss, U., Fan, F., & Altmann, G. (2008): *Problems in Quantitative Linguistics 1.* Lüdenscheid: RAM-Verlag.

Wimmer, G., & Altmann, G. (1996): The Theory of Word Length Distribution: Some Results and Generalizations. In Peter Schmidt (Hrsg.), *Glottometrika 15* (S. 112-133). Trier: Wissenschaftlicher Verlag Trier.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994): Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1, 98-106.*

# Hapax Legomena and Language Typology, a Case Study

*Fan Fengxiang (Dalian, China)*

## *1. INTRODUCTION*

The distributional characteristics of hapax legomena (hereafter referred to as hapax or hapaxes) have been noticed by quantitative linguists and have been used in studying textual lexical relationships (Good 1953; Holms 1991; Honore 1979; Tweedie & Baayen 1998; Baayen 2001). In two recent studies, Popescu, Mačutek and Altmann (2008) and Popescu and Altmann (2008) find that texts in a synthetic language have more hapaxes than ones in an analytic language. The reason behind this is that in a synthetic language, a word has many different forms through inflexion and derivation, but not all of them can appear in a text regardless of its length. English is one of the least synthetic languages but still one word can produce quite a number of different related words. Take the word *establish* as an example, it has 39 possible related forms (excluding those formed through compounding), but even in a mega-corpus such as the British National Corpus (BNC), only some of them occur, and two of them are hapax legomena, as shown in table one.

Taking compounding into consideration, the number of different word forms of *establish* would be still larger.

On the other hand, in an analytic language such as Chinese, a word is invariable, i.e., it does not have inflexions or derivations. Each word is a stand-alone morpheme – free morpheme, either a functional or lexical one. The English word *establish* corresponds to the Chinese word 建. However, the latter remains the same in whatever linguistic context, although it can form a large number of compounds with other words. This is why texts in Chinese have a much smaller vocabulary size (the number of different characters) and still fewer hapaxes than texts in English. For example, the *Dream of the Red Chamber*, a Chinese classic novel written in the 18$^{th}$ century, has 496408 word tokens but a vocabulary size of 4282, of which 718 are hapaxes, accounting

for 16.77 % of the vocabulary. While an English text of this length and nature would have a much larger vocabulary, over 40 % of which would be hapaxes.

Table 1
Different word forms of *establish* and their occurrences in the BNC

| | | | |
|---|---|---|---|
| established | 11989 | antidisestablishable | 0 |
| establishment | 4012 | antidisestablish | 0 |
| establishing | 2107 | unestablishes | 0 |
| establishes | 341 | unestablishers | 0 |
| disestablishment | 30 | unestablisher | 0 |
| disestablished | 13 | unestablishing | 0 |
| unestablished | 7 | unestablishable | 0 |
| antidisestablishmentarianism | 5 | unestablish | 0 |
| establishmentarian | 2 | disestablishmentarians | 0 |
| establisher | 2 | disestablishmentarianism | 0 |
| disestablish | 1 | disestablishmentarian | 0 |
| establishable | 1 | disestablishing | 0 |
| antidisestablishmentarians | 0 | disestablishes | 0 |
| antidisestablishmentarian | 0 | disestablishers | 0 |
| antidisestablishment | 0 | disestablisher | 0 |
| antidisestablishers | 0 | disestablishable | 0 |
| antidisestablisher | 0 | establishmentarians | 0 |
| antidisestablished | 0 | establishmentarianism | 0 |
| antidisestablishing | 0 | establishers | 0 |
| antidisestablishes | 0 | | |

An interesting question then ensues. In computing vocabulary size and the number of hapaxes of a text in a synthetic language, if only the free morphemes are considered, i.e., removing all the inflectional and derivational word endings, prefixes, and breaking compounds into separate free morphemes, would there still be a significant difference between a synthetic language and an analytic one as far as the number of free morphemes and hapaxes is concerned? This paper attempts to address this question.

In this paper, *word* and *word token* are used interchangeably; both refer to an alpha-numeric string separated on either side by a space for English, or a single character for Chinese. *Word type* refers to different word forms for English, or different characters for Chinese; while *word base*, used only for English, is the part of a word that

is a free morpheme (excluding compounds, which have more than one free morpheme). For example the following set of English words *establish*, *established*, *established*, *established*, *establishing*, *establisher*, *establisher*, *establishment*, *disestablish*, *disestablish* are ten words or word tokens, but six word types, all having the same word base *establish*, which is a free morpheme. The following set of Chinese words 建, 建, 不, 他, 人, 的, 的, 有, 有, 就 are ten words or word tokens but seven different word types; in this sense, the Chinese word type corresponds to the English word base.

## *2. DATA AND METHOD*

The languages we chose to study are English and Chinese. In choosing the data, the following factors were taken into consideration: text length, genre and contents. We used Dickens's *Great Expectations* and its mirror image in Chinese, translated by a Chinese scholar Luo Zhiye (2002). This way, if any significant difference results from the morpheme- and hapax-number comparison, it would be attributed to the difference in language only.

### *2.1 Tokenization*

The two versions of *Great Expectations* were first tokenized. Computerized tokenization of English texts is easy since words are separated on either side by a space, except for a few cases such as *ad hoc*, *foie gras*, *per cent* etc, which were taken care of by using a list of such words. Computerized tokenization of Chinese texts is more complicated since normally there is no space separating Chinese characters in a running text. So a two-byte-cut method for tokenization was devised and used since a Chinese word consists of two bytes. But there is a hitch in this method because Chinese texts, apart from two-byte characters, also have unprintable one-byte characters representing carriage returns (decimal ASCII code 13), end-of-line codes (decimal ASCII code 10), tabs (decimal ASCII code 9), etc. Suppose a line of Chinese words begins with a one-byte tab but a two-byte cut is made to tokenize this line, then the tab, as well as half of the following Chinese character would be cut, and the remaining part of the character with the first half of the second character would be cut too, and this

would continue until the end of the line is reached. The result would be a mess of garbled codes. To avoid this, all possible one-byte codes were removed first from the Chinese text before the two-byte-cut tokenization.

### *2.2 Stemming*

Chinese words do not need stemming since they are invariable. For English stemming proves to be quite a challenge. We devised a comparison stemming algorithm, which uses a list of prefixes, inflexional and derivational word endings, and a 42000-word dictionary. Words (non-compound words) are regarded as having a free morpheme (the base) with or without prefixes, inflectional or derivational endings. If a word contains a prefix and derivational ending, such as *incompleteness*, *in* and *ness* are both in the list of prefixes and endings and are removed, and the remaining part finds a match in the dictionary, so *incompleteness* is stemmed into *complete*. Words such as *receive*, *unify* are not stemmed, because although *re* and *un* are in the list of prefixes and endings, *ceive* and *ify* are not in the dictionary. For words such as *colony*, *colonize*, *colonization* and *colonial*, the verb form is taken as the base and the words are stemmed into it. If a word contains a prefix or an inflexional or derivational ending but its base is unique in the text, then the word is left unstemmed since this will not affect the number of free morphemes of the text.

### *2.3 Decompounding*

Compounds such as *two-byte-cut* are easy to decompose, but words such as *boyfriend*, *warship* etc are difficult to decompose automatically. The comparison algorithm is also used in compound decomposition. Take the word *boyfriend* as an example, *boy* is contained in the dictionary, and after the removal of *boy*, *friend* is also in the dictionary, so *boyfriend* is decomposed into *boy* and *friend*. But *boyish* is not, since *ish* is not in the dictionary.

A set of programs in Foxpro for tokenization (for the Chinese data), stemming and decompounding (for the English data) was written using the comparison algorithm. The accuracy is about 90 %, and manual checking was performed to weed out errors. Table 2 and Table 3 are respectively some of the tokenized Chinese words and the stemmed English words.

Table 2
Part of the tokenized Chinese words with frequencies

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 阿 | 9 | 盎 | 1 | 把 | 1417 | 罢 | 13 | 矮 | 7 | 懊 | 10 | 跋 | 2 | 拜 | 30 |
| 啊 | 177 | 凹 | 10 | 靶 | 3 | 霸 | 2 | 蔼 | 6 | 八 | 61 | 案 | 64 | 扳 | 2 |
| 哎 | 5 | 熬 | 5 | 坝 | 4 | 白 | 271 | 艾 | 12 | 巴 | 130 | 半 | 105 | 班 | 22 |
| 哀 | 10 | 翱 | 1 | 爸 | 48 | 百 | 62 | 爱 | 329 | 叭 | 5 | 伴 | 38 | 般 | 115 |
| 唉 | 10 | 鳌 | 1 | 暗 | 114 | 柏 | 17 | 碍 | 19 | 吧 | 289 | 办 | 201 | 斑 | 12 |
| 埃 | 334 | 傲 | 50 | 黯 | 2 | 摆 | 67 | 安 | 250 | 疤 | 3 | 拔 | 18 | 搬 | 18 |
| 挨 | 17 | 奥 | 78 | 肮 | 16 | 呗 | 1 | 鞍 | 57 | 笆 | 4 | 按 | 62 | 板 | 113 |
| 嗳 | 16 | 澳 | 4 | 昂 | 9 | 败 | 20 | | | | | | | | |

## *3. RESULTS*

### *3.1 Lexical description of Great Expectations*

*Great Expectations* has 184291 word tokens represented by 10983 word types. Of these word types, 4762 are hapaxes, 1777 dis legomena (words occurring twice), and 952 tris legomena (words occurring three times), respectively accounting for 43.36 %, 16 % and 9 % of the word types. However, many of the hapaxes are words with the same word base but different inflexions or derivations. For example, in the novel, the word *accuse* has the following forms: *accuse*, *accused*, *accuses*, *accusing* and *accusatory*. Of these word forms, *accused* occurs twice, but the rest all occur only once.

After stemming and decompounding, the 10984 word types were reduced to 4786 word bases, while the number of hapaxes, dis legomena and tris legomena dropped to 1534, 586 and 364 respectively, accounting for 32.05 %, 12.24 % and 7.61 % of the word bases.

In order to see the relationship between word types, word bases, hapaxes, dis legomena, tris legomena and text length, i.e., how their number changes as text length increases, the entire novel was divided into 100 chunks, each about 1843 words in length. The novel has 59 chapters; however, these chapters are not equal in length; so these natural divisions could not be used for the said purpose. These text chunks were put together one by one in a random order so as to compute the growth of word types, word bases, hapaxes, dis legomena and tris legomena. Figure 1 and Figure 2 respectively display the growth curves before and after stemming and decompounding.

Table 3
Part of the stemmed and decompounded English word bases with frequencies

| | | | | | |
|---|---|---|---|---|---|
| All | 736 | Always | 173 | Anchovy | 1 |
| Alleviate | 1 | Amateur | 4 | Ancient | 5 |
| Alley | 1 | Amaze | 3 | And | 7097 |
| Allot | 4 | Amble | 1 | Angel | 3 |
| Allow | 17 | Ambush | 1 | Anger | 14 |
| Alloy | 1 | Amelia | 1 | Angle | 2 |
| Allude | 4 | Amen | 2 | Animal | 2 |
| Ally | 3 | Amenity | 1 | Animate | 3 |
| Almanack | 1 | America | 1 | Animosity | 3 |
| Almighty | 1 | Ami | 1 | Ankle | 5 |
| Almost | 44 | Amiable | 4 | Anne | 1 |
| Alonger | 9 | Among | 77 | Announce | 8 |
| Alphabet | 5 | Amost | 5 | Annoy | 1 |
| Already | 39 | Amount | 5 | Annual | 2 |
| Also | 35 | Amphibious | 3 | Annum | 1 |
| Altar | 1 | Ample | 4 | Anonymous | 2 |
| Alter | 4 | Amuse | 4 | Another | 163 |
| Alternate | 1 | Anatomy | 1 | Answer | 111 |
| Although | 21 | Ancestor | 2 | Ant | 9 |
| Altogether | 17 | Anchor | 3 | | |

Before stemming and decompounding, the mean number of word types per 1843 words is 633.72, and the mean word type/word token ratio is 0.3435. The mean number of hapaxes per 1843 words is 415.34. The mean hapax/word type ratio is 0.6554.

After stemming and decompounding, the mean number of word bases per 1843 words is 493.89, while that of hapaxes is 278.94. The mean word base/word token ratio per 1843 words is 0.268, while the mean hapax/word base ratio is 0.5648.

### *3.2 Lexical description of the Chinese version*

The Chinese version of *Great Expectations* contains 323634 words represented by 3219 word types. Of these word types, 402 are hapaxes, 224 dis legomena and 172 tris legomena, accounting for 12.49 %, 6.96 % and 5.34 % of the word types respectively. The number of word to-

kens is much larger than that of the English version, but the rest are all much smaller than their English counterparts.

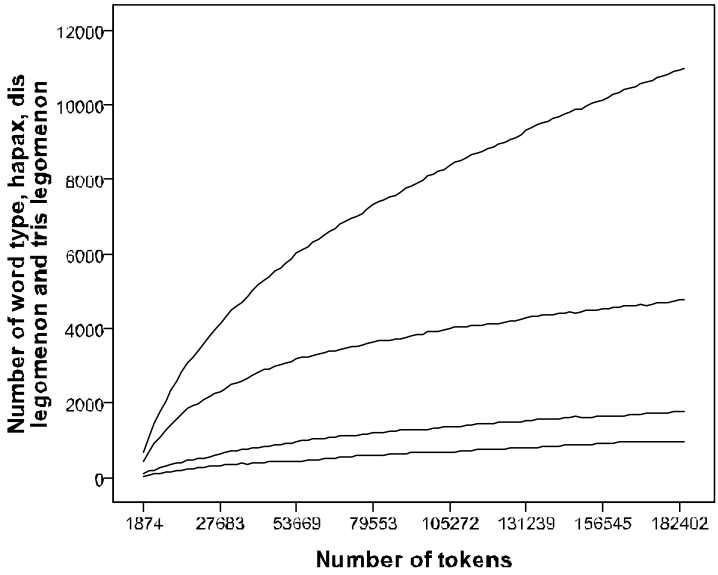



Fig.1 Growth curves of word types, hapaxes, dis legomena and tris legomena (From top down) before stemming and decompounding

As with the English version, the Chinese version was divided into 176 chunks, each about 1839 words in length. The text chunks were put together one by one in a random order and the growth of word types, hapaxes, dis legomena and tris legomena was computed. Figure 3 shows these growth curves.

The mean number of word types and hapaxes per 1839 words is respectively 525.91 and 254.96. The mean word type/word token ratio per 1839 words is 0.286, and the mean hapax/word type ratio per 1839 words is 0.4848.

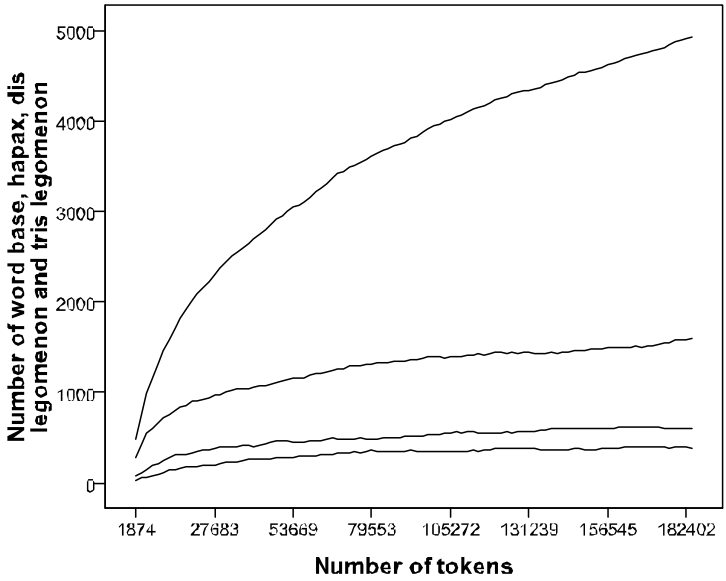



Fig. 2 Growth curves of word bases, hapaxes, dis legomena and tris legomena (From top down) after stemming and decompounding

### *3.3 Comparison*

The mean per chunk word type number of the Chinese version (525.91) is larger than the mean per chunk word base number of the English version (493.89). The mean per chunk hapax number of the Chinese version (254.96) is smaller than that of the English version (278.94). To check whether these differences are statistically significant, a *t*-test was performed since it is very robust and does not have strict requirement on the shape of the data distribution when the sample number is large. The results show that these differences are highly significant.

However, when we consider the two versions in their entirety, the differences in the number of word bases/word types and hapaxes are very striking. In the Chinese version, 323634 word tokens produce 3219 word types, of which only 402 are hapaxes; while in the English version 184291 word tokens produce 4786 word bases, of which 1534 are hapaxes. With such great differences, no statistical tests for significance are necessary.

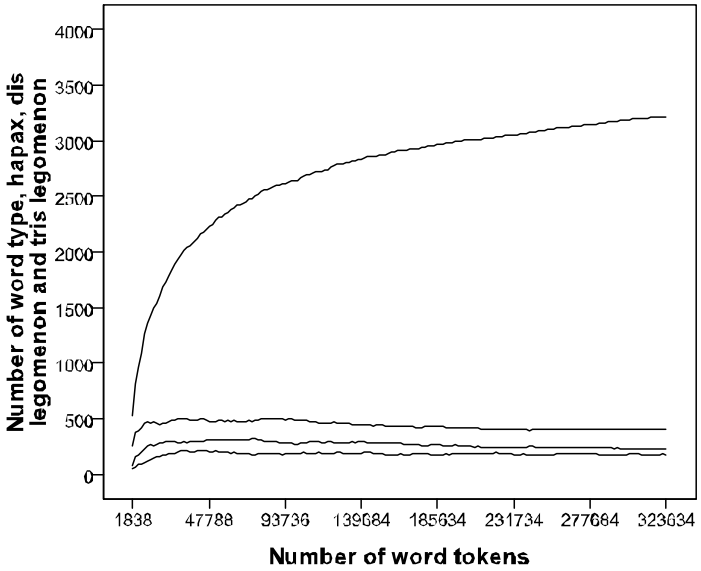



Fig. 3 Growth curves of word types, hapaxes, dis legomena and tris legomena (from top down) at an interval of about 1839 words

Table 4
*t*-test results

| | Levene's Test for Equality of Variances | | t-test for Equality of Means | | |
|---|---|---|---|---|---|
| | F | Sig. | t | df | Sig. (2-tailed) |
| Base/type | 2.170 | .142 | -7.043 | 274 | .000 |
| Hapaxes | .880 | .349 | 6.284 | 274 | .000 |

As shown in Figure 4, in the Chinese version, the number of hapaxes, dis legomena and tris legomena starts to drop after the number of word tokens reaches 47788, while that of the hapaxes, dis legomena and tris legomena of the English version keeps rising, as shown in Figure 5.

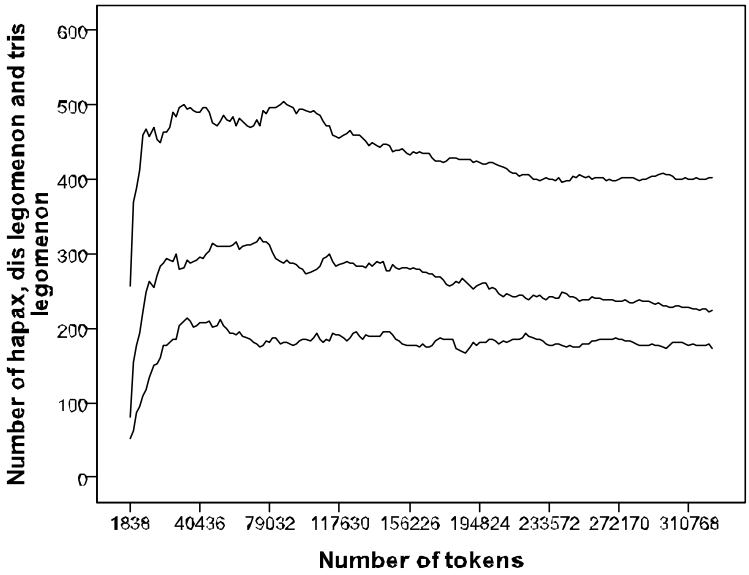



Fig. 4 Growth curves of hapaxes, dis legomena and tris legomena (from top down) of the Chinese version

## *4. CONCLUSION*

By returning words of the English version to free morphemes, the number of word types and hapaxes were sharply reduced. The number of free morphemes per chunk is smaller than that of the Chinese version, while the number of hapaxes per chunk is larger. Taking the two versions in their entirety, both the number of free morphemes and hapaxes of the Chinese version is considerably smaller than that of the English version. Part of the reason is that the Chinese language almost always uses compounding in word formation. For example, both in English and Chinese there are free morphemes for *pig*, *meat*, *fly*, *machine*, *two*, *ten*, etc, but in English there are also words such as *pork*, *plane*, *twenty*, while in Chinese *pig meat*, *fly machine*, *two ten* are used. This results in much fewer free morphemes in Chinese. The results of the study show that the number of hapaxes can be used as a typological marker for the Chinese language provided that the length of the text under examination is long enough.

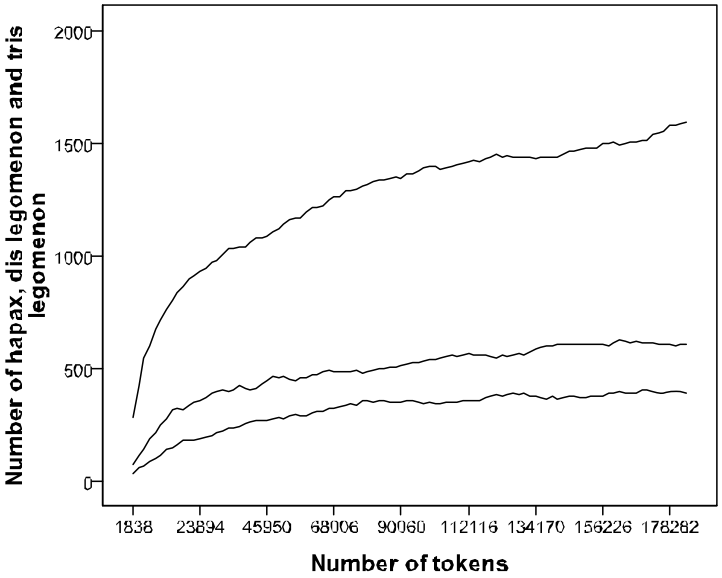



Fig. 5 Growth curves of hapaxes, dis legomena and tris legomena (from top down) of the English version

## *REFERENCES*

Baayen, H. (2001): *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Good, J. (1953): The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrica*, 40, 237-264.

Holmes, D. (1991): Vocabulary Richness and the Prophetic Voice, *Literary & Linguistic Computing* 6, 259-268.

Honoré, A. (1979): Some simple measures of richness of vocabulary. *Association of Literary and Linguistic Computing Bulletin, 7, 2,* 172-179.

Luo, Zhiye (2002): *Yuan Da Qian Cheng* (Chinese translation of *Great Expectations*), Nanjing: the Yilin Translation Press.

Popescu, I-I.; Mačutek, J.; Altmann, G. (2008): Word frequency and arc length. *Glottometrics* 17, 18-42.

Popescu, I.; Mačutek, J.: Altmann, G. (2008): Hapax legomena and language typology. *Journal of Quantitative Linguistics* 15 (4), 370-378.

Tweedie, J. & Baayen, H. (1998): How variable may a constant be? Measure of Lexical richness in Perspective. *Computer and the Humanities*, 32, 323-52.

*APPENDIX*

**1. Lexical statistics of Great Expectations (English version, before stemming and decompounding)[1]**

Great Expectations were divided into 100 text chunks, each about 1843 words in length.

| TOKENS | CLEN | CTYPE | CHAP | CDIS | CTRIS | GTYPE | GHAP | GDIS | GTRIS |
|---|---|---|---|---|---|---|---|---|---|
| 1874 | 1874 | 636 | 418 | 93 | 28 | 636 | 418 | 93 | 28 |
| 3742 | 1868 | 623 | 407 | 83 | 38 | 1037 | 667 | 137 | 72 |
| 5571 | 1829 | 643 | 433 | 77 | 40 | 1399 | 908 | 185 | 75 |
| 7412 | 1841 | 607 | 398 | 80 | 35 | 1690 | 1071 | 244 | 85 |
| 9298 | 1886 | 661 | 428 | 104 | 35 | 1978 | 1236 | 288 | 112 |
| 11118 | 1820 | 610 | 404 | 91 | 21 | 2210 | 1371 | 325 | 128 |
| 13007 | 1889 | 602 | 379 | 86 | 32 | 2446 | 1489 | 372 | 154 |
| 14855 | 1848 | 673 | 467 | 91 | 36 | 2696 | 1625 | 425 | 167 |
| 16678 | 1823 | 646 | 425 | 98 | 30 | 2924 | 1750 | 446 | 190 |
| 18552 | 1874 | 595 | 364 | 84 | 47 | 3106 | 1833 | 464 | 210 |
| 20426 | 1874 | 657 | 441 | 85 | 36 | 3319 | 1946 | 508 | 205 |
| 22213 | 1787 | 585 | 358 | 96 | 40 | 3488 | 2045 | 524 | 208 |
| 24044 | 1831 | 584 | 346 | 104 | 38 | 3640 | 2101 | 567 | 216 |
| 25918 | 1874 | 634 | 403 | 106 | 33 | 3782 | 2144 | 620 | 228 |
| 27683 | 1765 | 682 | 467 | 98 | 36 | 3980 | 2221 | 685 | 244 |
| 29474 | 1791 | 634 | 423 | 81 | 38 | 4131 | 2291 | 712 | 250 |
| 31350 | 1876 | 649 | 440 | 89 | 28 | 4289 | 2356 | 745 | 274 |
| 33272 | 1922 | 650 | 423 | 102 | 41 | 4434 | 2420 | 757 | 304 |
| 35137 | 1865 | 621 | 388 | 97 | 38 | 4531 | 2448 | 755 | 336 |
| 37055 | 1918 | 599 | 371 | 87 | 40 | 4647 | 2487 | 783 | 349 |
| 38901 | 1846 | 662 | 444 | 87 | 42 | 4775 | 2527 | 799 | 385 |
| 40704 | 1803 | 621 | 410 | 78 | 38 | 4911 | 2580 | 823 | 395 |
| 42543 | 1839 | 647 | 425 | 96 | 48 | 5057 | 2656 | 843 | 401 |
| 44530 | 1987 | 602 | 378 | 87 | 45 | 5182 | 2709 | 859 | 410 |
| 46299 | 1769 | 652 | 447 | 96 | 32 | 5340 | 2775 | 896 | 431 |
| 48161 | 1862 | 629 | 403 | 95 | 37 | 5478 | 2844 | 922 | 438 |

[1] Tokens: Cumulative number of tokens; CLEN: Number of word tokens per text chunk; CTYPE: Number of word types per text chunk; CHAP: Number of hapxes per text chunk; CDIS: Number of dis legomena per text chunk; CTRIS: Number of tris legomena per text chunk; GTYPE: Cumulative number of word types; GHAP: Cumulative number of hapaxes; GDIS: Cumulative number of dis legomena.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 50063 | 1902 | 602 | 378 | 79 | 43 | 5561 | 2877 | 921 | 455 |
| 51867 | 1804 | 598 | 391 | 81 | 47 | 5682 | 2928 | 945 | 466 |
| 53672 | 1805 | 629 | 412 | 83 | 40 | 5791 | 2970 | 950 | 495 |
| 55453 | 1781 | 597 | 367 | 109 | 32 | 5878 | 3001 | 964 | 498 |
| 57359 | 1906 | 605 | 376 | 84 | 49 | 5973 | 3031 | 984 | 502 |
| 59197 | 1838 | 652 | 450 | 73 | 41 | 6095 | 3090 | 1005 | 506 |
| 60990 | 1793 | 712 | 504 | 95 | 37 | 6249 | 3151 | 1043 | 532 |
| 62830 | 1840 | 635 | 414 | 90 | 45 | 6354 | 3204 | 1051 | 535 |
| 64628 | 1798 | 682 | 479 | 82 | 39 | 6481 | 3268 | 1064 | 545 |
| 66444 | 1816 | 707 | 499 | 98 | 30 | 6627 | 3347 | 1076 | 560 |
| 68241 | 1797 | 685 | 479 | 90 | 37 | 6756 | 3428 | 1074 | 573 |
| 70190 | 1949 | 604 | 367 | 98 | 44 | 6819 | 3438 | 1083 | 580 |
| 72121 | 1931 | 620 | 389 | 95 | 34 | 6915 | 3478 | 1090 | 592 |
| 73989 | 1868 | 589 | 357 | 90 | 35 | 6981 | 3492 | 1110 | 591 |
| 75736 | 1747 | 680 | 461 | 101 | 35 | 7082 | 3535 | 1117 | 601 |
| 77661 | 1925 | 607 | 381 | 81 | 47 | 7154 | 3559 | 1119 | 617 |
| 79556 | 1895 | 643 | 419 | 82 | 45 | 7264 | 3612 | 1147 | 620 |
| 81384 | 1828 | 658 | 454 | 91 | 37 | 7368 | 3644 | 1190 | 602 |
| 83215 | 1831 | 604 | 403 | 74 | 34 | 7443 | 3661 | 1213 | 613 |
| 85099 | 1884 | 621 | 380 | 101 | 47 | 7516 | 3690 | 1226 | 607 |
| 86939 | 1840 | 651 | 441 | 92 | 29 | 7610 | 3730 | 1240 | 621 |
| 88746 | 1807 | 641 | 416 | 94 | 34 | 7688 | 3769 | 1244 | 624 |
| 90566 | 1820 | 616 | 398 | 82 | 44 | 7743 | 3772 | 1274 | 615 |
| 92427 | 1861 | 685 | 475 | 88 | 47 | 7854 | 3823 | 1295 | 611 |
| 94350 | 1923 | 626 | 409 | 81 | 36 | 7912 | 3842 | 1291 | 629 |
| 96161 | 1811 | 668 | 457 | 92 | 36 | 8006 | 3884 | 1300 | 636 |
| 97962 | 1801 | 689 | 483 | 70 | 46 | 8119 | 3937 | 1313 | 639 |
| 99805 | 1843 | 642 | 437 | 92 | 26 | 8198 | 3968 | 1316 | 657 |
| 101622 | 1817 | 687 | 472 | 93 | 39 | 8281 | 3990 | 1341 | 660 |
| 103409 | 1787 | 650 | 435 | 86 | 41 | 8363 | 4003 | 1366 | 668 |
| 105277 | 1868 | 634 | 416 | 78 | 39 | 8431 | 3999 | 1399 | 673 |
| 107078 | 1801 | 689 | 471 | 96 | 37 | 8526 | 4027 | 1434 | 678 |
| 108966 | 1888 | 576 | 362 | 73 | 42 | 8569 | 4028 | 1450 | 678 |
| 110791 | 1825 | 650 | 421 | 98 | 38 | 8662 | 4079 | 1452 | 693 |
| 112589 | 1798 | 638 | 420 | 87 | 35 | 8732 | 4103 | 1466 | 700 |
| 114443 | 1854 | 542 | 322 | 82 | 29 | 8775 | 4116 | 1467 | 703 |
| 116350 | 1907 | 601 | 373 | 104 | 31 | 8830 | 4131 | 1481 | 694 |
| 118124 | 1774 | 634 | 433 | 80 | 32 | 8914 | 4170 | 1479 | 709 |
| 119999 | 1875 | 640 | 416 | 93 | 49 | 8987 | 4193 | 1487 | 718 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 121895 | 1896 | 599 | 375 | 76 | 47 | 9045 | 4209 | 1491 | 733 |
| 123797 | 1902 | 625 | 405 | 76 | 56 | 9091 | 4208 | 1518 | 731 |
| 125534 | 1737 | 719 | 508 | 106 | 35 | 9190 | 4237 | 1531 | 755 |
| 127487 | 1953 | 610 | 378 | 91 | 44 | 9242 | 4248 | 1548 | 756 |
| 129344 | 1857 | 590 | 366 | 91 | 42 | 9291 | 4254 | 1556 | 770 |
| 131245 | 1901 | 615 | 392 | 90 | 33 | 9339 | 4265 | 1564 | 784 |
| 133069 | 1824 | 613 | 387 | 105 | 33 | 9394 | 4272 | 1590 | 785 |
| 134974 | 1905 | 644 | 416 | 88 | 41 | 9469 | 4301 | 1607 | 792 |
| 136829 | 1855 | 612 | 393 | 76 | 48 | 9526 | 4316 | 1613 | 804 |
| 138640 | 1811 | 662 | 446 | 90 | 35 | 9586 | 4330 | 1620 | 814 |
| 140611 | 1971 | 626 | 385 | 103 | 43 | 9626 | 4323 | 1636 | 823 |
| 142456 | 1845 | 600 | 381 | 82 | 45 | 9674 | 4335 | 1643 | 832 |
| 144286 | 1830 | 689 | 474 | 92 | 43 | 9752 | 4363 | 1651 | 849 |
| 146053 | 1767 | 664 | 464 | 87 | 35 | 9815 | 4390 | 1638 | 874 |
| 147792 | 1739 | 637 | 435 | 88 | 29 | 9909 | 4426 | 1663 | 877 |
| 149719 | 1927 | 577 | 351 | 75 | 52 | 9947 | 4431 | 1671 | 871 |
| 151523 | 1804 | 605 | 387 | 93 | 38 | 9999 | 4450 | 1675 | 868 |
| 152829 | 1306 | 445 | 274 | 69 | 26 | 10029 | 4458 | 1676 | 878 |
| 154659 | 1830 | 655 | 445 | 81 | 41 | 10086 | 4465 | 1689 | 889 |
| 156553 | 1894 | 634 | 389 | 104 | 46 | 10154 | 4505 | 1686 | 886 |
| 158462 | 1909 | 593 | 375 | 93 | 33 | 10184 | 4498 | 1699 | 887 |
| 160258 | 1796 | 622 | 413 | 86 | 37 | 10258 | 4529 | 1714 | 890 |
| 162123 | 1865 | 658 | 429 | 98 | 39 | 10324 | 4542 | 1728 | 903 |
| 163954 | 1831 | 667 | 469 | 74 | 35 | 10385 | 4561 | 1727 | 912 |
| 165787 | 1833 | 633 | 420 | 91 | 36 | 10425 | 4565 | 1733 | 919 |
| 167619 | 1832 | 660 | 430 | 107 | 42 | 10478 | 4581 | 1740 | 919 |
| 169446 | 1827 | 643 | 418 | 107 | 31 | 10537 | 4599 | 1748 | 930 |
| 171304 | 1858 | 620 | 391 | 86 | 47 | 10591 | 4621 | 1749 | 938 |
| 173065 | 1761 | 683 | 474 | 98 | 31 | 10670 | 4659 | 1752 | 948 |
| 174911 | 1846 | 681 | 447 | 93 | 53 | 10726 | 4677 | 1752 | 960 |
| 176721 | 1810 | 668 | 470 | 74 | 38 | 10787 | 4704 | 1750 | 954 |
| 178581 | 1860 | 632 | 413 | 83 | 39 | 10860 | 4739 | 1761 | 956 |
| 180489 | 1908 | 653 | 423 | 87 | 52 | 10898 | 4741 | 1761 | 958 |
| 182411 | 1922 | 634 | 418 | 91 | 29 | 10941 | 4750 | 1770 | 960 |
| 184291 | 1880 | 611 | 386 | 102 | 37 | 10985 | 4764 | 1777 | 952 |

**2. Lexical statistics of Great Expectations (English version, after stemming and decompounding)**

Great Expectations were divided into 100 text chunks, each about 1843 words in length.

Total number of text chunks: 100

| TOKENS | CLEN | CMORPH[2] | CHAP | CDIS | CTRIS | GMORPH | GHAP | GDIS | GTRIS |
|---|---|---|---|---|---|---|---|---|---|
| 1874 | 1874 | 484 | 278 | 76 | 28 | 484 | 278 | 76 | 28 |
| 3742 | 1868 | 483 | 263 | 84 | 31 | 758 | 412 | 117 | 61 |
| 5571 | 1829 | 500 | 293 | 71 | 31 | 980 | 532 | 139 | 68 |
| 7412 | 1841 | 488 | 283 | 73 | 36 | 1154 | 589 | 188 | 86 |
| 9298 | 1886 | 526 | 303 | 77 | 43 | 1317 | 657 | 212 | 102 |
| 11118 | 1820 | 469 | 275 | 71 | 27 | 1440 | 702 | 244 | 116 |
| 13007 | 1889 | 458 | 247 | 64 | 34 | 1566 | 740 | 270 | 140 |
| 14855 | 1848 | 532 | 319 | 86 | 35 | 1693 | 780 | 309 | 144 |
| 16678 | 1823 | 514 | 289 | 87 | 34 | 1801 | 812 | 316 | 160 |
| 18552 | 1874 | 461 | 243 | 58 | 39 | 1884 | 834 | 313 | 177 |
| 20426 | 1874 | 519 | 311 | 68 | 35 | 1987 | 869 | 331 | 176 |
| 22213 | 1787 | 439 | 227 | 68 | 47 | 2050 | 876 | 348 | 178 |
| 24044 | 1831 | 442 | 221 | 77 | 31 | 2117 | 893 | 354 | 187 |
| 25918 | 1874 | 483 | 250 | 96 | 35 | 2176 | 907 | 360 | 197 |
| 27683 | 1765 | 527 | 308 | 81 | 45 | 2270 | 940 | 373 | 202 |
| 29474 | 1791 | 479 | 268 | 72 | 33 | 2322 | 946 | 380 | 210 |
| 31350 | 1876 | 513 | 297 | 83 | 31 | 2396 | 972 | 386 | 221 |
| 33272 | 1922 | 510 | 278 | 93 | 40 | 2463 | 997 | 381 | 233 |
| 35137 | 1865 | 471 | 242 | 86 | 35 | 2500 | 1000 | 388 | 230 |
| 37055 | 1918 | 465 | 251 | 64 | 38 | 2543 | 1000 | 404 | 238 |
| 38901 | 1846 | 519 | 295 | 75 | 47 | 2586 | 1002 | 397 | 251 |
| 40704 | 1803 | 486 | 276 | 71 | 33 | 2637 | 1019 | 393 | 253 |
| 42543 | 1839 | 499 | 267 | 94 | 45 | 2687 | 1031 | 398 | 257 |
| 44530 | 1987 | 483 | 253 | 93 | 37 | 2738 | 1036 | 408 | 257 |
| 46299 | 1769 | 497 | 291 | 86 | 32 | 2789 | 1042 | 426 | 253 |
| 48161 | 1862 | 485 | 261 | 88 | 29 | 2848 | 1060 | 446 | 258 |
| 50063 | 1902 | 472 | 243 | 79 | 49 | 2883 | 1071 | 443 | 264 |
| 51867 | 1804 | 460 | 260 | 71 | 37 | 2927 | 1090 | 446 | 265 |
| 53672 | 1805 | 486 | 266 | 84 | 38 | 2969 | 1107 | 437 | 273 |
| 55453 | 1781 | 436 | 219 | 90 | 33 | 2994 | 1112 | 433 | 280 |

[2]CMORPH: Number of free morphemes per text chunk; GMORPH: Cumulative number of free morphemes; GTRIS: Cumulative number of tris legomena.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 57359 | 1906 | 496 | 283 | 65 | 47 | 3026 | 1112 | 444 | 277 |
| 59197 | 1838 | 502 | 298 | 64 | 38 | 3077 | 1137 | 448 | 277 |
| 60990 | 1793 | 559 | 344 | 89 | 35 | 3142 | 1155 | 460 | 287 |
| 62830 | 1840 | 486 | 273 | 76 | 36 | 3176 | 1157 | 464 | 298 |
| 64628 | 1798 | 517 | 302 | 93 | 37 | 3219 | 1167 | 473 | 296 |
| 66444 | 1816 | 564 | 346 | 87 | 37 | 3286 | 1189 | 483 | 305 |
| 68241 | 1797 | 529 | 310 | 92 | 38 | 3330 | 1206 | 478 | 309 |
| 70190 | 1949 | 466 | 241 | 89 | 35 | 3348 | 1203 | 472 | 320 |
| 72121 | 1931 | 478 | 259 | 77 | 35 | 3398 | 1230 | 472 | 324 |
| 73989 | 1868 | 454 | 233 | 79 | 27 | 3418 | 1231 | 472 | 326 |
| 75736 | 1747 | 510 | 280 | 100 | 36 | 3453 | 1239 | 480 | 327 |
| 77661 | 1925 | 481 | 256 | 72 | 46 | 3479 | 1252 | 462 | 344 |
| 79556 | 1895 | 497 | 278 | 70 | 40 | 3509 | 1256 | 471 | 345 |
| 81384 | 1828 | 515 | 299 | 92 | 31 | 3555 | 1275 | 473 | 341 |
| 83215 | 1831 | 462 | 265 | 63 | 29 | 3582 | 1278 | 479 | 346 |
| 85099 | 1884 | 484 | 245 | 92 | 43 | 3603 | 1277 | 484 | 338 |
| 86939 | 1840 | 505 | 278 | 83 | 47 | 3636 | 1289 | 487 | 334 |
| 88746 | 1807 | 473 | 243 | 75 | 41 | 3654 | 1292 | 491 | 334 |
| 90566 | 1820 | 464 | 249 | 71 | 40 | 3668 | 1290 | 493 | 337 |
| 92427 | 1861 | 544 | 330 | 76 | 40 | 3717 | 1308 | 500 | 342 |
| 94350 | 1923 | 483 | 267 | 70 | 40 | 3734 | 1309 | 506 | 339 |
| 96161 | 1811 | 519 | 311 | 71 | 39 | 3771 | 1324 | 506 | 332 |
| 97962 | 1801 | 533 | 313 | 70 | 41 | 3811 | 1337 | 512 | 329 |
| 99805 | 1843 | 504 | 298 | 75 | 33 | 3839 | 1346 | 514 | 332 |
| 101622 | 1817 | 521 | 301 | 86 | 36 | 3864 | 1347 | 520 | 322 |
| 103409 | 1787 | 499 | 292 | 72 | 31 | 3885 | 1337 | 528 | 321 |
| 105277 | 1868 | 492 | 275 | 72 | 35 | 3910 | 1338 | 533 | 323 |
| 107078 | 1801 | 543 | 324 | 90 | 33 | 3945 | 1345 | 543 | 325 |
| 108966 | 1888 | 422 | 215 | 70 | 25 | 3958 | 1348 | 535 | 329 |
| 110791 | 1825 | 508 | 278 | 91 | 35 | 3994 | 1358 | 541 | 333 |
| 112589 | 1798 | 493 | 275 | 75 | 32 | 4020 | 1363 | 548 | 334 |
| 114443 | 1854 | 419 | 213 | 70 | 35 | 4033 | 1369 | 541 | 332 |
| 116350 | 1907 | 475 | 256 | 84 | 33 | 4048 | 1366 | 540 | 334 |
| 118124 | 1774 | 495 | 286 | 73 | 34 | 4085 | 1377 | 540 | 344 |
| 119999 | 1875 | 503 | 289 | 77 | 38 | 4112 | 1385 | 536 | 353 |
| 121895 | 1896 | 473 | 256 | 84 | 28 | 4136 | 1397 | 531 | 356 |
| 123797 | 1902 | 486 | 262 | 76 | 42 | 4145 | 1387 | 541 | 352 |
| 125534 | 1737 | 557 | 344 | 84 | 49 | 4179 | 1392 | 537 | 360 |
| 127487 | 1953 | 479 | 251 | 81 | 42 | 4194 | 1386 | 540 | 362 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 129344 | 1857 | 463 | 247 | 87 | 35 | 4212 | 1389 | 543 | 361 |
| 131245 | 1901 | 481 | 261 | 83 | 36 | 4221 | 1389 | 540 | 366 |
| 133069 | 1824 | 470 | 254 | 80 | 31 | 4238 | 1388 | 549 | 358 |
| 134974 | 1905 | 493 | 261 | 81 | 35 | 4257 | 1385 | 563 | 351 |
| 136829 | 1855 | 486 | 274 | 66 | 45 | 4279 | 1388 | 570 | 346 |
| 138640 | 1811 | 508 | 285 | 82 | 36 | 4302 | 1391 | 579 | 343 |
| 140611 | 1971 | 481 | 253 | 88 | 38 | 4312 | 1388 | 576 | 352 |
| 142456 | 1845 | 469 | 260 | 76 | 40 | 4327 | 1392 | 581 | 345 |
| 144286 | 1830 | 542 | 327 | 70 | 46 | 4361 | 1405 | 584 | 350 |
| 146053 | 1767 | 512 | 304 | 75 | 38 | 4378 | 1410 | 579 | 355 |
| 147792 | 1739 | 492 | 293 | 74 | 27 | 4402 | 1412 | 583 | 357 |
| 149719 | 1927 | 452 | 231 | 73 | 43 | 4412 | 1414 | 584 | 347 |
| 151523 | 1804 | 460 | 258 | 70 | 34 | 4429 | 1421 | 580 | 350 |
| 152829 | 1306 | 354 | 196 | 56 | 22 | 4438 | 1421 | 580 | 353 |
| 154659 | 1830 | 515 | 304 | 78 | 33 | 4462 | 1425 | 583 | 357 |
| 156553 | 1894 | 503 | 272 | 81 | 42 | 4496 | 1445 | 582 | 359 |
| 158462 | 1909 | 459 | 245 | 85 | 33 | 4505 | 1445 | 575 | 368 |
| 160258 | 1796 | 475 | 269 | 73 | 39 | 4530 | 1451 | 585 | 367 |
| 162123 | 1865 | 519 | 299 | 80 | 36 | 4553 | 1442 | 600 | 375 |
| 163954 | 1831 | 534 | 330 | 73 | 33 | 4572 | 1449 | 595 | 369 |
| 165787 | 1833 | 495 | 287 | 73 | 39 | 4582 | 1454 | 591 | 366 |
| 167619 | 1832 | 516 | 287 | 86 | 47 | 4598 | 1451 | 598 | 367 |
| 169446 | 1827 | 504 | 287 | 89 | 37 | 4617 | 1457 | 593 | 374 |
| 171304 | 1858 | 479 | 246 | 88 | 41 | 4629 | 1458 | 590 | 377 |
| 173065 | 1761 | 524 | 312 | 88 | 31 | 4661 | 1481 | 591 | 371 |
| 174911 | 1846 | 537 | 294 | 93 | 50 | 4677 | 1485 | 590 | 366 |
| 176721 | 1810 | 526 | 323 | 72 | 37 | 4699 | 1494 | 587 | 364 |
| 178581 | 1860 | 499 | 284 | 76 | 31 | 4743 | 1519 | 587 | 370 |
| 180489 | 1908 | 517 | 289 | 83 | 39 | 4755 | 1523 | 581 | 371 |
| 182411 | 1922 | 491 | 282 | 73 | 27 | 4771 | 1527 | 586 | 368 |
| 184291 | 1880 | 467 | 240 | 90 | 41 | 4786 | 1534 | 586 | 364 |

### 3. Lexical statistics of Great Expectations (Chinese version)

The Chinese version was divided into 176 text chunks, each about 1839 words in length.

Total number of text chunks: 176

| TOKENS | CLEN | CMOR | PH CHAP | CDIS | CTRIS | GMORPH | GHAP | GDIS | GTRIS |
|---|---|---|---|---|---|---|---|---|---|
| 1838 | 1838 | 522 | 257 | 80 | 51 | 522 | 257 | 80 | 51 |
| 3676 | 1838 | 559 | 291 | 104 | 46 | 817 | 370 | 155 | 62 |
| 5514 | 1838 | 520 | 255 | 87 | 46 | 967 | 387 | 177 | 86 |
| 7352 | 1838 | 479 | 234 | 81 | 44 | 1091 | 413 | 194 | 95 |
| 9190 | 1838 | 605 | 304 | 121 | 63 | 1258 | 460 | 221 | 110 |
| 11028 | 1838 | 527 | 246 | 101 | 57 | 1362 | 467 | 249 | 118 |
| 12866 | 1838 | 526 | 263 | 91 | 48 | 1429 | 458 | 263 | 134 |
| 14704 | 1838 | 505 | 251 | 86 | 42 | 1500 | 469 | 254 | 150 |
| 16542 | 1838 | 462 | 211 | 81 | 47 | 1538 | 454 | 270 | 152 |
| 18380 | 1838 | 527 | 255 | 102 | 47 | 1598 | 448 | 283 | 161 |
| 20218 | 1838 | 571 | 289 | 106 | 52 | 1682 | 464 | 287 | 176 |
| 22056 | 1838 | 514 | 264 | 84 | 42 | 1723 | 463 | 294 | 177 |
| 23894 | 1838 | 481 | 215 | 95 | 47 | 1778 | 470 | 291 | 180 |
| 25732 | 1838 | 564 | 286 | 106 | 41 | 1840 | 490 | 290 | 184 |
| 27570 | 1838 | 562 | 274 | 118 | 48 | 1889 | 484 | 299 | 186 |
| 29408 | 1838 | 559 | 257 | 122 | 65 | 1933 | 497 | 280 | 204 |
| 31246 | 1838 | 565 | 294 | 86 | 49 | 1975 | 501 | 281 | 209 |
| 33084 | 1838 | 528 | 255 | 95 | 54 | 2010 | 493 | 291 | 214 |
| 34922 | 1838 | 522 | 249 | 98 | 48 | 2040 | 496 | 287 | 210 |
| 36760 | 1838 | 511 | 264 | 78 | 38 | 2058 | 491 | 289 | 201 |
| 38598 | 1838 | 545 | 265 | 97 | 55 | 2092 | 490 | 291 | 203 |
| 40436 | 1838 | 526 | 247 | 110 | 44 | 2126 | 489 | 296 | 208 |
| 42274 | 1838 | 514 | 255 | 86 | 55 | 2158 | 497 | 294 | 207 |
| 44112 | 1838 | 534 | 267 | 100 | 48 | 2181 | 497 | 299 | 208 |
| 45950 | 1838 | 528 | 247 | 106 | 53 | 2205 | 489 | 303 | 210 |
| 47788 | 1838 | 515 | 261 | 79 | 51 | 2227 | 476 | 314 | 201 |
| 49626 | 1838 | 530 | 265 | 101 | 45 | 2243 | 471 | 309 | 203 |
| 51464 | 1838 | 552 | 278 | 105 | 52 | 2280 | 478 | 309 | 211 |
| 53302 | 1838 | 560 | 274 | 106 | 54 | 2309 | 486 | 309 | 203 |
| 55140 | 1838 | 492 | 230 | 83 | 47 | 2319 | 480 | 309 | 199 |
| 56978 | 1838 | 567 | 289 | 111 | 50 | 2341 | 478 | 310 | 193 |
| 58816 | 1838 | 450 | 196 | 88 | 45 | 2353 | 483 | 311 | 193 |
| 60654 | 1838 | 595 | 309 | 108 | 54 | 2376 | 472 | 315 | 192 |
| 62492 | 1838 | 521 | 254 | 97 | 38 | 2398 | 481 | 306 | 195 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 64330 | 1838 | 630 | 344 | 114 | 56 | 2419 | 477 | 309 | 190 |
| 66168 | 1838 | 460 | 170 | 109 | 41 | 2427 | 472 | 312 | 187 |
| 68006 | 1838 | 479 | 231 | 82 | 46 | 2436 | 469 | 312 | 185 |
| 69844 | 1838 | 482 | 206 | 91 | 49 | 2450 | 472 | 314 | 180 |
| 71682 | 1838 | 522 | 244 | 101 | 48 | 2473 | 480 | 315 | 179 |
| 73520 | 1838 | 444 | 208 | 68 | 40 | 2478 | 472 | 323 | 174 |
| 75358 | 1838 | 638 | 339 | 109 | 68 | 2511 | 491 | 316 | 177 |
| 77196 | 1838 | 467 | 217 | 90 | 36 | 2522 | 488 | 315 | 183 |
| 79032 | 1836 | 554 | 257 | 111 | 59 | 2540 | 496 | 311 | 180 |
| 80870 | 1838 | 574 | 282 | 117 | 48 | 2555 | 497 | 301 | 187 |
| 82708 | 1838 | 562 | 278 | 106 | 50 | 2565 | 496 | 294 | 187 |
| 84546 | 1838 | 540 | 266 | 97 | 49 | 2575 | 500 | 290 | 179 |
| 86384 | 1838 | 476 | 207 | 87 | 61 | 2590 | 504 | 288 | 180 |
| 88222 | 1838 | 531 | 261 | 101 | 56 | 2602 | 501 | 291 | 180 |
| 90060 | 1838 | 491 | 221 | 90 | 44 | 2602 | 499 | 287 | 179 |
| 91898 | 1838 | 523 | 243 | 96 | 56 | 2607 | 496 | 287 | 177 |
| 93736 | 1838 | 473 | 214 | 91 | 40 | 2612 | 487 | 286 | 179 |
| 95574 | 1838 | 508 | 254 | 88 | 36 | 2626 | 493 | 282 | 182 |
| 97412 | 1838 | 526 | 252 | 102 | 39 | 2636 | 493 | 279 | 184 |
| 99250 | 1838 | 442 | 181 | 85 | 44 | 2637 | 492 | 272 | 184 |
| 101088 | 1838 | 527 | 272 | 94 | 44 | 2647 | 489 | 274 | 183 |
| 102926 | 1838 | 539 | 255 | 110 | 56 | 2667 | 492 | 278 | 187 |
| 104764 | 1838 | 498 | 237 | 106 | 40 | 2680 | 487 | 280 | 194 |
| 106602 | 1838 | 518 | 244 | 106 | 54 | 2689 | 485 | 284 | 184 |
| 108440 | 1838 | 517 | 243 | 90 | 57 | 2702 | 477 | 293 | 180 |
| 110278 | 1838 | 550 | 269 | 111 | 52 | 2711 | 472 | 295 | 184 |
| 112116 | 1838 | 567 | 288 | 92 | 65 | 2722 | 471 | 299 | 182 |
| 113954 | 1838 | 515 | 238 | 97 | 53 | 2725 | 460 | 289 | 193 |
| 115792 | 1838 | 499 | 235 | 88 | 57 | 2727 | 458 | 284 | 192 |
| 117630 | 1838 | 464 | 235 | 68 | 40 | 2731 | 456 | 285 | 191 |
| 119467 | 1837 | 568 | 289 | 101 | 42 | 2743 | 459 | 288 | 187 |
| 121305 | 1838 | 580 | 271 | 123 | 66 | 2762 | 462 | 289 | 183 |
| 123143 | 1838 | 538 | 257 | 107 | 58 | 2779 | 466 | 287 | 187 |
| 124980 | 1837 | 507 | 227 | 98 | 57 | 2786 | 459 | 287 | 193 |
| 126818 | 1838 | 425 | 185 | 78 | 44 | 2789 | 459 | 283 | 196 |
| 128656 | 1838 | 457 | 209 | 78 | 44 | 2797 | 460 | 283 | 190 |
| 130494 | 1838 | 567 | 291 | 114 | 45 | 2804 | 456 | 283 | 186 |
| 132332 | 1838 | 533 | 257 | 109 | 54 | 2807 | 452 | 282 | 191 |
| 134170 | 1838 | 542 | 271 | 95 | 48 | 2812 | 444 | 288 | 190 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 136008 | 1838 | 544 | 284 | 94 | 39 | 2821 | 449 | 284 | 190 |
| 137846 | 1838 | 546 | 264 | 100 | 53 | 2832 | 445 | 290 | 190 |
| 139684 | 1838 | 473 | 225 | 85 | 33 | 2835 | 442 | 288 | 190 |
| 141522 | 1838 | 568 | 291 | 109 | 49 | 2849 | 446 | 290 | 195 |
| 143360 | 1838 | 536 | 268 | 90 | 52 | 2855 | 447 | 277 | 196 |
| 145198 | 1838 | 504 | 236 | 83 | 49 | 2857 | 445 | 278 | 195 |
| 147036 | 1838 | 560 | 268 | 115 | 52 | 2860 | 436 | 285 | 191 |
| 148874 | 1838 | 548 | 259 | 107 | 51 | 2865 | 438 | 282 | 184 |
| 150712 | 1838 | 462 | 205 | 80 | 43 | 2868 | 439 | 280 | 183 |
| 152550 | 1838 | 445 | 140 | 131 | 32 | 2874 | 441 | 281 | 179 |
| 154388 | 1838 | 542 | 270 | 104 | 54 | 2878 | 434 | 282 | 176 |
| 156226 | 1838 | 578 | 301 | 97 | 54 | 2883 | 433 | 280 | 177 |
| 158064 | 1838 | 611 | 323 | 113 | 47 | 2899 | 436 | 282 | 176 |
| 159902 | 1838 | 535 | 261 | 108 | 54 | 2902 | 435 | 279 | 177 |
| 161740 | 1838 | 538 | 263 | 91 | 49 | 2908 | 436 | 279 | 175 |
| 163578 | 1838 | 514 | 259 | 91 | 51 | 2911 | 435 | 276 | 179 |
| 165416 | 1838 | 488 | 228 | 91 | 47 | 2912 | 435 | 274 | 174 |
| 167254 | 1838 | 599 | 301 | 117 | 61 | 2919 | 435 | 273 | 174 |
| 169092 | 1838 | 561 | 277 | 104 | 50 | 2922 | 428 | 273 | 177 |
| 170930 | 1838 | 532 | 257 | 98 | 38 | 2924 | 425 | 268 | 183 |
| 172768 | 1838 | 501 | 246 | 82 | 39 | 2927 | 424 | 268 | 187 |
| 174606 | 1838 | 513 | 224 | 124 | 61 | 2930 | 423 | 265 | 186 |
| 176444 | 1838 | 525 | 262 | 117 | 34 | 2936 | 424 | 259 | 185 |
| 178282 | 1838 | 549 | 249 | 112 | 61 | 2943 | 428 | 257 | 185 |
| 180120 | 1838 | 501 | 238 | 89 | 49 | 2949 | 429 | 258 | 184 |
| 181958 | 1838 | 527 | 263 | 95 | 52 | 2954 | 428 | 262 | 172 |
| 183796 | 1838 | 526 | 231 | 118 | 56 | 2958 | 427 | 261 | 171 |
| 185634 | 1838 | 565 | 283 | 105 | 53 | 2964 | 426 | 266 | 168 |
| 187472 | 1838 | 496 | 238 | 90 | 49 | 2967 | 427 | 263 | 166 |
| 189310 | 1838 | 534 | 253 | 113 | 42 | 2974 | 427 | 257 | 175 |
| 191148 | 1838 | 475 | 225 | 75 | 50 | 2976 | 423 | 253 | 181 |
| 192986 | 1838 | 521 | 254 | 94 | 47 | 2983 | 424 | 256 | 177 |
| 194824 | 1838 | 513 | 247 | 92 | 53 | 2988 | 423 | 258 | 181 |
| 196662 | 1838 | 479 | 222 | 82 | 49 | 2992 | 421 | 261 | 180 |
| 198500 | 1838 | 510 | 244 | 90 | 43 | 2996 | 421 | 260 | 180 |
| 200338 | 1838 | 533 | 251 | 96 | 49 | 2997 | 422 | 252 | 184 |
| 202326 | 1988 | 614 | 299 | 126 | 57 | 3003 | 422 | 254 | 185 |
| 204164 | 1838 | 522 | 250 | 99 | 55 | 3004 | 421 | 253 | 182 |
| 206002 | 1838 | 569 | 300 | 107 | 47 | 3007 | 419 | 247 | 179 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 207840 | 1838 | 533 | 264 | 90 | 48 | 3008 | 416 | 243 | 182 |
| 209678 | 1838 | 551 | 291 | 94 | 40 | 3015 | 414 | 247 | 180 |
| 211516 | 1838 | 509 | 240 | 88 | 57 | 3015 | 410 | 245 | 182 |
| 213354 | 1838 | 522 | 259 | 104 | 41 | 3019 | 409 | 243 | 185 |
| 215192 | 1838 | 515 | 235 | 100 | 43 | 3022 | 409 | 242 | 185 |
| 217030 | 1838 | 518 | 223 | 105 | 60 | 3022 | 403 | 244 | 186 |
| 218868 | 1838 | 549 | 266 | 95 | 60 | 3028 | 405 | 245 | 187 |
| 220706 | 1838 | 525 | 270 | 89 | 41 | 3034 | 406 | 240 | 193 |
| 222544 | 1838 | 521 | 250 | 96 | 40 | 3036 | 405 | 239 | 190 |
| 224382 | 1838 | 550 | 295 | 81 | 58 | 3041 | 400 | 245 | 187 |
| 226220 | 1838 | 471 | 217 | 97 | 46 | 3044 | 399 | 242 | 186 |
| 228058 | 1838 | 569 | 284 | 101 | 60 | 3048 | 397 | 244 | 184 |
| 229896 | 1838 | 479 | 243 | 79 | 38 | 3050 | 399 | 241 | 183 |
| 231734 | 1838 | 513 | 245 | 102 | 54 | 3053 | 401 | 239 | 178 |
| 233572 | 1838 | 521 | 260 | 88 | 55 | 3057 | 399 | 243 | 177 |
| 235410 | 1838 | 499 | 224 | 109 | 44 | 3062 | 399 | 243 | 176 |
| 237248 | 1838 | 509 | 252 | 81 | 50 | 3065 | 398 | 240 | 179 |
| 239086 | 1838 | 511 | 235 | 109 | 50 | 3070 | 401 | 240 | 178 |
| 240924 | 1838 | 527 | 264 | 100 | 51 | 3075 | 396 | 248 | 176 |
| 242762 | 1838 | 527 | 255 | 98 | 58 | 3079 | 397 | 247 | 175 |
| 244600 | 1838 | 504 | 227 | 105 | 55 | 3081 | 398 | 243 | 177 |
| 246438 | 1838 | 553 | 279 | 109 | 48 | 3089 | 403 | 242 | 175 |
| 248276 | 1838 | 580 | 286 | 116 | 48 | 3090 | 402 | 240 | 174 |
| 250114 | 1838 | 534 | 258 | 101 | 63 | 3094 | 405 | 237 | 174 |
| 251952 | 1838 | 557 | 266 | 108 | 49 | 3101 | 403 | 239 | 178 |
| 253790 | 1838 | 508 | 246 | 98 | 41 | 3103 | 402 | 239 | 178 |
| 255628 | 1838 | 441 | 193 | 78 | 46 | 3107 | 404 | 239 | 178 |
| 257466 | 1838 | 566 | 269 | 117 | 57 | 3112 | 399 | 242 | 182 |
| 259304 | 1838 | 479 | 231 | 86 | 43 | 3115 | 402 | 240 | 183 |
| 261142 | 1838 | 525 | 242 | 116 | 48 | 3120 | 402 | 241 | 184 |
| 262980 | 1838 | 523 | 264 | 96 | 41 | 3121 | 401 | 238 | 185 |
| 264818 | 1838 | 469 | 222 | 90 | 41 | 3122 | 398 | 239 | 186 |
| 266656 | 1838 | 539 | 256 | 112 | 43 | 3126 | 399 | 238 | 184 |
| 268494 | 1838 | 539 | 255 | 85 | 64 | 3127 | 397 | 238 | 185 |
| 270332 | 1838 | 521 | 257 | 89 | 53 | 3130 | 397 | 236 | 187 |
| 272170 | 1838 | 515 | 257 | 88 | 50 | 3135 | 400 | 237 | 185 |
| 274008 | 1838 | 518 | 230 | 113 | 54 | 3140 | 401 | 236 | 184 |
| 275846 | 1838 | 444 | 188 | 80 | 42 | 3144 | 401 | 238 | 183 |
| 277684 | 1838 | 538 | 264 | 91 | 47 | 3145 | 401 | 235 | 183 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 279522 | 1838 | 508 | 249 | 88 | 55 | 3147 | 401 | 235 | 181 |
| 281360 | 1838 | 541 | 278 | 98 | 40 | 3151 | 400 | 237 | 179 |
| 283198 | 1838 | 537 | 263 | 103 | 52 | 3156 | 398 | 239 | 177 |
| 285036 | 1838 | 518 | 260 | 102 | 49 | 3159 | 400 | 237 | 176 |
| 286874 | 1838 | 526 | 272 | 86 | 44 | 3164 | 400 | 236 | 176 |
| 288712 | 1838 | 442 | 201 | 67 | 42 | 3169 | 402 | 236 | 177 |
| 290550 | 1838 | 512 | 263 | 79 | 52 | 3171 | 403 | 234 | 178 |
| 292388 | 1838 | 571 | 278 | 109 | 59 | 3175 | 404 | 233 | 177 |
| 294226 | 1838 | 526 | 262 | 88 | 55 | 3178 | 405 | 234 | 176 |
| 296064 | 1838 | 591 | 292 | 123 | 51 | 3184 | 408 | 229 | 174 |
| 297902 | 1838 | 571 | 301 | 98 | 52 | 3188 | 407 | 230 | 173 |
| 299740 | 1838 | 533 | 261 | 100 | 46 | 3192 | 407 | 227 | 177 |
| 301578 | 1838 | 516 | 250 | 92 | 51 | 3192 | 403 | 227 | 181 |
| 303416 | 1838 | 623 | 346 | 102 | 54 | 3194 | 400 | 231 | 180 |
| 305254 | 1838 | 508 | 243 | 94 | 52 | 3195 | 399 | 230 | 181 |
| 307092 | 1838 | 524 | 268 | 89 | 46 | 3197 | 400 | 228 | 180 |
| 308930 | 1838 | 547 | 274 | 112 | 53 | 3200 | 402 | 228 | 178 |
| 310768 | 1838 | 555 | 268 | 101 | 59 | 3201 | 400 | 227 | 177 |
| 312606 | 1838 | 588 | 321 | 89 | 50 | 3204 | 399 | 225 | 178 |
| 314444 | 1838 | 506 | 250 | 104 | 38 | 3208 | 401 | 226 | 177 |
| 316282 | 1838 | 514 | 240 | 94 | 47 | 3208 | 400 | 224 | 177 |
| 318120 | 1838 | 566 | 285 | 117 | 47 | 3212 | 400 | 226 | 176 |
| 319958 | 1838 | 508 | 228 | 105 | 52 | 3213 | 400 | 225 | 176 |
| 321796 | 1838 | 529 | 248 | 100 | 49 | 3216 | 401 | 222 | 179 |
| 323634 | 1838 | 501 | 252 | 88 | 41 | 3219 | 402 | 224 | 172 |

# Zur Homogenität von Graphemhäufigkeiten in Texten: Evidenz aus dem Russischen

*Emmerich Kelih (Graz, Österreich)*

## *0. EINLEITUNG*

Ein zentrales Problem einer methodologisch an der Anwendung statistisch/quantitativer Verfahren orientierten Textlinguistik ist die Frage nach einer adäquaten Textbasis bzw. Stichprobe. In der Regel – sofern eine Untersuchung nicht explizit literaturwissenschaftlich-stilistischen Fragestellungen gewidmet ist – kann im Grunde genommen ein (beliebiger) sprachlicher Akt (Text) dazu dienen, um postulierte Hypothesen und Gesetzesannahmen empirisch zu überprüfen. Eine theoretische Grundvoraussetzung ist, dass die untersuchten Texte dem Kriterium der Homogenität zu genügen haben. In diesem Beitrag soll die Homogenität von Texten anhand eines ausgewählten Merkmals, nämlich des quantitativen Verhaltens von Graphemhäufigkeiten näher diskutiert werden. Verglichen werden zehn Kapitel eines russischen Romans, die als einzelne Entitäten (Kapitel), als Ganzes (Gesamtkorpus) und als Kumulation von Kapiteln (Dynamik der Texte) analysiert werden.

## *1. PROBLEMSTELLUNG*

Spätestens seit den Arbeiten des georgischen Kybernetikers Ju.K. Orlov ist der quantitativen Sprach- und Textanalyse bekannt, dass nicht beliebiges sprachliches Material quantitativen Gesetzmäßigkeiten (z.B. Zipf-Mandelbrotsches Gesetz) unterliegt, sondern vornehmlich sprachliche Texte, die ganzheitlich und abgeschlossen (vgl. Orlov 1982a, 1982b) sind. Eine derartige, holistische Interpretation eines Textes ist sicherlich nicht unproblematisch hinsichtlich der Definition und Operationalisierbarkeit der semantischen „Ganzheit“ von Texten. Dennoch ist die Untersuchung von Texten, die „in einem Redeakt“ erschaffen worden sind und in einem Akt der Perzeption erfasst werden können, als ein

gangbarer Weg zu sehen. Letztendlich hängt dies mit dem Problem der Homogenität eines Textes und daraus gewonnener Daten zusammen[1].

Beide Probleme – die Abgeschlossenheit und die damit zusammenhängende Homogenität von Texten – stehen wiederum in engem Zusammenhang mit den „Produktionsbedingungen", die als eine Randbedingung der Textproduktion anzuführen ist. Vereinfacht gesagt wirken auf die Textproduktion unter anderem drei unterschiedliche Faktoren: Erstens, ein Text wird durch die inhaltlich-formale Struktur determiniert, d.h. ein individueller Text ist immer ein Vertreter einer größeren Gruppen von ähnlichen Texten (Textsorten, Funktionalstilen). Zweitens ist die Disposition des Schreibers zu beachten, der durch lange Pausen bzw. Selbstkorrekturen die Struktur eines Textes – sei es nun bewusst oder unbewusst – steuert (vgl. Popescu/Altmann 2006: 29) und ändert. Drittens – und dies ist eine Randbedingung, die bislang wenig beachtet wurde – sind die Umfänge von Inventaren unterschiedlicher sprachlicher Ebenen, die zur Texterzeugung verwendet werden, hervorzuheben. Ausgehend von einem mehr oder weniger abgeschlossenen und überschaubaren Graphem- bzw. Phonemsystem sind alle hierarchisch darüber liegenden Ebenen von ihrem Umfang her weitaus umfangreicher (Morpheminventar, Lexeme usw.) und ermöglichen einem Autor eine weitaus größere Wahl- und Selektionsmöglichkeit, womit auf einer höheren sprachlichen Ebene auch die Heterogenität zunimmt.

Wie dem auch sei: Das Resultat der Textproduktion muss aber in letzter Instanz so gestaltet sein, dass ein Hörer/Leser dieses mit mehr oder weniger viel Dekodierungsaufwand aufnehmen und verarbeiten kann. Die Einführung des Inventarumfangs sprachlicher Einheiten als zusätzliche Randbedingung führt dazu, dass man kaum von einer „absoluten" Homogenität eines Textes sprechen kann. Vielmehr ist von einer relationalen Homogenität auszugehen. Diese sollte auf eine bestimmte sprachliche Ebene bezogen werden, da sie in Abhängigkeit davon unterschiedlich ausfällt. So ist die Homogenität des Graphemsystems aufgrund des abgeschlossenen und überschaubaren Inventars vermutlich höher anzusetzen als die des lexikalischen Systems, welches durch ein hohes Inventar charakterisiert ist.

In engem Zusammenhang mit den Produktionsbedingungen und der Homogenität steht darüber hinaus die Frage von linguistischen

[1] Als ideal für quantitative Untersuchungen werden „mittellange" Privatbriefe, Tagebuchaufzeichnungen, kurze Prosatexte u.ä. bezeichnet. Vgl. dazu Best (1994) und Best (2006: 37).

„Stichproben“ bzw. deren Verhältnis zu einer wie auch immer definierten „Grundgesamtheit“ (bezogen auf die entsprechenden Überlegungen aus dem Bereich der russischen quantitativen Linguistik vgl. Kelih 2008: 238ff). Die Dichotomie Stichprobe vs. Grundgesamtheit kann als eine Relation zwischen „Teil“ und „Ganzem“ verstanden werden, wobei aus statistischer Sicht ein Teil bereits die Eigenschaften des Ganzen widerspiegeln müsste. Die herkömmliche Interpretation einer sprachlichen Grundgesamtheit als „repräsentative Gesamtsprache“, die mit einer Sammlung einer Vielzahl von unterschiedlichen Texten und Textsorten gleichgesetzt wird, ist aber aus heutiger Sicht nicht (mehr) plausibel. Ein derartiges Vorgehen führt nämlich dazu, dass immer mehr heterogenes sprachliches Material zusammengeführt wird und man so zu einer Mischung unterschiedlicher Sprachschichten[2] – also Quasi-Texten – gelangt. Diese Heterogenisierung des sprachlichen Materials kann unter Umständen sogar ein verzerrtes Bild von Regulationsmechanismen sprachlicher Texte liefern. Vorstellbar ist des Weiteren auch, dass der Nachweis von quantitativen Gesetzmäßigkeiten in einem derartigen gemischten Material erschwert wird bzw. gänzlich unmöglich ist.

An dieser Stelle setzt der vorliegende Beitrag – der vor dem Hintergrund der angedeuteten Problematik nur bescheiden ausfallen kann – ein. Die Frage der linguistischen und statistischen Homogenität von Texten wird ausschließlich auf der vermeintlich „untersten“ Ebene – der Graphemebene – untersucht. Die Datenbasis ist ein russischer Roman, von dem zehn einzelne Kapitel untersucht werden. Es geht um die Frage, ob sich eine Änderung des Frequenzverhaltens von Graphemen zwischen den einzelnen Kapiteln bzw. in der Relation Kapitel – Gesamtkorpus beobachten lässt. Als methodologisch sinnvolle Instrumentarien werden diskutiert: (1) die Wiederholungsrate, die auf der Basis von Graphem-Ranghäufigkeiten errechnet wird, (2) das Verhalten von Parametern aus geeigneten theoretischen Modellen der Modellierung von Graphemranghäufigkeiten und (3) der bislang für Untersuchungen von Graphemhäufigkeiten nicht diskutierte Hirschsche-Punkt (h-Punkt) unter Einschluss der Bogenlänge und daraus berechneter Koeffizienten. Und (4) wird – neben den einzelnen Kapiteln und dem Gesamtkorpus – das Verhalten der Wiederholungsrate beim sukzessiven

[2] Selbst die in den 60er Jahren propagierte Untersuchung von „repräsentativen“ Text-Stichproben (Fragmente von Texten) erwies sich in letzter Instanz als nicht zielführend, da inferenzstatistische Methoden aus unterschiedlichen Gründen bei der Untersuchung von Sprache/Text nicht immer angewandt werden können.

Anwachsen der Graphemhäufigkeiten (Kapitelkumulationen) näher untersucht. Letzterer Aspekt des Verhaltens von Textkumulationen ist in der bisherigen Forschung vernachlässigt worden. Es wird daher ein neues methodologisches Instrumentarium vorgeschlagen, welches potentiell in der Lage sein sollte, Heterogenitäten/Homogenitäten im kumulativen Zuwachs von Graphemhäufigkeiten aufzuzeigen.

## *2. TEXTAUSWAHL*

Die Frage des Verhaltens von „Ganzes" und „Teile eines Ganzen" und das Verhalten der Einzelkomponenten wird anhand des russischen Romans „Kak zakaljas' stal' / Wie der Stahl gehärtet wurde" aus den Jahren 1932-1934 untersucht. Von diesem Roman, der insgesamt 18 Kapitel umfasst, werden die ersten neun Kapitel des ersten Teiles und das erste Kapitel des zweiten Teiles untersucht. Insgesamt stehen somit zehn Kapitel zur Verfügung. Aus diesen Texten wird darüber hinaus ein Gesamttext, der im Folgenden als Gesamtkorpus (GK) bezeichnet wird, erstellt und analysiert.

Aus inhaltlicher Sicht (vgl. dazu auch Kelih 2009a) handelt es sich um zehn vollständige Kapitel des bekannten sozialrealistischen Romans[3] aus den 30er Jahren, der hauptsächlich aus deskriptiven Erzählsequenzen und mündlicher Rede besteht. Formal auffällig ist das neunte Kapitel, welches im Gegensatz zu den anderen Kapiteln, fast ausschließlich aus Tagebuchaufzeichnungen einer Krankenschwester über den Gesundheitszustand eines der Romanhelden besteht. Dieses Kapitel kann man aus inhaltlich-formaler Sicht – in Relation zu den anderen Kapiteln – als „heterogen" bezeichnen.[4]

In den untersuchten Texten wird die Häufigkeit von Graphemen bestimmt. Ausgegangen wird für das Russische von einem Inventar K = 32, da dass <ë> im Text an keiner Stelle realisiert wird. Die absoluten

[3] Unklar bleibt bis heute, in welchem Ausmaß der Text von N. Ostrovskij stammt bzw. wie hoch der Anteil von Eingriffen seitens der Redaktion der Molodaja Gvardija – dem Verlag, in dem dieses Buch zuerst in Teilen erschien – ist. Darüber hinaus wird berichtet, dass N. Ostrovskij die letzten Kapitel seines Romans aufgrund einer beginnenden Erblindung nur mehr diktiert habe. Details dazu in Guski (1981). Beim Lesen der heutigen Version fallen jedoch keinerlei Stilbrüche oder ähnliches auf, die auf einen massiven Eingriff in die Textstruktur schließen lassen.

[4] Die im Text vorkommenden Gedichte und Lieder wurden von der Untersuchung ausgeschlossen.

Häufigkeiten in den 10 Texten und im Gesamtkorpus sind in Tab. 1 zusammengefasst. Ohne Berücksichtigung des Gesamtkorpus beträgt die minimale Kapitellänge 17.533 Grapheme (Kapitel 9), während das längste, sechste Kapitel 39.820 Grapheme umfasst.

Tabelle 1
Absolute Häufigkeiten von russischen Graphemen in 10 Texten und Gesamtkorpus

| **Kapitel** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **Gesamt-Korpus GK** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **а** | 1935 | 1984 | 3123 | 1791 | 1654 | 3572 | 2924 | 2525 | 1476 | 2522 | 23506 |
| **б** | 339 | 359 | 587 | 401 | 300 | 670 | 545 | 491 | 317 | 489 | 4498 |
| **в** | 1041 | 1052 | 1605 | 1061 | 981 | 1731 | 1596 | 1400 | 797 | 1428 | 12692 |
| **г** | 361 | 370 | 599 | 424 | 373 | 800 | 608 | 561 | 372 | 557 | 5025 |
| **д** | 642 | 770 | 944 | 589 | 609 | 1260 | 1002 | 923 | 527 | 881 | 8147 |
| **е** | 1519 | 1768 | 2676 | 1586 | 1644 | 3112 | 2745 | 2314 | 1518 | 2323 | 21205 |
| **ж** | 169 | 253 | 349 | 208 | 226 | 350 | 362 | 259 | 196 | 295 | 2667 |
| **з** | 377 | 417 | 645 | 380 | 398 | 776 | 607 | 576 | 320 | 549 | 5045 |
| **и** | 1299 | 1459 | 2101 | 1384 | 1246 | 2413 | 2246 | 2035 | 1067 | 1890 | 17140 |
| **й** | 227 | 272 | 397 | 270 | 241 | 460 | 429 | 348 | 230 | 413 | 3287 |
| **к** | 870 | 857 | 1309 | 771 | 723 | 1474 | 1145 | 1177 | 559 | 1119 | 10004 |
| **л** | 1070 | 1026 | 1623 | 1079 | 1036 | 2095 | 1560 | 1509 | 852 | 1414 | 13264 |
| **м** | 577 | 662 | 1018 | 629 | 582 | 1132 | 1024 | 881 | 478 | 851 | 7834 |
| **н** | 1127 | 1302 | 2077 | 1184 | 1229 | 2551 | 2009 | 1818 | 1112 | 1734 | 16143 |
| **о** | 2211 | 2276 | 3474 | 2244 | 2178 | 4373 | 3549 | 3008 | 1876 | 3116 | 28305 |
| **п** | 725 | 594 | 967 | 634 | 549 | 1158 | 894 | 881 | 514 | 816 | 7732 |
| **р** | 925 | 1124 | 1574 | 1023 | 991 | 1876 | 1679 | 1585 | 871 | 1455 | 13103 |
| **с** | 1108 | 1182 | 1756 | 1124 | 1002 | 2065 | 1819 | 1480 | 941 | 1503 | 13980 |
| **т** | 1186 | 1254 | 1849 | 1101 | 1060 | 2322 | 1890 | 1526 | 1038 | 1641 | 14867 |
| **у** | 728 | 666 | 1121 | 656 | 665 | 1218 | 968 | 921 | 517 | 936 | 8396 |
| **ф** | 47 | 19 | 18 | 29 | 26 | 31 | 27 | 52 | 28 | 35 | 312 |
| **х** | 243 | 208 | 297 | 215 | 200 | 353 | 282 | 288 | 135 | 285 | 2506 |
| **ц** | 70 | 120 | 113 | 144 | 60 | 138 | 157 | 155 | 50 | 91 | 1098 |
| **ч** | 332 | 281 | 469 | 275 | 244 | 581 | 469 | 334 | 282 | 412 | 3679 |
| **ш** | 259 | 251 | 372 | 236 | 258 | 467 | 312 | 295 | 146 | 263 | 2859 |
| **щ** | 69 | 102 | 125 | 73 | 68 | 130 | 149 | 90 | 49 | 116 | 971 |
| **ъ** | 2 | 5 | 1 | 11 | 4 | 8 | 8 | 8 | 4 | 8 | 59 |
| **ы** | 353 | 432 | 580 | 504 | 380 | 760 | 648 | 613 | 351 | 570 | 5191 |
| **ь** | 447 | 413 | 601 | 344 | 385 | 773 | 613 | 514 | 349 | 518 | 4957 |
| **э** | 36 | 43 | 83 | 50 | 36 | 74 | 51 | 55 | 46 | 65 | 539 |
| **ю** | 129 | 114 | 197 | 160 | 111 | 224 | 172 | 164 | 102 | 183 | 1556 |
| **я** | 415 | 413 | 687 | 415 | 419 | 873 | 683 | 601 | 413 | 558 | 5477 |
| | 20838 | 22048 | 33337 | 20995 | 19878 | 39820 | 33172 | 29387 | 17533 | 29036 | 266044 |

Die absoluten Häufigkeiten werden in Ranghäufigkeiten transformiert. In Abb. 1 sind – als Illustration für das verwendete Material – zwei Datensätze graphisch dargestellt: Zum einen die relativen Ranghäufigkeiten des Kapitel 1 und zum anderen die relativen Ranghäufigkeiten des Gesamtkorpus.

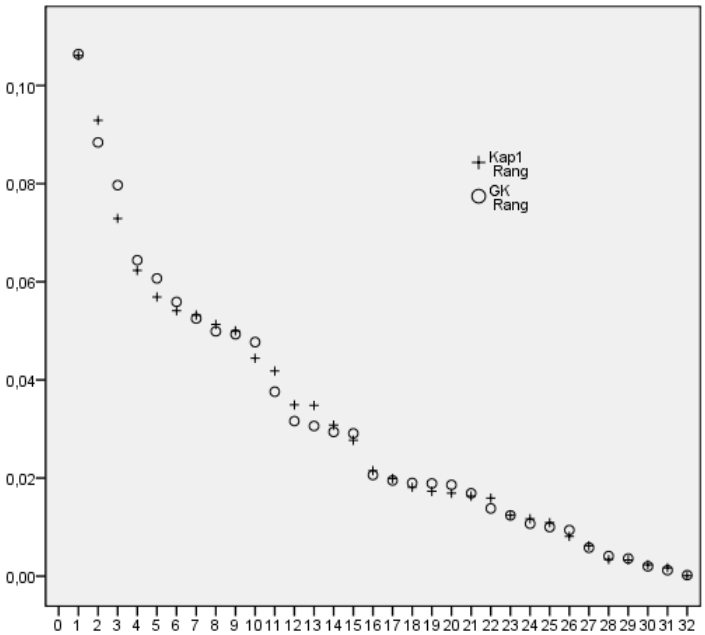



Abb. 1: Ranghäufigkeiten: Text 1 und Gesamtkorpus

Die graphische Darstellung zeigt zwar an der Oberfläche einen ähnlichen Verlauf der Ranghäufigkeiten. Betrachtet man Abb. 1 etwas ausführlicher, so wird unter anderem auffallen, dass sich zwischen dem Gesamtkorpus und dem Text 1 einige Unterschiede auftun: Die relative Frequenz in Rang 2 und 3 ist deutlich unterschiedlich. Auch in den Rängen 10, 11, 12 und 13 zeigt sich, dass ein Kapitel nicht unbedingt das Verhalten des Gesamtkorpus aufweist. Selbstverständlich ist aber eine derartige graphische Interpretation selektiver Punkte auf einer Rangverteilungskurve aus systemlinguistischer Sicht nicht befriedigend und muss durch eine Reihe von Systemeigenschaften ergänzt werden. Erst

dann bekommt man nachhaltige Einsichten in das globale Verhalten von Graphemhäufigkeiten.

### *2.1. Wiederholungsrate*

Als zentrale Kenngrößen, die das globale Verhalten von relativen Ranghäufigkeiten beschreibt, wird die Wiederholungsrate (RR) (vgl. Altmann/Lehfeldt 1980) angeführt. Die Wiederholungsrate lässt sich als $RR = \sum_{r=1}^{n} p_r^2$ berechnen, d.h. als die Summe der quadrierten relativen Häufigkeiten. Dieses Maß wurde bereits des Öfteren für die Charakterisierung von Graphem- bzw. Phonemhäufigkeiten herangezogen (vgl.Grzybek/Kelih/Altmann 2005: 122f.). Die für die 10 Kapitel berechnete Wiederholungsrate findet sich in Tab. 2.

Tab. 2:

| Kapitel | RR |
|---|---|
| 1 | 0,0539 |
| 2 | 0,0542 |
| 3 | 0,0547 |
| 4 | 0,0532 |
| 5 | 0,0545 |
| 6 | 0,0549 |
| 7 | 0,0551 |
| 8 | 0,0539 |
| 9 | 0,0546 |
| 10 | 0,0544 |
| GK | 0,0543 |

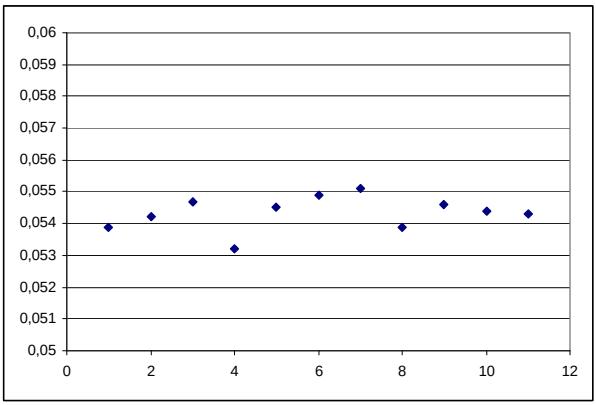



Abb. 2: Wiederholungsrate in 10 Kapiteln und im Gesamtkorpus (Datenpunkt 11)

Betrachtet man RR in den einzelnen Kapiteln und im Gesamtkorpus, so ist es durchaus gerechtfertigt von einer allgemeinen Stabilität der Werte zu sprechen: Das $RR_{min}$ beträgt 0.0532, während das Maximum bei 0.0551 liegt; der Mittelwert $\overline{RR}$ beträgt 0.0543. Damit sind Unterschiede erst auf der dritten Kommastelle zu beobachten.

Es sind auch keinerlei Ausreißer und sonstige Auffälligkeiten zu verzeichnen. Dieser Befund lässt somit die vorläufige Interpretation zu, dass in den einzelnen Kapiteln und im Gesamtkorpus hinsichtlich der Wiederholungsrate ein homogenes Gesamtbild zu beobachten ist. Dies zeigt auch die graphische Darstellung in Abb. 2. Die Wiederholungsrate innerhalb ein und desselben Romans, der in unterschiedliche Teile (Kapitel) gegliedert ist, unterscheidet sich – ohne an dieser Stelle einen passenden Test auf Signifikanz zu diskutieren – zumindest nicht „auffällig" voneinander.

### *2.2. Theoretische Modellierung von Ranghäufigkeiten*

Es kann hier die allgemeine Frage nach einem passenden (diskreten bzw. stetigen) Modell für Graphemranghäufigkeiten nicht im Detail diskutiert werden (vgl. dazu Grzybek/Kelih/ Altmann 2004, Grzybek/ Kelih 2005 u.v.m.). Aus einer vorangehenden Studie (Kelih 2009b) zu Graphemranghäufigkeiten in zwölf slawischen Sprachen (Datenbasis ist der in dieser Arbeit untersuchte Roman und die Übersetzungen in weitere elf slawische Standardsprachen) hatte sich aus einer Vielzahl von stetigen Funktionen jenes von Popescu/Altmann/Köhler (2009) $y = 1+ae^{-bx}$ besonders bewährt[5]. Daher soll an dieser Stelle ausschließlich dieses Modell empirisch gestestet[6] werden.

Es sollte jedoch nicht die Modellierbarkeit an sich ins Zentrum gerückt werden, sondern vielmehr das Verhalten der Parameter *a* und *b*. Diese können als Gradmesser für die Homogenität der Graphemranghäufigkeit in den untersuchten Texten dienen.

Zu beginnen ist mit der empirischen Validität des vorgeschlagenen Modells (vgl. dazu Tab. 3): Der durchschnittliche Determinationskoeffizient $\overline{R^2}$ für die 10 Texte und das Gesamtkorpus liegt bei $\overline{R^2} = 0.9765$. Dies ist ein sehr hoher Wert und bezeugt, dass insgesamt

[5] Die von Popescu/Altmann/Köhler (2009) vorgeschlagene Funktion beinhaltet die „Mischung mehrerer Komponenten". Es konnte in Kelih (2009b) gezeigt werden, dass jedoch eine Komponente, wie oben angeführt, „ausreicht", um slawische Graphemhäufigkeiten adäquat erfassen zu können. Das von den Autoren angesetzte Modell – dieses ist eigentlich für Worthäufigkeiten entwickelt worden – der „Mischung von Strata" muss für Häufigkeiten von Graphemen somit nicht unbedingt angesetzt werden.

[6] Als Gradmesser für die Anpassungsgüte des Modells wird der übliche Determinationskoeffizient ($R^2$) herangezogen.

von einer überzeugenden Übereinstimmung zwischen den empirischen und theoretischen Werten gesprochen werden kann. Darüber hinaus zeigt sich, dass alle Texte und das Gesamtkorpus offensichtlich durch ein einziges Modell – auch dies ist ein zentraler Hinweis auf die ähnliche Struktur der Graphemhäufigkeiten in diesem Roman – beschrieben werden können. Einzig und allein für das dritte Kapitel wurde ein etwas geringeres, aber immer noch zufriedenstellendes $R^2 = 0.9320$ festgestellt. Ansonsten haben alle anderen Texte und das Gesamtkorpus ein $R^2 > 0.97$. Auch gibt es sowohl hinsichtlich des Verhaltens des Gesamtkorpus und der Einzelkapitel keinerlei „Auffälligkeiten".

Tabelle 3
Determinationskoeffizient und Parameter *a* und *b*

| **Kapitel** | **a** | **b** | $R^2$ |
|---|---|---|---|
| 1 | 2214,41 | 0,0960 | 0,9734 |
| 2 | 2377,63 | 0,0976 | 0,9866 |
| 3 | 3635,16 | 0,0989 | 0,9320 |
| 4 | 2206,01 | 0,0951 | 0,9765 |
| 5 | 2145,40 | 0,0978 | 0,9752 |
| 6 | 4367,50 | 0,0997 | 0,9818 |
| 7 | 3659,02 | 0,1002 | 0,9866 |
| 8 | 3127,16 | 0,0958 | 0,9858 |
| 9 | 1909,99 | 0,0990 | 0,9806 |
| 10 | 3137,18 | 0,0979 | 0,9813 |
| 11 | 28734,17 | 0,0978 | 0,9822 |

Neben der empirischen Validität des Modells ist – wie bereits gesagt – auch die quantitative Ausprägung der Parameter *a* und *b* aus der obigen Funktion von Interesse. Folgendes ist dazu zu bemerken: Der Parameter *a* ist offensichtlich für die „Verschiebung" auf der y-Achse zuständig und gibt in etwa die „Höhe" des Beginns der theoretischen Kurve vor. Damit ist eine Wechselbeziehung zwischen den Parameter *a* und der absoluten Häufigkeit in Rang 1 ($p_1$) zu erwarten. Diese Vermutung lässt sich problemlos empirisch nachweisen.

In Abb. 3 zeigt sich ein linearer Zusammenhang zwischen der Anzahl von Graphemen im Rang 1 ($p_1$) und dem Parameter *a*. Auch in diesem Fall werden nur die einzelnen Kapitel näher untersucht, da das Gesamtkorpus durch seinen Umfang von 266044 Graphemen im Gegensatz zu den Teilkapitel, die im Durchschnitt 26604 Grapheme lang sind, a priori als Ausreißer auftritt. Mit $a = 1.0266*p_1 - 27.832$

ergibt sich ein $R^2 = 0.9935$. Dieser Zusammenhang ist in Abb. 3 dargestellt.

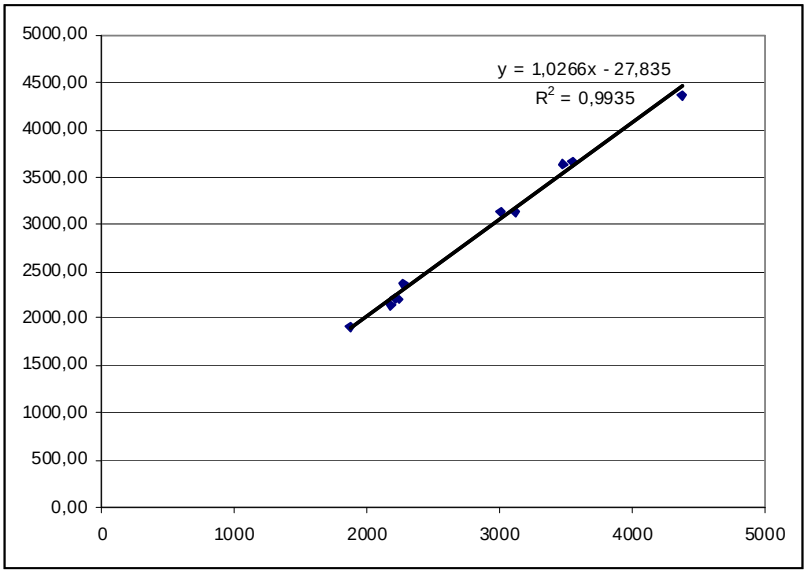



Abb. 3: Zusammenhang von $p_1$ und Parameter *a*

Der „unangenehme" Nebeneffekt dieses Befundes ist, dass der Parameter *a* für die hier zugrundeliegende Frage nach der „Homogenität" der einzelnen Kapitel nicht geeignet ist, da mit ihm (trivialerweise) gleichzeitig die Anzahl von Graphemen in einem bestimmten Rang und damit indirekt die Textlänge erfasst wird. Die Textlänge kann aber aus linguistischer Sicht jedoch kaum als Einflussvariable für die Form einer Rangverteilung von Graphemen verantwortlich gemacht werden.[7]

Geeigneter für die von uns verfolgte Fragestellung ist der Parameter *b*, der Exponent in dem obigen Modell. Der Parameter *b* weist keinerlei Beziehungen zur Textlänge auf und seine Ausprägung kann

[7] Die (alte) Idee, dass mit zunehmendem Stichprobenumfang eine „Stabilisierung" der einzelnen Graphem- oder auch Phonemhäufigkeiten eintreten sollte, ist bislang nicht nachgewiesen worden. Ein „je mehr, desto besser" kann aber selbst auf dieser untersten Ebene nicht gelten, da eine Hinzufügung von immer mehr Text zu einem Korpus die Häufigkeiten der einzelnen Grapheme/ Phoneme in die eine oder andere Richtung verschieben würde.

als Gradmesser für die Homogenität/Heterogenität der einzelnen Kapitel und des Gesamtkorpus herangezogen werden.

Er zeigt ein durchgehend homogenes Gesamtbild in dem Sinne, dass er in den einzelnen Kapiteln keinerlei starken Schwankungen ausgesetzt ist. Des Weiteren weist der Parameter *b* des Gesamtkorpus keine Besonderheiten in Bezug auf die Parameter-Werte der Teilkapitel auf. Vgl. dazu die graphische Darstellung in Abb. 4.

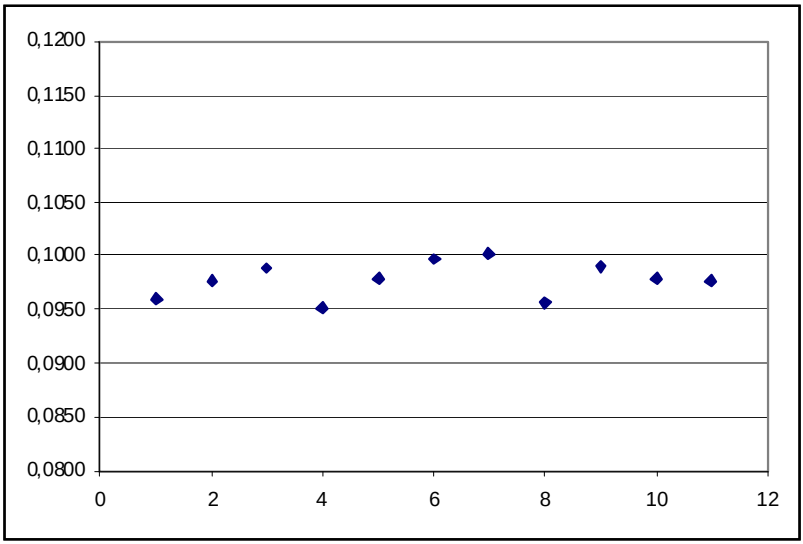



Abb.4: Konstanz des Parameter *b* in 10 Teilkapiteln
und im Gesamtkorpus (Datenpunkt 11)

Damit kann folgendes Zwischenergebnis präsentiert werden: Hinsichtlich der Wiederholungsrate und des Parameters *b* aus der Funktion von Popescu/Altmann/Köhler (2009) ist anzunehmen, dass sich das „quantitative" Verhalten der Teilkapitel untereinander nicht auffällig unterscheidet. Es ist eine Konstanz bzw. vor allem Stabilität der einzelnen Merkmale zu beobachten. Gestützt wird dieser Befund auch dadurch, dass ein gemeinsames theoretisches Modell für die Graphemhäufigkeiten gefunden werden konnte, welches für alle einzelnen Kapitel gültig ist. Darüber hinaus gibt es keinerlei Anzeichen, dass die Relation von „Teilkapitel vs. Ganzes" – systemlinguistisch gesprochen – „gestört" ist.

### *2.3. h-Punkt und Bogenlänge*

#### *2.3.1. h-Punkt*

In letzter Zeit sind für die Charakterisierung von Ranghäufigkeiten eine ganze Reihe von neuen, operational bestimmbaren Eigenschaften und Indikatoren eingeführt worden. Eine wichtige Eigenschaft stellt der sogenannte Hirschsche Punkt (im Folgenden h-Punkt) dar. Ursprünglich in die Szientometrie eingeführt, wird er mittlerweile auch in der quantitativen Linguistik intensiv diskutiert (vgl. Popescu 2006, Popescu/Altmann 2006, Popescu/Altmann 2007, Mačutek/Popescu/Altmann 2007, Popescu/Altmann 2008 u.a.).

Dieser Punkt ist die Stelle einer Rangverteilung in dem der Rang und die Frequenz zusammenfallen, d.h. es handelt sich um einen auf der Rangverteilung fixierbaren Punkt. In Fällen, in denen dieser Punkt nicht eindeutig identifiziert[8] werden kann (da kein eindeutiger Punkt vorliegt für den gilt, dass f(r) = r), so wird als h-Punkt jene Stelle genommen, an der das Produkt aus Rang und Frequenz das Maximum erreicht (vgl. Martináková et al. 2008: 93)). Wählt man letztere Vorgangsweise, so kann der h-Punkt ohne größeren rechnerischen Aufwand bestimmt werden. Für die vorliegende Untersuchung wurde aber die in Popescu/Altmann (2008: 95) vorgeschlagene, exakte mathematische Bestimmung des h-Punktes verwendet.

Bevor auf die Bedeutung dieses Punktes für Graphemrangverteilungen eingegangen wird, sei kurz auf seine Interpretation in der Textologie, insbesondere in lexikalischen Studien, verwiesen. Nach Popescu/Altmann (2006: 30) würde der h-Punkt, bezogen auf Wortfrequenzen, eine Möglichkeit darstellen, um eine Rangfolge in zwei Bereiche zu trennen: einen vorderen „rapiden" Bereich (figurativ gesprochen der „gesättigte" Bereich einer Rangverteilung) und einen zweiten „gemächlichen" Teilbereich des „langsamen Auslaufens" einer Rangverteilung.

[8] Es gibt mehrere Ansätze zur Bestimmung und zur Berechnung des h-Punktes. So wird der h-Punkt in Popescu/Altmann 2006: 29) folgendermaßen definiert: „The h-point is defined as that point of the rank-frequency distribution which is the nearest to the [0, 0], i.e. to the origin" (vgl. Popescu/Altmann 2006: 29). Eine alternative Bestimmung des h-Punktes findet sich in Mačutek/Popescu/Altmann (2007: 45). Dort wird vorgeschlagen, jenen Rangplatz als h-Punkt zu bestimmen zu nehmen, deren absolute Differenz zu f(r) minimal ist. Zu einer weiteren Bestimmung des h-Punkts vgl. Martináková et al. (2008: 97).

Bezogen auf die Verteilung lexikalischer Einheiten wird der h-Punkt als Trennmarke zwischen Auto- und Synsemantika interpretiert.

Abgesehen davon, dass sich bezogen auf Graphemhäufigkeiten allzu voreilige Analogieschlüsse verbieten würden, sei zumindest auf folgende Interpretation hingewiesen: Sofern der h-Punkt in den unterschiedlichen Kapiteln in einem bestimmten Intervall liegt, kann davon ausgegangen werden, dass eine Ähnlichkeit der einzelnen Kapitel hinsichtlich der Graphemhäufigkeiten vorliegt.

Das Ergebnis für die untersuchten russischen Texte – die Daten sind in Tab. 4. und graphisch in Abb. 5 dargestellt – lautet: Der durchschnittliche h-Punkt liegt bei 31.01 und einem Intervall von <30.52, 31.88>. Ob diese Spannweite hoch oder niedrig ist, kann zum gegenwärtigen Stand der Forschung nicht gesagt werden.

Tab.4: h-Punkt

| Kapitel | h |
|---|---|
| 1 | 31,14 |
| 2 | 30,52 |
| 3 | 30,80 |
| 4 | 30,91 |
| 5 | 30,55 |
| 6 | 31,00 |
| 7 | 30,84 |
| 8 | 31,47 |
| 9 | 30,84 |
| 10 | 31,14 |
| GK | 31,88 |

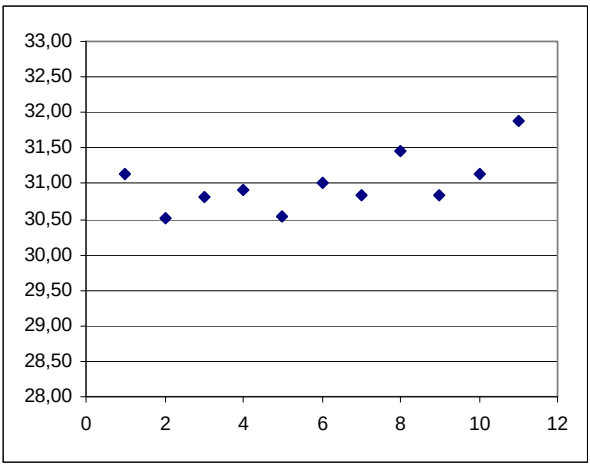



Abb. 5: h-Punkt in 11 Texten
und im Gesamtkorpus (Text 11)

In jedem Fall wird deutlich, dass sich die Fluktuation des h-Punktes in einem abgeschlossenen Bereich abspielt und keine allzu großen Auffälligkeiten zu beobachten sind. Ob damit der Schluss gezogen werden kann, dass die untersuchten Teilkapitel hinsichtlich ihres h-Punktes als homogen zu verstehen sind, soll einstweilen offen gelassen werden. In jedem Fall ist aber die Brauchbarkeit des h-Punkts zumindest angedeutet und seine Sinnhaftigkeit für Graphemuntersuchungen nachgewiesen.

*2.3.2. Bogenlänge*

Ein weiteres Charakteristikum einer Ranghäufigkeit ist die von Popescu/Mačutek/Altmann (2008: 19) eingeführte „arc-length/Bogenlänge". Das ist die Summe der Euklidischen Distanzen zwischen zwei benachbarten Ranghäufigkeiten der Sequenz und wird folgendermaßen berechnet:

$$L = \sum_{r=1}^{R-1} [(f(r) - f(r+1))^2 + 1]^{1/2} \ .$$

Hier ist f(r) die Graphemfrequenz im Rang r und R = $r_{max}$ ist der höchste bestimmbare Rang. Zu berechnen ist auch das $L_{min} = [(R-1)^2 + (f_1-1)^2]^{1/2}$ und $L_{max} = [(f_1-1)^2+1]]^{1/2} + R-2$.[9]

Die Bogenlange wird in Popescu/Mačutek/Altmann (2008) als ein Maß für die quantitative Charakterisierung von Worthäufigkeiten eingeführt. Sie zeigt direkte Zusammenhänge zum Stichprobenumfang (*N*): Je länger ein Text (Stichprobengröße in Graphemen), desto länger die Bogenlänge. Dieser Zusammenhang[10] kann mit einem einfachen linearen Modell L = 0.1082* N – 53.2 zufriedenstellend ($R^2$ = 0.99) erfasst werden Vgl. dazu die graphische Darstellung in Abb. 6.

Es ergibt sich somit ein Zusammenhang zwischen der Bogenlänge und dem Stichprobenumfang (*N*). Derartiges konnte bereits für den Parameter *a* des in Abschnitt 2.3. verwendeten theoretischen Modells festgestellt werden. Demnach ist die Bogenlänge für die von uns verfolgte Fragestellung nach der Homogenität von Graphemranghäufigkeiten in unterschiedlichen Kapiteln eines Romans nicht geeignet.

Geeigneter sind demgegenüber zwei in Popescu/Kelih/Altmann/Best (2009) eingeführte Indikatoren, die mit dem h-Punkt und der Bogenlänge L in direktem Zusammenhang stehen. Es sind dies die Indikatoren $c = \frac{R + f(1) - f(R) + 1 - L}{h}$ und $p = \frac{L_{max} - L}{h - 1}$. *R* ist der höchste Rang und *f*(1) das häufigste Graphem. Da *c* und *p* linear miteinander verbunden sind, wird im Folgenden ausschließlich der Indikator *c* näher besprochen.

[9] Zu einer alternativen Berechung von $L_{max}$ = R -1 + f(1)-f(R) vgl. Popescu/Kelih/Best/Altmann (2009).

[10] In diesem Fall wird das Gesamtkorpus nicht untersucht.

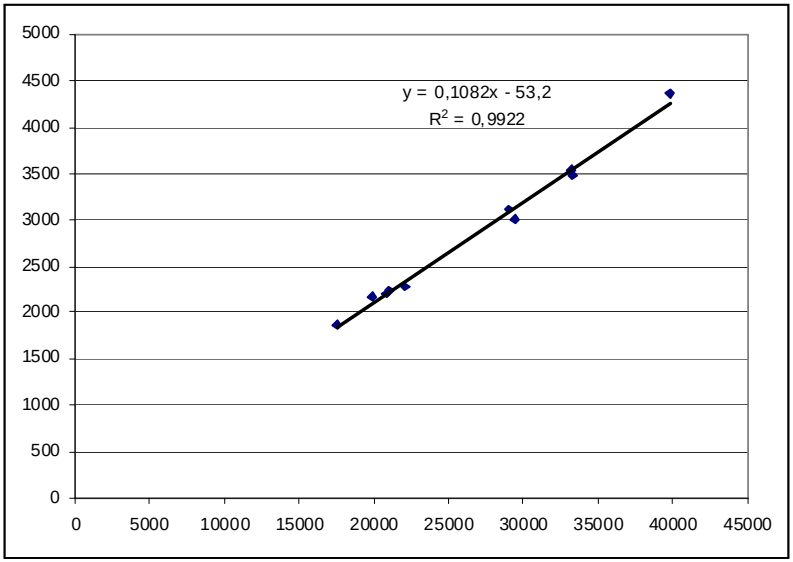



Abb.6: Stichprobengröße N (in Graphemen) vs. Bogenlänge (L)

Betrachtet man die *c*-Werte in den einzelnen Kapiteln (vgl. Tab. 5), so zeigt sich eine fast durchgehende „Stabilität“: Die Werte für *c* liegen zwischen 0.990 und 1.050 und haben einen Mittelwert von $\overline{c}$ = 1.030 (vgl. dazu auch Abb. 7). Da Erfahrungswerte für andere Sprachen bislang nicht vorliegen, kann nicht postuliert werden, ob es sich im Fall der russischen Daten um eine niedrige oder hohe Spannweite handelt. In jedem Fall nehmen sie wiederum ein recht enges Intervall ein. Im Grunde genommen zeigt nur das Kapitel 8 ein geringfügig abweichendes Verhalten.

Es zeigt sich, dass der c-Indikator geeignet ist, um als Gradmesser für die Homogenität der Daten hinsichtlich des h-Punktes und der Bogenlänge verwendet zu werden. Solange keine weiteren Vergleichswerte zu anderen Texten bzw. auch Sprachen vorliegen, kann vorläufig die enge Bandbreite der Werte als Nachweis für die globale Ähnlichkeit der einzelnen Kapitel angesehen werden.

Tab. 5: c- und p-Werte

| Kapitel | c | p |
|---|---|---|
| 1 | 1,024 | 0,992 |
| 2 | 1,017 | 0,984 |
| 3 | 1,050 | 1,018 |
| 4 | 1,044 | 1,012 |
| 5 | 1,049 | 1,017 |
| 6 | 1,043 | 1,011 |
| 7 | 1,043 | 1,010 |
| 8 | 0,990 | 0,957 |
| 9 | 1,020 | 0,987 |
| 10 | 1,019 | 0,986 |
| Gesamt-korpus | 1,033 | 1,002 |

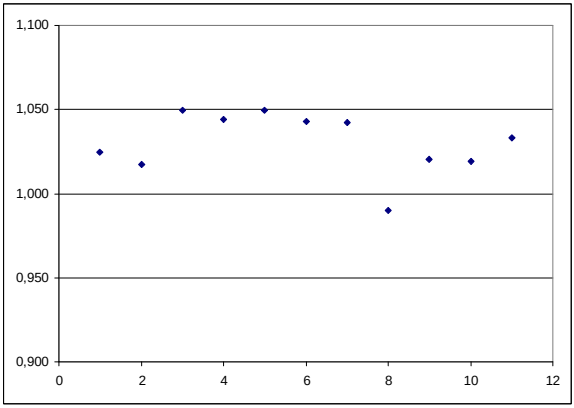



Abb. 7: c-Wert in 10 Kapiteln
und Gesamtkorpus (Datenpunkt 11)

### *3. WIEDERHOLUNGSRATE IN KUMULIERTEN KAPITELN*

Abschließend soll, wie einleitend angekündigt, eine weitere Analysemöglichkeit eingeführt werden. Diese besteht darin, nicht nur die einzelnen Texte, sondern auch das sukzessive kumulative Anwachsen der Ranggraphemhäufigkeiten in den Texten näher zu betrachten. Zur Illustration wird das Verhalten der Wiederholungsrate (RR) in den kumulierten Kapiteln näher untersucht.

Es werden zu diesem Zweck die Graphemhäufigkeiten in den Kapitel(n) 1+2, 1+2+3 … usw. bestimmt. Damit stehen 10 „neue“ Datensätze zur Verfügung, wobei letzter Datenpunkt (Nr. 10) nichts anderes ist als der gesamte analysierte Roman (vgl. die Daten in Tab. 6).

Trägt man nun die Werte für die Wiederholungsrate – und nur diese wird im Folgenden auf diese Art und Weise untersucht – gegen die einzelnen kumulierten Kapitel auf, so lässt sich bereits auf graphischer Ebene (vgl. Abb. 8) ein durchaus interessanter Befund ablesen: Offensichtlich weist der dritte Datenpunkt (das sind die kumulierten Kapitel 1, 2 und 3) einen Bruch hinsichtlich der kumulativen Wiederholungsrate auf.

Dieses Phänomen kann einstweilen nur damit erklärt werden, dass sich durch das Hinzunehmen des dritten Kapitels plötzlich das globale Häufigkeitsverhalten ändert. D.h. offensichtlich wechseln einzelne oder mehrere Grapheme gegenüber den vorangehenden Kapiteln ihre Rangplätze und rufen somit einen Bruch innerhalb der kumulierten Wiederho-

lungsrate hervor. Zu beachten ist auch, dass die Wiederholungsrate bereits im dritten Kapitel den Wert (0.0543) erreicht. Dieser Wert ergibt sich dann auch für das Gesamtkorpus (Datenpunkt Nr. 10) mit RR = 0.0543, wobei ähnliche Wiederholungsraten auch die Datenpunkte Nr. 7, 8 und 9 haben.

Tab. 6: Wiederholungsrate in Kapitelkumulation

| Nr. | RR | $\lvert RR_j - RR_{11} \rvert * 1/x$ |
|---|---|---|
| 1 | 0,0539 | 0,000041563335 |
| 2 | 0,0540 | 0,000034437474 |
| 3 | 0,0543 | 0,000003744763 |
| 4 | 0,0540 | 0,000031032710 |
| 5 | 0,0541 | 0,000025742859 |
| 6 | 0,0542 | 0,000006646790 |
| 7 | 0,0544 | 0,000006582796 |
| 8 | 0,0543 | 0,000001424966 |
| 9 | 0,0543 | 0,000000401167 |
| 10 | 0,0543 | 0 |

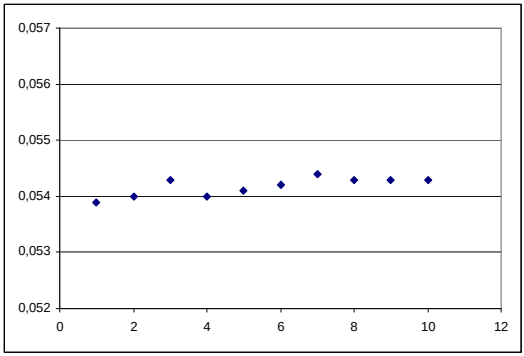



Abb. 8: Kumulierte Wiederholungsrate

Diese bisherige auf graphischen Darstellungen beruhende Interpretation lässt sich jedoch durch eine verfeinerte Untersuchungsmethode stützen. Zu diesem Zweck wird für jeden Datenpunkt die durchschnittliche absolute Abweichung von $RR_{1,2,3\ldots j}$ minus $RR_{10}$ berechnet:

$$K_x = \frac{1}{x}\sum_{i=i}^{x} \lvert RR_j - RR_{10} \rvert \ .$$

Das heißt, nach dieser Methode wird im Gesamtkorpus (Datenpunkt 10) der Wert 0 erreicht. Theoretisch müsste sich für $K_x$ in Relation zu den kumulierten Kapiteln eine fallende Kurve ergeben. Tatsächlich lässt sich graphisch zeigen, dass ein globaler Trend Richtung Null erkennbar ist. Sobald aber Kapitel 3 kumulativ[11] erfasst wird, ist wiederum die bereits beobachtete Bruchstelle zu beobachten.

[11] An dieser Stelle muss erinnert werden, dass für dieses Kapitel (Nr. 3) hinsichtlich

Dieser Bruch lässt sich auch in Abb. 9a ablesen. Eine erfolgreiche Modellierung der einzelnen kumulierten Kapitel und $K_x$ lässt sich allerdings nur dann erzielen, sofern der Datenpunkt Nr. 3 ausgeschlossen wird. In diesem Fall erreicht man mit einem einfachen linearen Modell y = -0.0000051603x + 0.0000462407 ein äußerst zufriedenstellendes $R^2$ = 0.9147. Dieses Modell und die empirischen Daten sind in Abb. 9b abgebildet. Unter Einschluss von Datenpunkt 3 ergibt sich allerdings nur ein $R^2$ von 0.6631. Daraus kann geschlossen werden, dass aufgrund des großen Absinkens des Determinationskoeffizienten (von 0.91 auf 0.66) der „richtige" Text identifiziert worden ist, und somit dieser als heterogen hinsichtlich des Verhaltens der Wiederholungsrate bezeichnet werden kann.

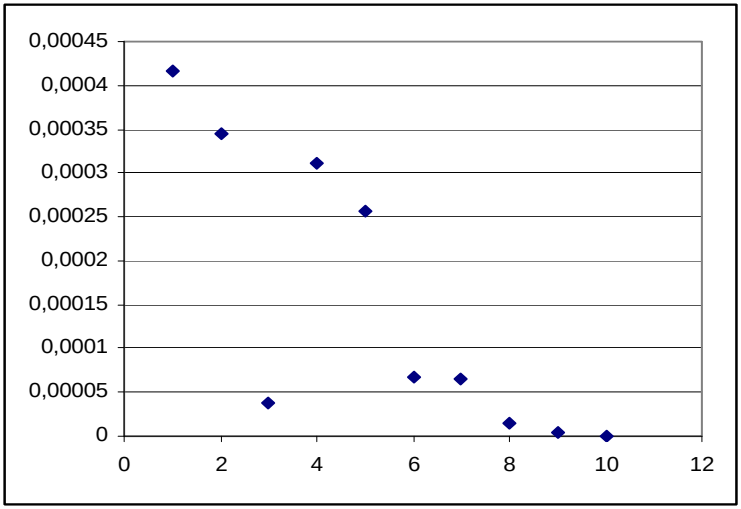



Abb. 9a: Kumulierte Texte vs. $K_x$ (alle Datenpunkte)

Diese vorgeschlagene Methode erweist sich demnach bei der Untersuchung von kumulierten Ranghäufigkeiten von Graphemen als durchaus „effektiv". Offen bleibt aber, ob auch höhere sprachliche Ebenen (man denke in etwa an Wortranghäufigkeiten) mit dieser

der Güte der Anpassung ein schlechteres Ergebnis festgestellt werden konnte, als für die restlichen Kapitel. Nunmehr zeigt sich auch, dass gerade dieser Text hinsichtlich der kumulierten Wiederholungsrate von den anderen Kapiteln abweicht.

Methode untersucht werden können, da in diesem Bereich keine geschlossenen Inventare vorliegen und die Wiederholungsrate a priori über eine höhere Schwankungsbreite verfügen müsste.

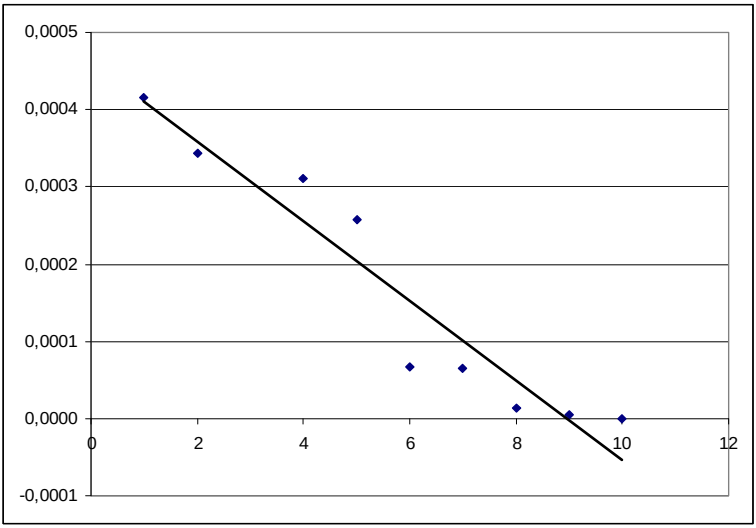



Abb. 9b: Kumulierte Texte vs. $K_x$ (ohne Datenpunkt 3)

## *4. ZUSAMMENFASSUNG*

Die Frage der Texthomogenität wurde in diesem Beitrag anhand einer vergleichenden Untersuchung der Graphemhäufigkeiten von mehreren Kapiteln aus einem Roman diskutiert. Darüber hinaus konnte das Wechselverhältnis von Teiltext vs. Ganzes (Summe einzelner Texte, Gesamtkorpus) thematisiert werden. Die Untersuchung zeigt, dass für das statistische Verhalten von Graphemhäufigkeiten – für diese sprachliche Ebene können abgeschlossene Inventare angesetzt werden – je nach Perspektive auf die Datenstruktur sowohl Homogenitäten als auch Heterogenitäten angesetzt werden können.

Betrachtet man die einzelnen Kapitel als abgeschlossene Entitäten so zeigt sich, dass sich beispielsweise die Wiederholungsraten darin nicht auffällig voneinander unterscheiden. Dies gilt auch für die Relation „Teiltext – Gesamtkorpus“. Ein homogenes Verhalten zeigen ebenfalls ausgewählte Parameter aus theoretischen stetigen Funktionen, die in der

Lage sind, die empirischen Graphem-Ranghäufigkeiten adäquat zu erfassen. Dieser Befund gilt sowohl für die Parameter der Kapitel als auch die des Gesamtkorpus.

Der Hirsch-Punkt (h-Punkt), der erstmals für Graphemranghäufigkeiten berechnet worden ist, bewegt sich in allen Texten innerhalb einer bestimmten Bandbreite (Intervall). Auch wenn weitere Studien in anderen Texten und Sprachen erfolgen müssten, kann die „relative" Konstanz des h-Punktes einstweilen als Hinweis auf Daten- und Texthomogenität gewertet werden.

Eine Stabilität der Werte konnte auch für die Indikatoren *c* und *p* nachgewiesen werden, die auf dem h-Punkt und der sogenannten Bogenlänge aufbauen. Da beide Indikatoren – die bislang hauptsächlich für Worthäufigkeiten berechnet worden sind – nicht mit der Länge der Kapitel (in Graphemen) in Zusammenhang stehen, sind sie ebenfalls geeignet, um Aufschluss über die Homogenität der einzelnen Kapitel zu geben.

Ein etwas anderes Bild lässt sich gewinnen, wenn man die Textdynamik untersucht. Dies kann dadurch bewerkstelligt werden, indem man das schrittweise Anwachsen, d.h. die Kumulation der einzelnen Kapitel, näher beobachtet. In diesem Fall kann – anhand einer ausgewählten Eigenschaft, nämlich der Wiederholungsrate – gezeigt werden, dass in bestimmten Kapiteln „Bruchstellen" zu lokalisieren sind. Dieser Befund steht allerdings nicht im Widerspruch zur allgemeinen Homogenität der Graphemhäufigkeiten in den einzelnen Kapiteln, sondern zeigt auf, dass es hinsichtlich des sukzessiven Anwachsens der Graphemhäufigkeiten offensichtlich immer wieder zu dynamischen Verschiebungen innerhalb der einzelnen Ranghäufigkeiten kommt. Diese Dynamik ist aber nicht so stark ausgeprägt, dass das globale Verhalten (Gültigkeit eines theoretischen Modells, Stabilität des h-Punkt usw.) nachhaltig beeinträchtig wird.

## *LITERATUR*

Altmann, G.; Lehfeldt, W. (1980): *Einführung in die Quantitative Phonologie.* Bochum: Brockmeyer. [= Quantitative Linguistics, 7]

Best, K.-H. (1994): Word class frequencies in contemporary German short prose texts, in: *Journal of Quantitative Linguistics* 1, 144-147.

Best, K.-H. (2006): *Quantitative Linguistik: Eine Annäherung. 3., stark überarbeitete und ergänzte Auflage.* Göttingen: Peust & Gutschmidt.

Grzybek, P.; Kelih, E. (2005b): „Towards a General Model of Grapheme Frequencies in Slavic Languages. In: Garabík, Radovan (ed.), *Computer*

*Treatment of Slavic and East European Languages.* Bratislava: Veda, 73-87.

Grzybek, P.; Kelih, E.; Altmann, G. (2004): „Häufigkeiten russischer Grapheme. Teil II: Modelle der Häufigkeitsverteilung", in: *Anzeiger für Slavische Philologie, 32*; 25-54.

Grzybek, P.; Kelih, E.; Altmann, G. (2005): „Graphemhaufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – eine Nebenbemerkung zur Diskussion um das 'ë'", in: *Anzeiger fur Slavische Philologie* 33, 117-140.

Guski, A. (1981): N. Ostrovskij „Kak zakaljalas' stal': biographisches Dokument oder sozial-realistisches Romanepos, in: *Zeitschrift fur slavische Philologie*, 42, 116-145.

Kelih, E. (2008): *Geschichte quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft.* Kovač: Hamburg. [= Studien zur Slavistik, 19]

Kelih, E. (2009a): Slawische parallele Texte als Datenbasis für systemlinguistische Untersuchungen: Projektvorstellung von „Kak zakaljalas' stal' (KZS)". [in diesem Band]

Kelih, E. (2009b): Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle, in: Glottometrics 18, 53-69.

Mačutek, J.; Popescu, I-I.; Altmann, G. (2007): Confidence intervals and tests for the h-point and related text-characteristics, in: *Glottometrics*, 15, 42-52.

Martináková, Z.; Mačutek, J.; Popescu, I-I.; Altmann, G. (2008): Some problems of musical texts, in: *Glottometrics* 16, 2008, 80-110.

Orlov, Ju.K.: (1982a): Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache-Rede" in der statistischen Linguistik). In: Orlov, Ju.K.; Boroda, M.G.; Nadarejšvili, I.Š. (eds.) (1982), 1-55.

Orlov, Ju.K. (1982b): Dynamik der Häufigkeitsstrukturen. In: Orlov, Ju.K.; Boroda, M.G.; Nadarejšvili, I.Š. (eds.) (1982), 82-117.

Orlov, Ju.K.; Boroda, M.G.; Nadarejšvili, I.Š. (eds.) (1982): *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer. [= Quantitative Linguistics, 15]

Popescu, I.-I.; Altmann, G., Köhler, R. (2009). Zipf's law. Another view. [im Druck]

Popescu, I.-I. (2006). The ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.): *Exact methods in the study of language and texts.* Berlin: Mouton-de Gruyter, 553-562.

Popescu, I.-I., Altmann, G. (2006): Some aspects of word frequencies, in: *Glottometrics* 13, 23-46.

Popescu, I.-I.; Altmann, G. (2008): On the regularity of diversification in language, in: Glottometrics, 17, 94-108.

Popescu, I.-I., Mačutek, J. Altmann, G. (2008): Word frequency and arc length, in: *Glottometrics*, 17, 18-42.

Popescu, I.-I.; Kelih, E.; Altmann, G.; Best, K-H. (2009). Diversification of the case, in: *Glottometrics*, 18, 32-39.

# Slawisches Parellel-Textkorpus: Projektvorstellung von „Kak zakaljalas' stal' (KZS)"

*Emmerich Kelih (Graz, Österreich*)

## *0. EINLEITUNG*

In der Korpuslinguistik, der Universalienforschung, der vergleichenden Sprachwissenschaft und Textlinguistik wird immer wieder die besondere Eignung von parallelen Texten für die linguistische Forschung hervorgehoben. Unter einem Paralleltext wird ein Text/Textausschnitt verstanden, der in mehreren übersetzten Varianten vorliegt (vgl. McEnery/Xiao/Tono 2006: 47ff; Lemnitzer/Zinsmeister 2006: 121ff; Teubert/Čermáková 2007: 73, die in diesem Zusammenhang von translation corpora sprechen).

Cysouw/Wälchli (2007: 95) führen den Begriff von „massively parallel texts" ein, worunter sie Vorhandensein von vielen und möglichst sprachtypologisch unterschiedlichen Übersetzungen ein und desselben Textes[1] verstehen. Ohne Zweifel sind parallele Texte – sei es nun mit oder ohne Annotation und Tagging – eine wertvolle Datenbasis für vielfältige sprachwissenschaftliche Analysen.

Auch im slawistischen Kontext erfreuen sich Parallel-Texte immer größerer Popularität. Erwähnt seien folgende Projekte mit Bezug zu slawischen Sprachen: Das älteste Parallel-Korpus dürfte „Multext East"

[1] Als Beispiele für derartige Textsammlungen werden u.a. die Berichte des Europäischen Parlaments (mit Übersetzungen in über 20 europäische Sprachen), die „Universal Declaration of Human Rights" (online in 329 Sprachen verfügbar), entsprechende Passagen aus der Bibel (ca. in 1300 Sprachen online verfügbar) und der „Index Translationum" der Unesco (Verzeichnis übersetzter Werke mit Angabe der technischen Verfügbarkeit) genannt. Weiters sind unterschiedliche Übersetzungen von „Le petit prince" und von „Harry Potter" digital verfügbar. Einigkeit scheint im Grunde darüber zu bestehen, dass biblische Texte (im übrigen eine beliebter Vergleichstext der historisch-vergleichenden Sprachwissenschaft) aufgrund der zum Teil ungeklärten Originalsprache, Autorenschaft usw. mehr oder weniger ungeeignet sind (vgl. dazu de Vries 2007 für vergleichende Studien).

sein, welches Übersetzungen von George Orwells „1984“ in das Bulgarische, Tschechische, Estnische, Ungarische, Rumänische und Slowenische beinhaltet (vgl. Erjavec et al. 1995, Dimitrova et al. 1998). Ambitiös ausgerichtet ist das „Regensburg Parallel Corpus of Slavic Languages“ (vgl. dazu Waldenfels 2006), das eine Reihe von unterschiedlichen Übersetzungen in und aus dem Russischen, Weißrussischen, Kroatischen, Polnischen, Serbischen, Slowakischen, Tschechischen und Ukrainischen (Verfassung der Europäischen Union, literarische Prosa von M. Bulgakov und S. Lem, u.v.m.) enthält.

Auffällig an den zuletzt genannten „großen“ Parallel-Text-Projekten ist, dass zwar für einen Text zwei oder mehrere Übersetzungen in slawische Sprachen existieren, aber offensichtlich bislang keine Textsammlung[2] vorliegt, die die Übersetzung eines Textes in alle slawischen Standard- bzw. Literatursprachen umfassen würde.

An dieser Stelle setzt das nunmehr vorzustellende Projekt ein: Vorrangiges Ziel ist der Aufbau eines Textkorpus, welches für eine Vielzahl von interessanten und bislang wenig untersuchten Fragestellungen der sprachlichen Selbstregulation (Köhler 1986, 2005) geeignet ist. Unter anderem geht es konkret um die Rolle des Einflusses des Umfanges von Phoneminventaren auf die Phonemhäufigkeiten und auf weitere, darüber liegende sprachliche Ebenen. Zu denken ist hierbei an Wechselbeziehungen zwischen dem Umfang von Phoneminventaren und phonotaktischen Gegebenheiten, die ihrerseits wiederum einen Einfluss auf die Silben- und Morphemstruktur einer Sprache ausüben können.

Um derartige Fragestellungen, die methodologisch wesentlich auf der Verwendung statistischer und quantitativer Methoden beruhen, verfolgen zu können, wurde vom Autor eine Sammlung von parallelen Texten aus 12 slawischen Standardsprachen (Weißrussisch, Ukrainisch, Russisch, Tschechisch, Polnisch, Slowakisch, Obersorbisch, Bulgarisch, Kroatisch, Makedonisch, Serbisch, Slowenisch) erstellt. Es handelt sich dabei um den russischen Roman „Kak zakaljalas' stal'“ (N.A. Ostrovskij)

[2] Für weitere Paralleltext-Sammlungen slawischer Sprachen, die in der Regel neben dem Original nur die Übersetzung in eine slawische Sprache beinhalten vgl. Waldenfels (2006: 123). Vgl. auch Garabík et al. (2007), die eine Sammlung (Wortlisten) aus westslawischen Sprachen vorstellen. Dieses Projekt ist auf eine bestimmte Art der Kindersprache fokussiert und beinhaltet weniger Übersetzungen, als vielmehr Textmaterial aus unterschiedlichen Lehrbüchern der gleichen Schulstufe. In Solz/Stroh/Urdze (2007) wird von einer Sammlung von Übersetzungen des „Le Petit Prince“ – darunter auch vielen slawischen Übersetzungen – berichtet.

aus den Jahren 1932-1934. Bevor diese Textsammlung im Folgenden näher vorgestellt wird, seien allgemein die linguistischen Einsatzmöglichkeiten, die Probleme der Erstellung, die Textauswahl und das Problem der Repräsentativität diskutiert. Im Anschluss daran werden erste Untersuchungsergebnisse vorgestellt, insbesondere im Hinblick auf den Textumfang (Anzahl der lexikalischen Tokens) und die durchschnittliche Wortlänge (gemessen in der Anzahl der Grapheme pro Wort).

## *1. PARALLEL-(TEXT)-KORPORA: LINGUISTISCHE EINSATZMÖGLICHKEITEN*

Aus linguistischer Sicht[3] sind Parallel-Texte eine wertvolle empirische Basis für die Untersuchung einer Vielzahl von theoretischen Fragestellungen. Zu denken ist u.a. an folgende Bereiche:

1. Sprachtypologische und „cross-linguistic" Untersuchungen der Phonologie, Morphologie, Syntax und Lexik (vgl. dazu Johannson 1998; Véronis 2000; Cysouw/Wälchli 2007, Wälchli 2007). Insbesondere hervorzuheben ist die quantitative Sprachtypologie (vgl. dazu Altmann/Lehfeldt 1973: 63-64), die ebenfalls an der Auswertung identischer Texte in mehreren Sprachen interessiert ist. Erste Anregungen zu quantitativen Analysen in diese Richtung finden sich in Stolz (2007), der ebenfalls die herausragende Bedeutung von Parallel-Texten für vergleichende quantitativ-linguistische Untersuchungen hervorhebt.
2. Translationswissenschaftliche Untersuchungen mit Fokus auf die Qualität und die Struktur von Übersetzungen (vgl. dazu Altenberg/Aijmer 2000; Johannson 2003; Gellerstamm 1996, Teubert 2002).
3. Hypothesen der quantitativen und synergetischen Linguistik können durch die vergleichende Untersuchung mehrerer Sprachen tiefgehende Einblicke in die sprachliche Selbstregulation von Sprachen und Texten geben (vgl. dazu allgemein Köhler 2005). Übersetzungen sind aus einer synergetischen Perspektive bislang nur wenig untersucht worden (vgl. Mohanty 2008). Offen bleibt die Frage, inwiefern für Übersetzungen ebenfalls allgemein

[3] Die technische und computerlinguistische Komponente der Erstellung von Parallel-Korpora wird an dieser Stelle bewusst nicht behandelt. Vgl. dazu den Sammelband zu Satz-Alignment, POS-Tagging usw. von Parallel-Korpora in Véronis (2000).

gültige quantitative Gesetzmäßigkeiten (wie z.B. das Zipf-Mandelbrotsche, Menzerath-Altmannsche Gesetz usw.) gelten.

4. Vergleichende quantitative textlinguistische Untersuchungen, unter Einschluss stilistischer, lexikologischer, syntaktischer und ähnlicher Fragen, können auf der Basis von Paralleltexten untersucht werden.
5. Zu denken ist auch an die – bislang wenig untersuchte – Probleme der Interkomprehension und der Text-Verständlichkeit.

Diese breiten Anwendungsmöglichkeiten sollen jedoch nicht darüber hinwegtäuschen, dass es einige gewichtige Argumente[4] gegen die Verwendung von Parallel-Texten zu geben scheint. Bemängelt wird zum Teil, dass es den übersetzten Texten an „Authentizität" (vgl. Stolz 2007: 102) mangelt. Dieses Problem wird unter dem Stichwort der „translationese" (vgl. Mauranen 2002, McEnery/Xiao/Tono 2006: 49) intensiv diskutiert. Allgemein wird auch auf die zum Teil fehlende sprachliche „Qualität" der Übersetzungen (Wälchli 2007: 128) verwiesen.

Abgesehen von diesen, im Detail erst zu untersuchenden, Problemen liegt der besondere Reiz in der Untersuchung von Parallel-Texten darin, dass eine direkte Vergleichbarkeit identischer Textpassagen aus unterschiedlichen Sprachen gegeben ist (vgl. Wälchi 2007: 130). Im Idealfall hat man es mit Texten und Textpassagen zu tun, die ein jeweils ähnliches „semantisches Feld" erfassen. In der Regel sind die Übersetzungen – abgesehen von allfälligen stilistischen Vorlieben und Interpretationen des Übersetzers – in der Regel zumindest grammatikalisch und morphologisch korrekt. Die Frage von stilistischen Abweichungen und sprachlichen Interferenzen kann, sofern man an einer empirischen Absicherung derartiger Phänomene interessiert ist, anhand von Parallel-Texten untersucht werden. In jedem Fall ist die Untersuchung von Parallel-Texten eine zentrale Möglichkeit für systematische vergleichende Studien in den unterschiedlichsten linguistischen Bereichen.

[4] Allgemein wird angeführt, dass der Originaltexte bzw. Übersetzungen immer nur eine Form der geschrieben Sprache wiedergeben. Thematisiert wird auch der hohe zeitliche und technische Aufwand für den Aufbau von umfangreichen Paralleltext-Sammlungen. Dies gilt allerdings auch für den Aufbau beliebiger Sprach- und Textkorpora, die sich auf schriftlich vorhandene Texte stützen.

## *2. „WIE DER STAHL GEHÄRTET WURDE“ IN ZWÖLF SLAWISCHEN STANDARDSPRACHEN*

Der Vergleich von Parallel-Texten ist keine prinzipiell methodologische Innovation. So werden beispielsweise in Lehr- und Handbüchern zur vergleichenden Grammatik (Phonetik/Phonologie, Morphologie) der slawischen Sprachen (vgl. Carlton 1990, Nahtigal 1961) zur Illustration jeweils kurze Passagen identischer Texte in den unterschiedlichen slawischen Standardsprachen präsentiert. Anhand ausgewählter Beispiele werden sodann Unterschiede zwischen den slawischen Sprachen aufgezeigt. So finden sich z.B. in der *Historischen Phonologie der slawischen Sprachen* von Carlton (1990) und Nahtigal (1961) Textfragmente aus dem Neuen Testament, in Kondrašov (1956) bzw. Mel'nyčuk (1966) Textausschnitte aus literarischen Romanen und Wortlisten mit dem Basis-Wortschatz der einzelnen slawischen Sprachen. Derartige Fragmente sind aber nur eine eingeschränkt taugliche Datenbasis für empirische, insbesondere jedoch statistische Untersuchungen, da diese in der Regel nicht mehr als 200 bis 300 Wörter umfassen, und somit für weit reichende Schlussfolgerungen zu kurz sind.

Für ein konkretes Forschungsprojekt zur vergleichenden quantitativen Phonologie (Phonemhäufigkeit, Phonotaktik, Silbenstruktur) der slawischen Sprachen wurde daher der Entschluss gefasst, ein slawisches Paralleltext-Korpus aufzubauen, welches eine möglichst große Anzahl von slawischen Standardsprachen umfasst. Die Entscheidung fiel auf den bekannten und vielfach übersetzten russischen Roman „Kak zakaljalas' stal' // Wie der Stahl gehärtet wurde“ (im folgenden KZS). Das Buch wurde von N.A. Ostrovskij (1904-1936) in den Jahren 1930-1934 auf Russisch verfasst und dem Stil nach dem sozialistischen Realismus[5] zuzuordnen ist.

Zu diesem Roman liegen für viele slawische Standardsprachen entsprechende Übersetzungen vor. Genauer gesagt für das Weißrussische, Ukrainische, Russische, Tschechische, Polnische, Slowakische,

[5] Ohne eine literaturwissenschaftliche Analyse durchführen zu wollen, sei auf folgendes verwiesen: Dieser autobiographisch geprägte Roman behandelt anhand des Schicksal von Pavel Korčagin die Geschichte der russischen Arbeiterbewegung am Ende des 19. Jhd.s bis in die 20 Jahre des 20. Jhd.s. Heute wird dieser Roman als zentrales Werk des sozialistischen Realismus gesehen und hatte bereits in den 60er Jahren die Auflage von 10 Millionen überschritten (in über 200 Auflagen). Der Roman ist in über 200 Sprachen übersetzt.

Obersorbische, Bulgarische, Kroatische (ijekavisch), Makedonische, Serbische (ekavisch) und das Slowenische. Besonderes Augemerk wurde darauf gerichtet, dass auch eine der sorbischen[6] Sprachen in einer Übersetzung vorliegt.

Allerdings sind an dieser Stelle einige problematische Punkte zu diskutieren: Erstens die nicht eindeutig geklärte Autorschaft, zweitens, die Rolle von editorischen Eingriffen, und drittens die unklare Ausgangslage hinsichtlich der verwendeten russischen Quell- und Originaltexte für die Übersetzungen. Die Autorschaft kann relativ eindeutig N. Ostrovskij zugeschrieben werden. Allerdings wird betont, dass an der endgültigen Fassung dieses Romans bis zu elf Redakteure gearbeitet haben sollen (vgl. Guski 1981: 121, ähnlich auch Anninskij 1989: 8), die offenbar diesen Text einem entsprechenden stilistischen bzw. politischen Schliff unterzogen haben.

Darüber hinaus – und dies hat bereits Guski (1981: 121f) detailliert aufgezeigt – ist die kanonische monographische Ausgabe aus dem Jahr 1934 gegenüber dem 1932 bis 1934 in Teilen in der sowjetischen Literaturzeitschrift „Molodaja Gvardija" veröffentlichten Roman nachträglich verändert worden. So wurden politisch verfängliche Textpassagen (der Hauptheld als Mitglied der ukrainischen Arbeiteropposition und ähnliches) ohne jeglichen Ersatz gestrichen. Dieser Eingriff betrifft – die Makrostruktur des gesamten Romans teilt sich auf je neun Kapitel in zwei Teilen – das neunte Kapitel des ersten Teiles und das erste Kapitel des zweiten Teiles. Diese Information ist, um damit zur Erstellung des Parallel-Textkorpus zurückzukommen, insofern von Bedeutung, da in der Regel die kanonisierte Textversion (also die monographische Ausgabe aus dem Jahr 1934) die Grundlage für die Übersetzung in alle anderen slawischen Sprachen darstellt. Angemerkt sei auch, dass es auch heutiger Sicht nicht immer möglich ist zu rekonstruieren, welche Originalvorlage für die Übersetzungen verwendet wurde, da großteils weder die Übersetzungsgrundlage (d.h. konkrete Ausgabe/Edition) noch der Übersetzer genannt werden[7].

[6] In van der Auwera/Schallea/Nuyts (2005) – eine der wenigen konkreten linguistischen Untersuchungen an slawischen Paralleltexten – wird ebenfalls hervorgehoben, dass die von ihnen untersuchten Übersetzungen des bekannten Kinderromans „Harry Potter" eben für das Sorbische nicht vorliegt.

[7] Auch wenn an den Texten bislang kein Alignment durchgeführt wurde, kann mit sehr hoher Wahrscheinlichkeit angenommen werden, dass keine der Übersetzungen der früheren Version entspricht, sondern vielmehr standardisierte Ausgaben verwendet wurden. Dies wurde anhand der Schlusspassagen des neunten Kapitels des ersten Teiles überprüft.

Details zu den von uns verwendeten Ausgaben der einzelnen Übersetzungen sind im Anhang 1 zu finden.

Trotz der problematischen Produktions- Editions- und Übersetzungsgeschichte des KZS ist der Roman aus sprachlicher Sicht – der literarische Wert hält sich der allgemeinen Einschätzung nach in Grenzen – durchaus interessant. Nach Guski (1981: 140) ist KZS durch einen einfachen, linearen Satzaufbau, durch die hohe Frequenz von umgangssprachlichen Lexemen und zum Teil durch einen pathetischen Sprachstil („obščestvennaja reč'") gekennzeichnet. Dem ist hinzuzufügen, dass zumindest der erste Teil des Romans durch einen hohen Anteil an mündlicher Rede, wenig narrative Sequenzen und durch die Verwendung typisch sowjetischer Kurz- und Abkürzungswörter (vor allem aus der politischen Lexik) auffällt. Des Weiteren – und dies ist im Hinblick auf die sprachliche Heterogenität wichtig – sind in den Fließtext vereinzelt Gedichte, Tagebuchaufzeichnungen, Briefe und offizielle Ankündigungen eingeflochten, die dazu dienen, die Authentizität des Romans zu unterstreichen. Insgesamt ist man mit einem Prosa-Text konfrontiert, der in gewisser Weise repräsentativ für die literarische Sprache des sozialistischen Realismus der 30er Jahre ist.

### *2.1. Zur Homogenität/Heterogenität des Romans*

Aufgrund des Mangels an slawischen parallelen Textsammlungen fiel die Entscheidung bewusst auf diesen russischen Text der 30er Jahre des 20. Jahrhunderts. Aus Sicht der Funktionalstilistik und Textsortenforschung handelt es sich dabei um einen „längeren" Roman (nach Eigenklassifikation „povest´"), der jedoch in unterschiedliche Teilkomponenten wie mündliche Rede, narrative Sequenzen, Briefe, Tagebuchaufzeichnungen, Gedichte, Ankündigungen usw. gegliedert werden kann.

Der Text ist nicht literarisch stark „deformiert", sondern dieser – und davon zeugt auch der (vermutlich eher politisch propagierte) „Erfolg" als Kinder- und Jugendbuch – stellt eine relativ „einfache" Sprache[8] dar. In Anbetracht der Mischung von Stilen (bzw. unterschiedliche Textarten) ist man zwar mit sprachlicher Heterogenität konfrontiert, gleichzeitig ist jedoch KZS ein abgeschlossenes Werk und genügt damit der bekannten Forderung nach der Untersuchung von ganzen, semantisch

[8] Eine vorläufige Analyse der durchschnittlichen Satzlänge in dem Roman und den Übersetzungen hat gezeigt, dass sich diese durchschnittlich etwa zwischen 8-10 Wörtern pro Satz bewegt.

abgeschlossenen Texten (vgl. Orlov 1982a, 1982b) bei quantitativen Untersuchungen.

### *2.2. Stichprobengröße und Repräsentativität*

Die Stichprobengröße wird üblicherweise als ein entscheidendes Problem korpuslinguistischer Untersuchungen genannt. In engem Zusammenhang damit steht die Frage nach der Repräsentativität des gesammelten sprachlichen Materials. Aus unserer Sicht ist es in diesem Zusammenhang sehr hilfreich, auf die ältere Diskussion zu verweisen, die um eine „adäquate" Stichprobengröße von Häufigkeitswörterbüchern in der sowjetischen Linguistik der 60er und 70er Jahre geführt wurde.

Die einzelnen Etappen dieser Diskussion müssen an dieser Stelle nicht dargelegt werden (vgl. dazu Kelih 2008). Es gibt – sehr grob gesprochen – u.a. zwei „Verfahren": (1) Die Absicherung des Stichprobenumfangs durch statistische Verfahren, wie beispielsweise in Altmann/Lehfeldt (1980: 122ff.) skizziert. Vor dem Hintergrund der Untersuchung statistischer Sprachmodelle sind derartige Verfahren, die zum Teil auf die Verwendung der Normalverteilung und ähnliches abzielen, plausibel und notwendig. (2) Die eigentlichen linguistischen Fragestellungen werden in den Vordergrund gestellt. In Abhängigkeit von den zu untersuchenden Hypothesen muss entschieden werden, ob man an der Untersuchung von ganzen Sprachkorpora, von Funktionalstilen, von Textsorten oder von individuellen Texten interessiert ist.

In unserem Fall versteht sich das KZS als empirisches Fallbeispiel für die Illustration sprachlicher Gesetzmäßigkeiten auf phonologischer und morphologischer Ebene. Darüber hinaus sind arbeitstechnische Möglichkeiten und Kapazitäten ausschlaggebend. Da – bis auf den russischen Originaltext – kein einziger Text elektronisch zur Verfügung stand[9] – fiel die Entscheidung schlussendlich auf die Untersuchung von zehn Kapiteln (aus insgesamt 18) des KZS. Es sind dies die ersten

[9] Auch wenn heute moderne Scanner und OCR-Programme (Abby-Finereader, PDFtoTXT usw.) zur Verfügung stehen, darf der Aufwand für die Erstellung dieser Textsammlung nicht unterschätzt werden (ca. acht Monate). Problematisch ist vor allem die Papierqualität der Übersetzungen, die in der Regeln aus den 50er und 60er Jahren (einzelne auch früher) stammen und zum Teil bereits vergilbt bzw. unscharf gedruckt sind. Erschwerend kommt hinzu, dass für das Weißrussische und Obersorbische keine Rechtschreibprüfprogramme zur Verfügung stehen und demnach die Korrekturarbeiten sehr viel Zeit in Anspruch nehmen.

neun Kapitel des ersten Teiles und das erste Kapitel des zweiten Teiles. Die nunmehr elektronisch vorhandene Textbasis umfasst ca. 220 Druckseiten und steht für unterschiedlichste automatische Analysen zur Verfügung. Abschließend sei darauf hingewiesen, dass KZS noch in weiteren 200 Übersetzungen vorliegt und somit in Zukunft eine Erweiterung auf nichtslawische Sprachen möglich ist.

## *3. QUANTITATIVE EIGENSCHAFTEN VON KZS*

Die folgende Darstellung ausgewählter quantitativer Texteigenschaften des KZS ist als vorläufig und tentativ zu verstehen. Da nach unseren Informationen bislang noch keinerlei Vergleichswerte zu quantitativen Merkmalen von Parallel-Texten in 12 slawischen Sprachen vorliegen, soll es an dieser Stelle hauptsächlich um eine Vorstellung einiger weniger selektiver Kenngrößen gehen. In Zukunft werden systematische Untersuchungen durchzuführen sein, in denen die entsprechenden synergetischen Hypothesen im Vordergrund des Interesses stehen. Eine erste vergleichende Untersuchung der Graphemhäufigkeiten des KZS findet sich in Kelih (2009).

### *3.1. Anzahl von Wortformen-Tokens*

Die erste wichtige Information betrifft den lexikalischen Stichprobenumfang der untersuchten Texte. Dieser wird in der Anzahl von Wortformen-Tokens gemessen. Als Kriterium wurde das orthographische Wort herangezogen. Der Bindestrich gilt als worttrennendes Zeichen. In der Tabelle 1 sind die entsprechenden Daten zusammengefasst. Die Sprachen sind bereits aus typologischer Sicht, genauer gesagt, aufgrund genetischer und arealer Merkmale (Südslawisch, Ostslawisch und Westslawisch) geordnet.

Bevor die insgesamt relativ hohe Spannbreite der Stichprobengrößen detaillierter besprochen wird, sei darauf hingewiesen, dass die Parallel-Texte und der Originaltext bislang keinem Satz-Alignment unterzogen wurden.[10] Es kann aber mit sehr hoher Wahrscheinlichkeit davon ausgegangen werden, dass alle Übersetzungen sehr nah am

[10] Für das Russische, Serbische und Slowenische wurde bereits eine Alignment-Prozedur (mit der Hilfe von Duško Vitas, Belgrad) gemacht. In Zukunft wird an eine vollständige Bearbeitung aller Texte zu denken sein.

russischen Originaltext sind und in allen Texten eine in etwa gleich große „semantische Menge“ transportiert wird.[11]

Tabelle 1
Anzahl von Wortformen-Tokens: Parallel-Textkorpus KZS

| Nr. | Sprachgruppe | Sprache | Tokens |
|---|---|---|---|
| 1 | Südslawisch | Slowenisch | 62655 |
| 2 | | Serbisch | 56230 |
| 3 | | Kroatisch | 56424 |
| 4 | | Bulgarisch | 57174 |
| 5 | | Makedonisch | 58837 |
| 6 | Ostslawisch | Russisch | 49675 |
| 7 | | Ukrainisch | 49612 |
| 8 | | Weißrussisch | 50010 |
| 9 | Westslawisch | Tschechisch | 52180 |
| 10 | | Slowakisch | 52099 |
| 11 | | Polnisch | 52737 |
| 12 | | Sorbisch | 58484 |

Folgende Beobachtungen sind von allgemeinem Interesse: Die ostslawischen Sprachen (Russisch, Ukrainisch, Weißrussisch) haben alle in etwa die gleiche Anzahl von Tokens. Der Unterschied von 63 (Ukrainisch) bzw. 335 Wörtern (Weißrussisch) im Vergleich zum Russischen ist als marginal anzusehen. Diese in etwa gleiche Textlänge kann einstweilen derart interpretiert werden, dass sich die untersuchten Sprachen (Ukrainisch, Russisch, Weißrussisch) hinsichtlich der lexikalischen Struktur und der Verwendung morphologischer Verfahren kaum unterscheiden. Im Grunde genommen können somit auch massive stilistische Modifizierungen der Übersetzungen ausgeschlossen werden.

Ein Blick auf die Textlänge der westslawischen Sprachen zeigt einen ähnlichen Befund. Für das Slowakische, Polnische und

[11] Folgende Analyse – die Details müssen hier nicht diskutiert werden – wurde durchgeführt: Für jeden übersetzten Roman wurden die Stichprobengröße pro Kapitel immer paarweise mit dem russischen Original verglichen. Die jeweiligen Zusammenhänge können mit einfachen linearen Modellen beschrieben werden und zeigen, dass sich – je nach untersuchter Sprache – die Stichprobengrößen der einzelnen Kapitel, jeweils systematisch um einen bestimmten Anteil verschiebt. Der Determinationskoeffizient ist bei allen Vergleichen mit dem russischen Originaltext bei $R^2 > 0.98$.

Tschechische kann in etwa eine gleich hohe Anzahl von Wortformen-Tokens ausgemacht werden (≈ 52300 Tokens). Besonders auffällig ist die gleich große Textlänge für das Slowakische und das Tschechische. Diese beiden Sprachen haben einen Unterschied von 81 Wortformen. Etwas höher fällt – beispielsweise in Relation zum Tschechischen und Slowakischen – die Anzahl von Tokens im Polnischen aus: Es sind hier immerhin 550 Wortformen mehr zu verzeichnen als im Tschechischen. Es wäre noch im Detail zu klären, ob stilistische oder andere Faktoren für diesen Unterschied verantwortlich sind.

Als klarer „Ausreißer" innerhalb der westslawischen Sprachen erweist sich jedoch eindeutig das Obersorbische, welches gegenüber dem Tschechischen ca. 5700 Wörter mehr hat. Diese Differenz ist nicht durch die Verwendung unterschiedlicher Quelltexte bzw. durch stilistische Vorlieben, sondern durch grammatische bzw. morphologische Faktoren zu begründen. Zu denken wäre an die analytische Bildung des Präteritums im Obersorbischen bzw. an potentielle Unterschiede im morphologischen Ausdruck der Reflexivität. In jedem Fall – und dies wird deutlich – eröffnet sich selbst bei einer so einfachen Frage wie der Analyse der Textlänge von Übersetzungen ein neues und weites Forschungsfeld.

Dass in der Tat morphologische Charakteristika der Sprachen für „große" Unterschiede in der Anzahl von Tokens des gleichen Textes in unterschiedlichen Sprachen verantwortlich sind, kann anhand der südslawischen Sprachen gezeigt werden. Die beiden eng verwandten Sprachen Kroatisch (56424 Tokens) und Serbisch (56230 Tokens) unterscheiden sich hinsichtlich der Anzahl von Tokens fast nicht. Die Unterschiede zwischen dem Makedonischen und dem Bulgarischen hinsichtlich der Anzahl von Tokens (das Makedonische hat 1663 mehr Tokens als das Bulgarische) könnten mitunter auf unterschiedliche Formen des Ausdrucks im Temporalsystem zurückgeführt werden. Dies müsste allerdings erst im Detail systematisch untersucht werden.

Als absoluter Spitzenreiter innerhalb aller Übersetzungen erweist sich jedoch das Slowenische mit 62655 Tokens. Dieser große Unterschied zu anderen slawischen Sprachen, insbesondere aber zum russischen Original kann durch eine Vielzahl von Faktoren erklärt werden: die analytische Bildung des Präteritums (mit dem Hilfsverbum „sein"), die intensive Verwendung des analytisch gebildeten Plusquamperfekts (je bil videl), die Neigung zur intensiven Nutzung von Relativsätzen und ähnliches. All diese Faktoren bewirken, dass insbesondere synsemantische Wörter aus morphologischen und syntaktischen Gründen

überdurchschnittlich häufig gebraucht werden und sich daher eine hohe Anzahl von Tokens ergibt. Weiter Erkenntnisse zu dieser Problematik können aber erst Studien zur Wortfrequenz liefern.

Es ergibt sich, dass bereits eine einfache vergleichende Analyse der Textlänge bereits zu durchaus interessanten Ergebnissen führt, und gleichzeitig eine Vielzahl von offenen Fragen zu Tage fördert. Es kann aufgrund der bisherigen Ergebnisse geschlossen werden, dass übersetzte Texte in der Regel zwar länger sind, aber in unserem Fall für dieses Phänomen vermutlich hauptsächlich sprachtypologische Gründe verantwortlich sind. In Abb. 1 findet sich eine graphische Darstellung[12] der unterschiedlichen Stichprobengrößen.

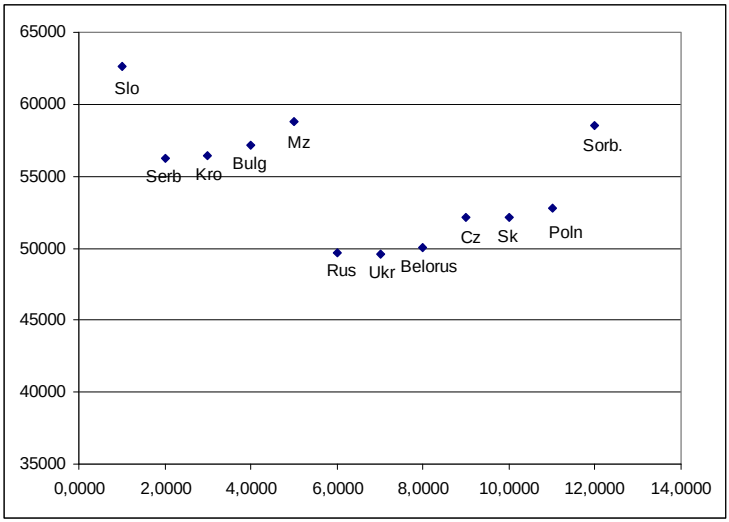



Abb. 1: Stichprobenumfänge des KZS (Wortformen)

[12] Folgende Abkürzungen werden verwendet: Slo = Slowenisch, Kro = Kroatisch, Serb = Serbisch, Bulg = Bulgarisch, Mz = Makedonisch, Sorb = Obersorbisch, Rus = Russisch, Ukr = Ukrainisch, Belorus = Weißrussisch, Cz = Tschechisch, Sk = Slowakisch und Poln = Polnisch.

### *3.2. Wortlänge als sprachtypologisches Merkmal?*

Die Wortlänge ist innerhalb der quantitativen Linguistik ein intensiv untersuchtes Merkmal (vgl. u.a. Grzybek 2006). Mit dem nunmehr vorliegenden Parallel-Korpus liegen erstmals vergleichende Angaben zur Textlänge in Wortformen-Tokens (N) für zwölf slawische Standardsprachen vor. Für eine erste Analyse der durchschnittlichen Wortlänge in slawischen Sprachen wird für alle Sprachen die Anzahl von Graphemen (NG) berechnet. Damit kann bereits die durchschnittliche Wortlänge in der Anzahl von Graphemen für jede einzelne Sprache berechnet werden. In vorliegenden Fall sollte die Wortlänge nicht als textsortenspezifisches Merkmal interpretiert werden, sondern als Kenngröße, die hauptsächlich Auskunft gibt über sprach- bzw. schrifttypologische Gegebenheiten der untersuchten Sprachen.

Auch wenn – aufgrund theoretischer Gründe, die hier nicht diskutiert werden müssen – üblicherweise die Wortlänge in Silben bzw. Morphemen zu messen wäre, ist die durchschnittliche Wortlänge in der Anzahl von Graphemen[13] eine erste interessante deskriptive Größe.

Diese erhaltenen Daten sind folgendermaßen zu kommentieren. Der Fokus liegt wiederum auf den einzelnen Sprachgruppen: Die ostslawischen Sprachen sind hinsichtlich der Wortlänge sehr kompakt und homogen. De facto unterscheidet sich das Ukrainische vom Weißrussischen nicht, während das Russische eine etwas höhere durchschnittliche Wortlänge in der Anzahl von Graphemen aufweist.

Ein ähnlich homogenes Bild ergibt sich für das Tschechische, Slowakische und Sorbische, die alle in etwa im Durchschnitt fünf Grapheme pro Wort aufweisen. Einzig das Polnische hat eine – in Relation zu denen anderen Sprachen – eine etwas höhere durchschnittliche Wortlänge von 5.53 Graphemen. Diese Abweichung lässt sich linguistisch leicht durch die vielen Digraphen in der polnischen Orthographie erklären. Bei der vorgenommenen Zählung wurden Digraphen als aus zwei Graphemen bestehende Einheit gezählt.

[13] Bei der Bestimmung der Anzahl von Graphemen wird ein Buchstabe des jeweiligen Alphabets gezählt, wie es in Referenzwerken zu den slawischen Sprachen (vgl. Rehder 1998, Comrie/Corbett 1993) angeführt ist. Anzumerken ist, dass bei der Zählung Digraphen, sofern sie auftreten (wie beispielsweise im Slowakischen, Kroatischen), in ihren Teilkomponenten (d.h. den sie konstituierende Buchstaben) gezählt werden. In Zukunft ist auch an eine phonologische Kodierung der Texte zu denken.

Tabelle 2
Wortlänge (in Graphemen) in zwölf slawischen Sprachen

| Nr. | Sprachgruppe | Sprache | Tokens (N) | Anzahl von Graphemen | Wortlänge in Graphemen |
|---|---|---|---|---|---|
| 1 | Südslawisch | Slowenisch | 62655 | 288879 | 4.6106 |
| 2 | | Serbisch | 56230 | 265344 | 4.7189 |
| 3 | | Kroatisch | 56424 | 269386 | 4.7743 |
| 4 | | Bulgarisch | 57174 | 276131 | 4.8297 |
| 5 | | Makedonisch | 58837 | 283510 | 4.8186 |
| 6 | Ostslawisch | Russisch | 49675 | 266055 | 5.3559 |
| 7 | | Ukrainisch | 49612 | 264283 | 5.3270 |
| 8 | | Weißrussisch | 50010 | 266239 | 5.3237 |
| 9 | Westslawisch | Tschechisch | 52180 | 258355 | 4.9512 |
| 10 | | Slowakisch | 52099 | 260707 | 5.0041 |
| 11 | | Polnisch | 52737 | 291979 | 5.5365 |
| 12 | | Sorbisch | 58484 | 297996 | 5.0953 |

Hinsichtlich der südslawischen Sprachen ergibt sich folgendes Bild: Die geringfügigen Unterschiede zwischen dem Kroatischen und Serbischen (4.77 vs. 4.71 Grapheme pro Wort) sind auf die kroatischen Digraphen (lj, nj, dž) zurückzuführen, die in unserer Zählung jeweils in ihren Teilkomponenten erfasst wurden. Im kyrillischen Serbischen stellen die Grapheme einen einzelnen Buchstaben dar. Zu bedenken ist auch die Schreibung des als jat-Reflex zu interpretierenden <ije> im Kroatischen, während im serbischen Text (ekavisch) dieses als <e> geschrieben wird. Die im Gegensatz zu den anderen südslawischen Sprachen sehr geringe Wortlänge des Slowenischen (im Durchschnitt 4,61 Grapheme pro Wort) kann einstweilen nur dadurch erklärt werden, dass im Text vermutlich eine hohe Anzahl von in der Regel kurzen Synsemantika vorkommt (analytische Bildung von Zeitformen: je, sem, usw.) und daher insgesamt die Wortlänge so kurz ausfällt.

Auffällig, aber linguistisch ebenfalls erklärbar ist die de facto gleiche Wortlänge im Bulgarischen (4.83) und Makedonischen (4.82), die etwas höher ist als im Kroatischen/Serbischen. Dies ist nicht weiter überraschend, sind doch beide Sprachen durch den Verlust von Deklinationsendungen und durch den Gebrauch des postpositiven Artikels gekennzeichnet. Abschließend wird die unterschiedliche Wortlänge in den slawischen Sprachen im KZS auch graphisch dargestellt. Vgl. Abb. 2.

In Zukunft ist an verfeinerte statistische Methoden (Diskriminanzanalysen, post-hoc Analysen) zu denken, um zu untersuchen, ob

sich auf der Basis der durchschnittlichen Wortlänge (unter Berücksichtigung weiterer Kenngrößen wie Standardabweichung, Variationskoeffizient usw.) Subgruppen feststellen lassen. Im Idealfall sollte dann natürlich die Wortlänge in der Anzahl von Phonemen, Silben oder Morphemen bestimmt werden.

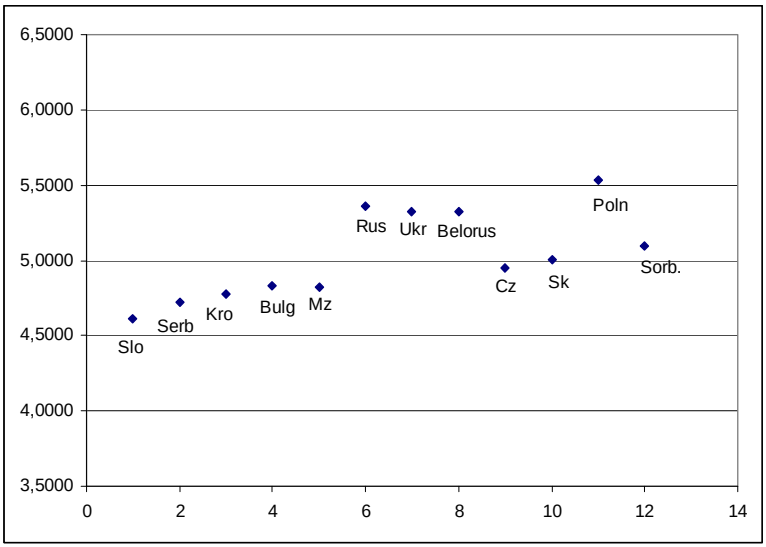



Abb. 2: Wortlänge im KZS-Korpus: 12 Slawische Sprachen

## *4. ZUSAMMENFASSUNG*

Die vorliegende Projektvorstellung ist nicht mehr und nicht weniger als eine erste Vorstellung des von uns erstellten Korpus von Parallel-Texten slawischer Standardsprachen. Es sollte deutlich geworden sein, dass die vorgelegten Texte sowohl aus textlinguistischer als auch sprachtypologischer Hinsicht von Interesse sind. In jedem Fall ist die Untersuchung von parallelen Texten ein Versuch, maximale Vergleichbarkeit des sprachlichen Materials zu erreichen.

Darüber hinaus zeigt sich, dass die quantitative Untersuchung von parallelen Texten im Grunde in ihren Anfängen steht. Bereits der Stichprobenumfang der Texte in den einzelnen slawischen Sprachen unterscheidet sich auffällig, wobei die aufgedeckten Unterschiede

durchaus plausibel durch sprachtypologische Merkmale erklärt werden können. Ähnliches gilt auch für die durchschnittliche Wortlänge in den untersuchten slawischen Sprachen.

## *LITERATUR*

Altmann; G.; Lehfeldt, W. (1973): *Allgemeine Sprachtypologie.* München: Fink. [= Uni-Taschenbücher, 250]

Altmann, G.; Lehfeldt, W. (1980): *Einführung in die Quantitative Phonologie.* Bochum: Brockmeyer. [= Quantitative Linguistics, 7]

Altenberg, B.; Aijmer, K. (2000): The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies. In: Mair, Ch.; Hundt, Chr. (2000): *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999.* Amsterdam [u.a.]: Rodopi, 15-33. [= Language and computers; 33]

Anninskij, L. (1989): Obručennyj s ideej. In: Ostrovskij, N.A. (1989): *Sobranie sočinenij v trech tomach. Tom 1. Kak zakaljalas' stal'.* Moskva: Molodaja Gvardija, 7-28.

van der Auwera, J.; Schallea, E.; Nuyts, J. (2005): Epistemic possibility in a Slavonic parallel corpus – a pilot study. In: Hansen, B.; Karlík, P. (eds.): *Modality in Slavonic languages. New perspectives*. München: Sagner, 201-17. [= Slavolinguistica, 6]

Carlton, T.R. (1990): *Introduction to the Phonological History of the Slavic Languages.* Columbus (OH): Slavica.

Comrie, B., Corbett, G.G. (eds.) (1993): *The Slavonic languages.* London/New York: Routledge.

Cysouw, M.; Wälchli, B. (2007): Parallel texts: using translational equivalents in linguistic typology, in: *Sprachypologie und Universalienforschung*, 60, 2, 95-99.

Dimitrova, L.; Ide, N.; Petkević, V.; Erjavec, T.; Tufis, D. (1998): Multext-East: Parallel and Comparable Corpora and Lexicons and Lexicons for six Central and Eastern European Languages. In: *Proceedings of the 36th annual meeting on Association for Computational Linguistics. Volume 1. Montreal/Quebec***,** 315-319.

de Vries; L. (2007): Some remarks on the use of Bible translations as parallel texts in linguistic research, in: *Sprachypologie und Universalienforschung*, 60, 2, 148-157.

Erjavec, T.; Ide, N.; Petkević. V.; Véronis, E. (1995): Multext-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. In: *Language Resources for Language Technology: Proceedings of the*

*TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995)*, o.O., 88-97.

Garabík R. (et al.) (2007): A Cross-linguistic Database of Children's Printed Words in Three Slavic Languages. In: Levická, J.; Garabík, R. (eds.) (2007): *Slovko 2007. Fourth International Seminar. Bratislava, Slovakia, 25-27 October 2007.* Tribun: Bratislava, 51-64.

Gellerstamm, M. (1996): Translations as a source for cross-linguistic studies. In. Aijmer, K.; Altenberg, B. and Johansson (eds.) (1996): *Language in Contrast: Papers from a symposium on Text based Cross-linguistic studies. Lund, March 1995.* Lund: Lund University Press, 53-62.

Grzybek, P. (2006): *Contributions to the science of language. Word Length Studies and Related Issues.* Dordrecht: Springer.

Guski, A. (1981): N. Ostrovskij: Kak zakaljalas' stal': biographisches Dokument oder sozial-realistisches Romanepos?, in: *Zeitschrift für slavische Philologie*, 42, 116-145.

Johannson, S. (1998): On the role of corpora in cross-linguistic research. In: Johannson, S.; Oksefell, S. (eds.) (1998): *Corpora and Cross-Linguistics Research*. Amsterdam: Rodopi, 3-24. [= Language and computers; 24]

Johannson, S. (2003): Reflections on Corpora and their Uses in Cross-linguistic research. Zanettin, F.; Bernardini. S.; Stewart, D. (eds.) (2003): *Corpora in translator education.* Manchester [u.a.]: St. Jerome Publisher, 135-144.

Kelih, E. (2008): *Geschichte quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft.* Kovač: Hamburg. [= Studien zur Slavistik, 19]

Kelih, E. (2009): Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle, in: *Glottometrics*, 18, 53-69.

Köhler, R. (1986): *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer. [= Quantitative Linguistics; 31]

Köhler, R. (2005): Synergetic linguistics. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.): *Quantitative Linguistik/Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin/New York: de Gruyter, 760-774. [= Handbücher zur Sprach- und Kommunikationswissenschaft, Band 27]

Kondrašov, N.A. (1956): *Slavjanskie jazyki.* Moskva.

Lemnitzer, L.; Zinsmeister, H. (2006): *Korpuslinguistik: eine Einführung.* Tübingen: Narr.

Mauranen, A. (2002): Will ‚translationese' ruin a contrastive study, in: *Languages in Contrast*, 2, 2, 161-186.

McEnery, T:, Xiao, R:, Tono, Y. (2006): *Corpus-based language studies: an advanced resource book.* London u.a.: Routledge.

Mohanty, P. (2008): The Semantic Differential Technique and Measurement of Translational Meaning. In: Altmann, G.; Zadorozhna, I.; Matskulyak, Y. (eds.): *Problems of General, Germanic and Slavic Linguistics. Papers for the 70th anniversary of Professor V.V. Levickij*. Chernivtsi: Knichi XXI, 215-225.

Mel'nyčuk, O.S. ed. (1966): *Vstup do porivnjal'no-istoryčnoho vyvčennja slov'janskych mov.* Kyïv: Naukova Dumka.

Nahtigal, R. (1961): *Die slavischen Sprachen.* Wiesbaden: Harrassowitz.

Orlov, Ju.K. (1982a): Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache-Rede" in der statistischen Linguistik). In: Orlov, Ju.K.; Boroda, M.G.; Nadarejšvili, I.Š. (eds.) (1982), 1-55.

Orlov, Ju.K. (1982b): Dynamik der Häufigkeitsstrukturen. In: Orlov, Ju.K.; Boroda, M.G.; Nadarejšvili, I.Š. (eds.) (1982), 82-117.

Orlov, Ju.K.; Boroda, M.G.; Nadarejšvili, I.Š. (eds.) (1982): *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer. [= Quantitative Linguistics, 15]

Rehder, P. (ed.) (1998): *Einführung in die slavischen Sprachen. (Mit einer Einführung in die Balkanphilogie).* Darmstadt: Wissenschaftliche Buchgesellschaft.

Stolz, Th. (2007): *Harry Potter* meets *Le petit prince*: On the usefulness of parallel corpora in crosslinguistic investigations, in: *Sprachypologie und Universalienforschung*, 60, 2, 100-117.

Stolz, Th.; Stroh, C.; Urdze, A. (2007): Nicht ganz ohne … In: Grzybek, P.; Köhler, R. (2006) (eds.): *Exact Methods in the Study of Language and Text. Dedicated to Professor Gabriel Altmann On the Occasion of His 75th Birthday.* Mouton de Gruyter: Berlin – New York, 633-646. [= Quantitative Linguistics, 62]

Teubert, W. (2002): The role of parallel corpora in translation and multilingual lexicography. In: Altenberg, B. and Granger, S. (eds.): *Lexis in Contrast.* Amsterdam: Benjamins, 189-214.

Teubert, W.; Čermáková, A. (2007): *Corpus linguistics: A short introduction.* London [u.a.]: Continuum.

Véronis, J. (2000). From the Rosetta Stone to the Information Society: A Survey of Parallel Text Processing . In: Véronis, J. (2000) (ed.): *Parallel Text Processing. Alignment and Use of Translation Corpora.* Dordrecht: Kluwer, 1-25. [= Text, Speech and Language Technology, 13].

Wälchli, B. (2007): Advantages and disadvantages of using parallel texts in typological investigations, in: *Sprachypologie und Universalienforschung*, 60, 2, 118-134.

Waldenfels von, R. (2006): Compiling a Parallel Corpus of Slavic Languages. Text strategies, Tools and the Question of Lemmatization in Alignment. In: Brehmer, B., Ždanova, V., Zimny, R. (eds.) (2006): *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9.* München, 123-138.

Anhang 1
Verwendete Textausgaben bzw. Übersetzungen

| Sprache | Ausgabe/Übersetzung |
|---|---|
| Slowenisch | Ostrovskij, N. (1945): *Kako se je kalilo jeklo.* Ljubljana: Mladinska knjiga. [Ü. von Vladimir Levstik] |
| Serbisch | Ostrovskij, N. (1949): *Kako se kalio čelik*. Beograd: Novo Pokolenje. [keine weiteren Angaben] |
| Kroatisch | Ostrovskij, Nikolaj (1945): *Kak se kalio čelik.* Zagreb. [keine weiteren Angaben] |
| Bulgarisch | Ostrovskij, N. (1976): *Kak se kaljavaše stomanata.* Sofia: Narodna mladež. [Ü. von Ljudmil Stojanov nach der Ausgabe von (1967): Moskva: Molodaja Gvardija.] |
| Makedonisch | Ostrovskij, N. (1988): *Kako se kaleše čelkiot*. Skopje: Detska radost. [Ü. von Kiril Koneski] |
| Russisch | Ostrovskij, N. (1966): *Kak zakaljalas' stal'.* Moskva. [dieser Text ist online auf lib.ru verfügbar] |
| Ukrainisch | Ostrovskij, M. (1974): *Jak hartuvalasja stal'.* Kyïv: Dnipro. [Ü. von Viktor Petrovs'kij nach der Ausgabe von (1966): Molodaja gvardija: Moskva.] |
| Weißrussisch | Ostroŭski, N. (1950): *Jak hartavalasja stal'.* Mins'k: Dzjaržaŭnae vydavectva BSSR. [Ü: von L. Gankin] |
| Tschechisch | Ostrovskij, N. (1951): *Jak se kalila ocel.* Praha: Státní Nakladatelství Dětské Knihy. [Ü. von Jarmila Wagsteinóva der Ausgabe von (1948): Moskva-Leningrad: Detskaja Literatura] |
| Slowakisch | Ostrovskij, N.A. (1966): *Ako so kalila oceľ.* Bratislava: Smena. [Ü. von Magda Takáčová] |
| Polnisch | Ostrowski M. (1974): *Jak hartowała się stal.* Wyd. 9. poprawione. Warszawa: Wydawnictwo Ministerstwa Obrony Narodowej. [Ü. von Wacław Rogowicz der Ausgabe von (1951): Sovetskij pisatel'. Moskva.] |
| Obersorbisch | Ostrowski, N. (1960): *Kak so wocl kowaše.* Budyšin: Ludowe Nakładnistwo Domowina. [Ü. von Jan Žur] |

# Project Description: Designing and Constructing a Typologically Balanced Ukrainian Text Database

*Emmerich Kelih (Graz, Austria),*
*Solomija Buk (Lviv, Ukraine),*
*Peter Grzybek (Graz, Austria),*
*Andrij Rovenchak (Lviv, Ukraine)*

## *0. INTRODUCTION*

The present contribution reports about a bilateral Austrian-Ukrainian research project. Starting with 2009, this project is financially supported by the Austrian Agency for International Cooperation in Education and Research (ÖAD), in the framework program "Scientific and Technological Co-Operation" and the Ministry of Education and Science of Ukraine (Project No. M/6-2009). Project partners are the Institute for Slavic Studies at Graz University (Emmerich Kelih, Peter Grzybek), the Institute for Theoretical Physics at Lviv University (Andrij Rovenchak), and the Institute for General Linguistics at Lviv University (Solomija Buk).

The major objective of this project is the design and construction of a typologically balanced text database for Ukrainian. With regard to future linguistic analyses, directed to the quantitative study of Ukrainian texts and language, special attention is paid to the inclusion of individual texts, rather than building a comprehensive corpus of heterogeneous language material. Particular attention is paid to the balanced inclusion of texts of various functional styles and text types, in order to cover the broad spectrum of text genres.

## *1. SHORT OVERVIEW: OBJECTIVES AND AIMS*

For any empirical and/or statistical analysis of language and text the availability of digitized text material (language corpora) is an indispensable precondition. Especially in the field of corpus linguistics great efforts have been made to develop and build text archives, text collections and

general language corpora. For Ukrainian[1], however, there are only a few projects pursuing this direction, and they shall be initially described here.

Two projects of general language corpora of Ukrainian are built independently at the Institute of Ukrainian Language (*National Corpus of the Ukrainian Language*) and at the Ukrainian Language-Information Fund (*Ukrainian National Linguistic Corpus*); both Institutes are parts of the National Academy of Sciences of Ukraine. Unfortunately, none of these two projects is available via Internet. *National Corpus of the Ukrainian Language* is planned as a balanced corpus, aiming to cover all text genres, and containing one million word occurrences (Dems'ka-Kul'čyc'ka 2005). The *Ukrainian National Linguistic Corpus* is designed as a non-balanced corpus, constantly being enlarged by texts from various domains (Šyrokov et al. 2005), and presently containing ca. 60 million word occurrences.

Some parallel corpora of Ukrainian are represented in the Internet. The project of the Ukrainian-Polish Corpus[2] is carried out at the Institute of Slavic Studies of the Polish Academy of Sciences (Warsaw). It covers 70 (2 × 35) texts, which are aligned at the paragraph level. The Ukrainian-Russian parallel corpus[3] of web publications uses automatic alignment by keywords.

Several projects from the domain of author lexicography are known for specific Ukrainian writers. The text corpus of Taras Shevchenko, a great Ukrainian poet and artist, was created in Canada for the compilation of the concordance of Shevchenko's poetry (Ilnytzkyj/Hawrysch 2001). The same team under the supervision of O. Ilnytzkyj is presently working on the text corpus of H. Skovoroda, a Ukrainian poet and philosopher. The corpus of works of V. Shevchuk, a Ukrainian writer, was recently designed as a research tool for the study of key concepts in his works (Monakhova 2006).

[1] To give an impression of the situation for other Slavic standard languages, in particular the following "bigger" corpora projects can be mentioned: "Slovenian Fida-Corpus" (http://www.fida.net/slo/index.html), "Hrvatski nacionalni korpus/ Croatian National Corpus" (http://www.hnk.ffzg.hr/), "Slovenský národný korpus/ Slovak National Corpus" (http://korpus.juls.savba.sk/index.en.html), "Nacional'nyj korpus russkogo jazyka/Russian National Corpus" (http://www.ruscorpora.ru); "Český národní korpus/Czech National Corpus" (http://ucnk.ff.cuni.cz), "Korpus Języka Polskiego IPI PAN /IPI PAN Corpus of Polish" (http://www.korpus.pl); "Korpus Języka Polskiego Wydawnictwa Naukowego PWN/ PWN Corpus of Polish" (http://korpus.pwn.pl); "Project PELCRA" (http://pelcra.ia.uni.lodz.pl) and many more.

[2] http://corpus.domeczek.pl.

[3] http://ling.infostream.ua.

The corpus of Ivan Franko's texts is planned to cover all his works (estimated size: ca. seven million word occurrences), representing the Western variant of the Ukrainian language from the turn of the 20th century (Buk 2007). The morphological tagging of some works has already been accomplished; thus far, this work has provided an online concordance[4] of *Perekhresni stezhky (The Cross-Paths)*, a novel by Ivan Franko; on this basis, some statistic features of Franko's text have successfully been studied (Buk/Rovenchak 2007a; 2007b).

What is problematic with such corpora is the fact that analyses done on their basis (collocation analysis, word form frequency, etc.), mostly ignore the specifics of individual texts, since the latter are not accessible as such.

From a quantitative linguistics perspective, however, the availability of semantically coherent and closed texts is of high importance and represents a primary interest. The analysis of individual texts can be seen as an attempt to minimize "internal" textual heterogeneity, characteristic of corpus analyses (cf. Altmann 1982; Kelih 2009, this volume). As Ju.K. Orlov has convincingly shown, particular quantitative regularities (e.g., the Zipf Law), hold true only for complete, "closed", semantically coherent texts; as compared to this, text mixtures (and any corpus is such a mixture!), represent some kind of a quasi text, for which a harmonious behavior of frequency and length behavior cannot be expected.

With regard to the fact that for many Slavic languages there are no systematically constructed database, from which individual texts would be available, work in this direction has started in 2002 in the framework of the Graz project on Quantitative Text Analysis (QuanTA), to fill this gap. Up to now a text database, containing individual texts form different text types and functional styles (for details see below) for Slovene, Croatian, Serbian, Russian and Slovak with over 6000 individual texts, has been collected and pre-processed. With the present project, concentrating on Ukrainian only, another East Slavic language shall be added. The extension to Ukrainian is important for various reasons: irrespective of the fact that such a desirable text database has been generally missing, far-reaching options and possibilities will be provided for literary, linguistic, stylistic etc. analyses, particularly in a comparative perspective, with regard to other Slavic languages.

[4] Available at: http://www.ktf.franko.lviv.ua/~andrij/science/Franko/concordance.html.

### *1.1. Detailed Project Description*

The project's program includes two major tasks:

1. Construction of a web based database of Ukrainian texts, on the basis of a typologically structured and balanced design;
2. Annotation and Tagging of the chosen texts, partly supported by relevant software.

In this context, it should be emphasized that the planned text database does not include the naïve claim for representativeness of Ukrainian as a literary or standard language. Representativeness is one of the greatest illusions in linguistics, not only in corpus linguistics. Rather the intention is to represent the broad spectrum of different text types and functional styles typical for a given language. Given the experience from intensive work on word length frequencies in various Slavic languages – see, among others, Grzybek/Kelih/Stadlober 2005, Grzybek/Stadlober/Kelih/Antić 2005, Kelih/Antić/Grzybek/Stadlober 2005) – it has turned out to be efficient, to motivate the selection of texts for the database with regard to their attribution to a particular text type and functional style.

### *1.2. Text Type and Functional Style*

From a theoretical point of view, the concepts of functional style and text type are of crucial relevance with regard to the structure of the planned text database:

a. With reference to the achievements of the Prague School and other research in this direction, qualitatively different functional styles can be distinguished. The distinctions are made on the basis of general communicative (socially defined) functions of language and texts. Usually, a diverging number from five to eight functional styles are postulated, depending on the concrete theoretical position taken.
b. With regard to empirical research in the domain of text types, the whole spectrum of text types can be understood as a maximum differentiation of communicative utterances. This involves an enormous amount of pragmatic-communicative text situations, and results in an inventory of ca. 4000 text types which are assumed to play a crucial role in a given culture or language (cf. the bibliographical data given by Adamzik 1995).

It is hardly possible, of course, to take into consideration all text types when constructing a balanced text database. Having in mind sys-

tematic statistical analyses, however, this is not even necessary; rather, it must be guaranteed that the whole spectrum of different stylistic shapes is adequately represented. In this respect, a combined selection of different text types from all functional styles seems to be a pragmatic solution. Table 1 represents a schema with all functional styles and specific characteristic text types within each of them; this schema may serve as a starting point for the intended construction of a typologically balanced text database.

It seems reasonable that the schema represented in Table 1 is an adequate and sufficient orientation, and can serve as a basis for quantitative and statistical analyses. As a *minimal condition*, two different text types (boldfaced in Table 1) will be chosen for each of the seven functional styles, each of them represented by 30 individual texts. This results in a minimal number of 420 texts ($2 \times 7 \times 30$) as a lower limit for systematic analyses.

### *1.3. Project Management and Work-Flows*

The course of the project includes two phases, each of them planned to cover one year.

**Phase I:** The first steps are directed at text acquisition and selection. These steps include the analysis and documentation of available sources (internet, CD collections, etc.). In particular, usage of online versions of newspapers[5] and of electronic news portals[6] is planned, as well as of official web- ages of the Ukrainian Parliament[7], some Ukrainian Churches[8], freely available academic journals, electronic libraries of the Ukrainian literature[9], etc. For some text types, it may be additionally necessary to process electronic texts by way of OCR software.

The next steps include the unification of different character encoding, the unified attribution of meta data (author, time, title, text type, functional style) for each single text.

**Phase II** primarily includes the preparation of tagging procedures, i.e. the text-specific treatment and/or annotation of headlines, chapter

[5] http://www.day.kiev.ua, http://www.zn.kiev.ua, http://www.wz.lviv.ua, etc.
[6] http://www.elvisti.com, http://www.unian.net, http://www.korespondent.net, http://pravda.com.ua, etc.
[7] http://www.rada.kiev.ua.
[8] http://www.uaoc.info, http://www.ugcc.org.ua, http://www.ugcc.lviv.ua, http://www.cerkva.info.
[9] http://poetry.uazone.net, http://www.ukrlib.com.ua, etc.

headings, abbreviations, numbers (years), foreign words, etc. which might influence text analyses unless carefully integrated. Additionally, a unified treatment and annotation of sentence boundaries, of direct and indirect speech, of quotations, etc. is necessary here. Finally, all the text have to be migrated to and incorporated into a central database structure; so they can be made publicly available to the academic community (given there are no copyright conflicts).

Table 1:
Functional Styles and Text Types

| everyday style | scientific | administration | journalistic | prose | poetry | drama |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| **Private Letter** | Abstract | Instruction | News Agency communication | Autobiography | Elegy | **Drama** |
| Diary Entry | **Articles Social sciences Natural sciences** | Business Letter | Report | Biography | Epos | Comedy |
| Joke | | Legal Text | Professional Article | Epistolary Novel | **Poem** | Tragedy |
| **Cooking Recipe** | Diploma Thesis | Expertise | Feuilleton | Epilogue | Ode | Drama in Verse |
| … … | Ph.D.Thesis | Political Convention | Comment | Memoirs | **Sonnet** | … |
| | Presentation | **Sermon** | **Column** | **Short Story** | | |
| | Review | Contract | Reader's Letter | Fable | Novel in Verse | |
| | Conference Report | **Open Letter** | **Sport Reportage** | Parable | | |
| | … | Resolution | Weather Report | Fairy Tale | … | |
| | | **Parliament speech** | … | **Novel** | | |
| | | … | | Legend | | |
| | | … | | Myth | | |
| | | | | Diary Novel | | |
| | | | | … | | |

## *2. RESULTS (EXPECTED)*

The first concrete project result will be the text database itself, providing an adequate material basis for systematic stylistic, linguistic, statistical, etc. text analyses. The typologically balanced design, the pre-processing of all texts, and the specific annotations are essential characteristics of this database. In fact, a systematically constructed and publicly available collection of Ukrainian texts will be available in electronic form, representing the necessary pre-condition for automatic, typologically specific analyses of Ukrainian texts. In the long run, these analyses will provide theoretical insight into the synergetic functioning of texts and languages, and with its focus on Ukrainian, it will contribute to a deeper understanding of language-specific factors.

As to future perspectives, one should first think, by way of an example, of analyses of the Ukrainian writing system (cf. Grzybek/Kelih 2005, Buk/Mačutek/Rovenchak 2008); in case an automatic grapheme-phoneme (letter-allophone) conversion will be available, a reasonable next step would be a (comparative) analysis of the Ukrainian phonological system and phoneme frequencies. Another reasonable field of application can be seen in lexical and syntactic studies, including word length, sentence length, and the relation between sentence length and word length (Arens Law, Menzerath-Altmann Law) in various discourse styles (cf. Buk/Rovenckak 2008, Grzybek/Kelih/Stadlober 2008). When then the technical conditions will be fulfilled, and the sources will be available, as has been described above, it will be easy to systematically pursue these questions and, to be sure, quite a few others …

## *REFERENCES*

Adamzik, K. (1995): *Textsorten – Texttypologie. Eine kommentierte Bibliographie.* Münster: Nodus.

Altmann, G. (1992): Das Problem der Datenhomogenität, in: *Glottometrika* 13, 105-120.

Buk, S. (2007): Korpus tekstiv Ivana Franka: sproba vyznačennja osnovnykh parametriv [Ivan Franko Text Corpus: an Attempt to Determine Main Parameters]. In: Shyrokov, V. A. (ed.): *Applied Linguistics and Linguistic Technologies: MegaLing-2006.* Kyiv: Dovira, 72-82.

Buk S.; Mačutek J.; Rovenchak A. (2008): Some properties of the Ukrainian writing system, in: *Glottometrics* 16, 63-79.

Buk, S. N.; Rovenchak, A. A. (2004): Rank-Frequency Analysis for Functional Style Corpora of Ukrainian, in: *Journal of Quantitative Linguistics* 11, 161-171.

Buk, S.; Rovenchak, A. (2007a): Statistical Parameters of Ivan Franko's Novel *Perekhresni stežky* (*The Cross-Paths*). In: Grzybek, P.; Köhler, R. (Eds.): *Exact Methods in the Study of Language and Text.* Berlin; New York, 39-48.

Buk, S.; Rovenchak, A. (2007b): Častotnyj slovnyk romanu Ivana Franka *Perekhresni stežky* [Frequency dictionary of Ivan Franko's Novel *Perekhresni stežky* (*The Cross-Paths*)]. In: Bacevyč, F. (Sci. Ed.); Buk, S. N.; Procak, L. M.; Rovenchak, A. A.; Svaryčevs'ka, L. Ju.; Cikhoc'kyj, I. L.: Stežkamy Frankovoho tekstu (komunikatyvni, stylistyčni ta leksyčni vymiry romanu *Perekhresni stežky*). Lviv University Press, 138-369.

Buk, S.; Rovenchak, A. (2008): Menzerath-Altmann Law for Syntactic Structures in Ukrainian, in: *Glottotheory* 1, 10-17.

Dems'ka-Kul'čyc'ka, O. (2005): *Osnovy nacional'noho korpusu ukraïns'koï movy* [Basics of the National Coprus of Ukrainian]. Kyiv: Institute of the Ukrainian Language.

Grzybek, P.; Kelih, E. (2005): Graphemhäufigkeiten im Ukrainischen. Teil 1 Ohne Apostroph (’). In: Altmann, G.; Levickij, V.; Perebyinis, V. (Eds.): *Problems of Quantitative Linguistics 2005*. Černivci: Ruta, 159-179.

Grzybek, P.; Kelih, E.; Stadlober, E. (2005): Empirische Textsemiotik und quantitative Texttypologie. In: Bernard, J.; Fikfak, Ju.; Grzybek, P. (Eds.): *Text & Reality*. Ljubljana/Wien/Graz: ZRC, 95-120.

Grzybek, P., Stadlober, E., Kelih, E., and Antić, G. (2005): Quantitative Text Typology: The Impact of Word Length. In: Weihs, C.; Gaul, W. (Eds.): *Classification – The Ubiquitous Challenge.* Springer, Heidelberg; 53-64.

Grzybek, P.; Kelih, E.; Stadlober, E. (2008): The Relation Between (Linguistic) Units of Different Levels: An Intra-Systemic Perspective, in: *Glottometrics* 16, 111-121.

Ilnytzkyj, O. S.; Hawrysch, G. (eds. and comps.) (2001): *Konkordancija poetyčnykh tvoriv Tarasa Ševčenka* [A Concordance to the Poetic Works of Taras Shevchenko]. 4 vols. New York: Shevchenko Scientific Society; Toronto: Canadian Institute of Ukrainian Studies Press.

Kelih, E. (2009): Zur Homogenität von Graphemhäufigkeiten in Texten: Evidenz aus dem Russischen. [in this volume]

Kelih, E.; Antić, G.; Grzybek, P. and Stadlober, E. (2005) Classification of Author and/or Genre? The Impact of Word Length. In: Weihs, C.; Gaul, W. (Eds.): *Classification – The Ubiquitous Challenge.* Springer, Heidelberg, 498–505.

Monakhova, T.V. (2006): *Korpus tvoriv Valerija Ševčuka: format rozmitky bazovykh konceptiv* [Corpus of Valerij Shevchuk works: the mark-up format for basic concepts], in: *Leksykohrafičnyj bjuleten'* 13, 26-29.

Šyrokov, V.A.; Buhakov, O. V., Hrjaznuchina, T. O. et al. (2005): *Korpusna linhvistyka*. [Corpus Linguistics]. Kyiv: Dovira.

# A Densitometric Classification of Web Template Content

*Christian Kohlschütter (Hannover, Germany)*

What makes template content in the Web so special that we need to remove it? In this paper I present a large-scale aggregate analysis of textual Web content, corroborating statistical laws from the field of Quantitative Linguistics. I utilize the fact that Web pages, as opposed to plain text, actually contain two classes of text, "templates" and "full text" and derive a simple yet effective local strategy to identify and remove template content.

## *1 INTRODUCTION*

In contrast to plain text, documents in the Web expose some specific structural properties. A Web page's text is not limited to the actual "main content" but usually consists of several additional segments. These provide further information (a site map, a list of news headlines, a table of contents, links to related information, text-ads etc.) and are only meant to augment the full-text (Gibson et al., 2005). This additional information provided seems only be partially useful or probably even counterproductive for search and classification; the common solution to the problem is simply erasing template content or at least ignoring it. However, current approaches tackle the problem of template identification only heuristically or by means of machine learning. In this paper, the problem of template detection is approached from a Quantitative Linguistic perspective.

Contributions. (1) The structure of a large, representative Web corpus is analyzed, utilizing the segment-level text density metric presented in (Kohlschütter/Nejdl 2008). Through a densitometric classification, we see that Web content exposes two classes of text, covering full-text and navigational information respectively. (2) I show that this structure corroborates recent findings from the field of Quantitative Linguistics. (3) The findings are applied to template removal.

Organization. Section 2 describes the related work. Section 3 refers to the theoretical background of Quantitative Linguistics and its relation to the Web. In Section 4, we examine the shape patterns on corpus-level. The application of these findings for the purpose of template removal is evaluated in Section 5. Section 6 concludes with an outlook to future work.

## *2 RELATED WORK*

In this Section, I refer to the related work from the perspective of Information Retrieval only; related work from Quantitative Linguistics is described in Section 3.

The detection and removal of templates seems currently to be focused on exploiting DOM-level (*Document Object Model*) features or identifying quantitatively common (= frequently used) segments or patterns/shingles on a website (Bar-Yossef/Rajagopalan 2002; Gibson et al. 2005; Debnath et al. 2005; Vieira et al. 2006; Chen et al. 2003). Using a mixture of approaches, Gibson et al. quantified the amount of template content in the Web (40%-50%) (Gibson et al., 2005). Chakrabarti et al. determine the "templateness" of DOM nodes by a classifier based upon regularized isotonic regression (Chakrabarti et al. 2007). Yi et al. simplify the DOM structure by deriving a so-called Site Style Tree which is then used for classification (Yi et al. 2003). Baluja (Baluja 2006) employs decision tree learning and entropy reduction for template detection. Template detection is strongly related to the more generic problem of web page segmentation, which has been addressed at DOM-level (Chakrabarti et al. 2008), by exploiting term entropies (Kao et al. May 2005) or by using Vision-based features (Cai et al. 2003). Recently, Kohlschütter et al. exploited the statistical properties of text to identify page-level segments (Kohlschütter/ Nejdl 2008). As shown by Cai et al. (Cai et al. 2004) and more recently by Fernandes et al. (2007) the frequency of common segments can also be used to improve keyword-based search.

## *3 THEORETICAL BACKGROUND*

### *3.1 Quantitative Linguistic Text Theory*

Several observations have been made which corroborate the theory that natural language obeys the same principles as many other psycho-biological and natural phenomena, namely the class of power laws (Naranan/Balasubrahmanyan, 2005). George K. Zipf pioneered this model by his principle of least effort, which he said was inherent in human behavior (Zipf 1949). Numerous empirical observations confirm the hypothesis that the creation process of language, in particular text (spoken or written language), follows particular probabilistic regularities, which have been subsumed by statistical laws, in particular Zipf's law (the frequency of an object, e.g. a term, is inversely proportional to its rank), Frumkina's law (when dividing text into passages of words, the frequency of a particular linguistic entity follows the negative hyper-geometric distribution) and the Menzerath-Altmann law (the longer a linguistic construct, the smaller its constituents). The organization of text has been observed and successfully modeled statistical as urn trials at the level of various linguistic units such as phoneme, word, sentence, text segment etc. and for several features such as frequency, length, repeat rate, polysemy and polytextuality.

The levels of language seem to be strongly interdependent (cf. Menzerath-Altmann law). Köhler modeled this system as the so-called synergetic language control circuit and showed that it seems applicable to any linguistic level or aspect (Köhler 2005). He postulated language system requirements, amongst others the requirements of secure/reliable information transfer, leading to redundancy, and the requirements of economy, incorporating the principle of least effort, with its aspects like minimization of effort for encoding, decoding, memory capabilities/context-independence, ambiguity and so on. Köhler found that the system requirements mutually influence the variability of the system's pro-perties in cooperating as well as in competing ways; considering Zipf's theories, these requirements may also be called synergetic "forces" (Köh-ler 1990, 2005).

Any attempt to corroborate the established laws and models (or possibly to reject them) requires a quantitative, empirical analysis. Quantification is really not the aim, but a means to understanding the structures and processes of text and language (Grzybek 2006). The required statistical analysis has to be performed using an appropriate

text or corpus, otherwise one would neglect/hide the language-immanent heterogeneity (Altmann 1992). Once a representative baseline corpus is established, further analytical explorations can be attempted such as stylometric approaches to assign (with a given probability) a particular author to a specific document (or to exclude an author from consideration), a genre (newspaper text, political statement, scientific work etc.) or a readability score (e.g., boulevard news vs. legal articles) to a particular article, using scores like type-token-ratio, verb-adjective-ratio, vocabulary richness and so on (Tuldava 2005; Tweedie 2005; Popescu/Altmann 2006).

If one is able to closely fit a previously discovered distribution (e.g., negative hypergeometric, hyperpascal, negative binomial etc.) to the data, this contributes to corroborating the theory. Recently, Wimmer and Altmann presented a unified representation of many existing linguistic hypotheses (Wimmer/Altmann 2005; Grzybek 2006), a logical extension of Köhler's synergetic approach. They derive a common representation of the aforementioned distributions and relations by discussing the relation between a linguistic variable Y and another independent variable X which shapes the behavior of Y (i.e., also its rate of change, dx, which in turn is controlled by the aformentioned synergetic forces). Relations between X and Y for example are: polytextuality/polysemy, polysemy/length and also rank/frequency.

The relationship between X and Y can be seen as an infinite series of the form

$$\frac{dy}{y} = (a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \frac{a_3}{x^3} + \cdots)\, dx \qquad (1)$$

(with $a_0$, $a_1$, $a_2$,… being constant factors of the acting forces). The solution of 1 yields

$$y = C\, x^{a_1} e^{-a_0 x} exp(-\sum_{i=1}^{\infty} \frac{a_{i+1}}{x^i}) + d \qquad (2)$$

(with C and d being normalization parameters) which actually is a generalization of the commonly used form of the Menzerath-Altmann law

$$y = C\,x^{a_1} + d \tag{3}$$

his regularity was also discussed for discrete variables, in particular non-negative probability distributions with probability mass functions $\{P_0, P_1, \ldots\}$ of the form $P_x = g(x)P_{x-1}$. The discrete equivalent of the continuous model (Equation 1) is:

$$P_x = \left(a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \frac{a_3}{x^3} + \cdots\right) P_{x-1} \tag{4}$$

From this recurrence formula many well-known distributions observed in the field of linguistics can be derived, including the Katz/ Kemp-Dacey-hypergeometric families of distributions (Wimmer /Altmann 1999), whose limiting cases are (amongst others) the geometric, the Poisson, the hyperpascal and the negative-hypergeometric (including its limiting cases binomial and negative-binomial) distributions; all of them have already been discussed and empirically found for particular linguistic units.

### *3.2 Relation to the Web*

It would be surprising if the findings made on "plain text" would not be valid for text on the Web. In Kohlschütter/Nejdl (2008), it was shown that the discussed laws can be applied successfully to segment web pages into blocks of text. For conducting the segmentation, the block-level text density measure $\varrho(b)$ was introduced, derived from the pixel-based text density of Computer Vision-based approaches and transformed to the *token* level. Basically, it counts the number of tokens $|T(b)|$ in a particular text block b divided by the number of lines $|L(b)|$ covered after word-wrapping the text at a fixed column width $w_{max}$ (the empirically estimated optimal value for English text is between 80 and 90 characters). Due to the side effect of having an incompletely filled last line after wrapping, the latter is not taken into consideration unless it is the only line in the segment:

$$T'(b) = \left\{ t \,|\, t \in T(l),\ l_{first}(b) \le l < l_{last}(b) \right\}$$

$$\varrho(b) = \begin{cases} \dfrac{\left|T'(b)\right|}{\left|L(b)\right| - 1} & \left|L(b)\right| > 1 \\ \left|T(b)\right| & \text{otherwise} \end{cases} \tag{5}$$

The actual segmentation algorithm presented in Kohlschütter/ Nejdl (2008) is based on a merge-only strategy called Block Fusion. Adjacent text fragments of similar text density (interpreted as "similar class") are iteratively fused until the blocks' densities (and therefore the text classes) are distinctive enough. Using various settings, including a rule-based approach, it was shown that the resulting block structure closely resembles a manual segmentation.

Even though text density was derived from concepts of Computer-Vision, it appears that the exposed behavior of ϱ(b) in text is similar to existing linguistic measures. In particular, the ratio between the text densities of neighbored blocks follows the negative-hypergeometric distribution, corroborating Frumkina's law (Kohlschütter/Nejdl 2008). Further details about the density measure are discussed in Section 4.2.

## *4 CORPUS-LEVEL PATTERN ANALYSIS*

### *4.1 Setup*

The analysis is conducted on the Webspam UK-2007 test collection[1], a representative crawl of the .uk web graph with about 106 million pages. The collection has already partially been classified into spam and non-spam pages. Since spam pages tend to be automatically generated, may not necessarily obey the laws of natural language and could skew our results, let us focus on the non-spam part consisting of 356,437 pages with 316,448 documents containing extractable text (this was also the corpus used in Kohlschütter/Nejdl (2008)).

Since we are trying to understand the distinction between templates and main content, a statistical classification is performed on segment-level, under the assumption that each segment is sufficiently

[1] http://www.yr-bcn.es/webspam/datasets/uk2007/

homogeneous (i.e., either template or main content). As a manual segmentation appears infeasible at corpus-scale, the state-of-the-art BlockFusion segmentation algorithm is employed, BF-RuleBased in particular, which was shown to have a segmentation accuracy in terms of normalized mutual information (NMI) of 0.87 (Kohlschütter/Nejdl 2008) (BF-RuleBased is the most effective variant of the BlockFusion family; the segmentation boundaries are usually at the HTML block-level elements H1-H6, UL, DL, OL, HR, TABLE, ADDRESS, HR, IMG, SCRIPT but they may also exist between two neighbored segments which expose noticeably different text densities.

### *4.2 Density vs. Token Length*

Text density is a particularly useful measure when analyzing the Web's quantitative structure. It does not depend on the notion of "sentence", which one could hardly define for the Web's content – many portions of text simply do not contain sentences, nor anything meaningful which could be separable by full stop (this is especially true for template text). As for the text density a relationship to existing linguistic measures was already shown (cf. Kohlschütter/Nejdl 2008), we may assume that $\varrho(b)$ indeed is an adequate linguistic measure, too. Under the aspect that many linguistic measures obey the principles expressed by the Menzerath-Altmann law, we verify whether this law also holds for the text density. Actually, text density seems to follow this law per definitionem: the higher a text density, the shorter con-tained tokens must be. Thus, a strong relationship to the average token length is likely.

First, let us analyze the measures "average token length" and "text density" separately. For both measures, we compute per-document averages, normalize the scores to a maximum of 1.0 and sort them in decreasing order. Finally, Equation 3 is fitted to them. We may quantify the goodness of the fit by the correlation coefficient $R^2$ (the square of the correlation between the response values and the predicted response values) and the root-mean-square error RMSE. Indeed, for both measures – *average* token length and text density – we observe a high correlation. For average token length, we achieve $R^2 = 0.9335$; RMSE = 0.00002 with $a_1 = 0.51$, $c = 4.096 \cdot 10^{-5}$, $d = -1$. For the text density, with $a_1 = 7.5 \cdot 10^{-7}$, $c = 2.79 \cdot 104$, $d = -1$ the goodness of fit is $R^2 = 0.9654$, RMSE = 0.0154. Of note, in order to fit the rank sequence for average token length, the top 100 documents that had a very high average score had to be omitted; the skewness was caused

by very long tokens (the largest average token length encountered was 65.026). Second, let us analyze the ratio of text density to the average token length. As above, we normalize and sort the values; we then fit Equation 2 to them. With the parameters $a_0 = 0$, $a_1 = 0.024$, $a_2 \cdots a\infty = 0$, $c = 0.712$, $d = -0.98$, the resulting goodness of fit is $R^2 = 0.97$, RMSE = 0.0029. The three rank sequences are depicted in Figure 1.

As a conclusion, it seems that the text density measure (in combination with a good segmentation strategy) can be well integrated into the established theories. Moreover, text density appears to be less susceptible for noisy data than the average token length.

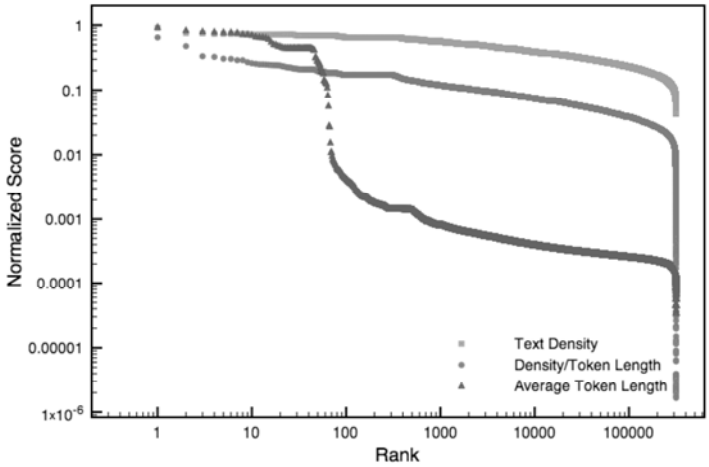



Figure 1: Text Density / Token Length Ranks

### *4.3 The Beta Distribution Model*

To reduce the impact of errors caused by a too fine-grained segmentation, the amount of text (= number of tokens) contained in segments of a particular text density $\varrho$ is examined. We can model this histographically by rounding the density to the nearest integer $\varrho'(b) = [\varrho(b)]$. According to Wimmer/Altmann (2005) switching between a continuous and a discrete representation as needs arise is valid under these circumstances, also see Equations 1 and 4.

Figure 2 depicts the retrieved token-level count/density distribution for the whole corpus. Apparently, two modal scores are visible, at $\varrho' = 2$ and $\varrho' = 12$ respectively. This indicates at least two classes of

text within the corpus. The superimposition of different classes ("strata") of text is already known in linguistics, from a theoretical perspective it may even be the normal case, even though empirically a separation may not seem necessary (Altmann 1992). To confirm the presence of multiple classes we need to find a corresponding distribution function.

As we have two visible modal scores, the distribution function that is to be retrieved is expected to be a combination of two Beta distributions:

$$c\cdot\left[p_1\cdot f_{beta}(x,a_1,b_1)+p_2\cdot f_{beta}(x,a_2,b_2)\right] \tag{6}$$

$$f_{\text{beta}}(x,a,b) \;=\; \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$$

$$B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}=\int_0^1 t^{a-1}(1-t)^{b-1}\,dt \tag{7}$$

$$\Gamma(x)=\int_0^\infty t^{x-1}\mathrm{e}^{-t}\mathrm{d}t = x\cdot\Gamma(x-1)=(x-1)!$$

$$x\in[0;1],\;\; x=\varrho\cdot\frac{1}{2}w_{max}$$

However the fit was unsatisfactory. I was able to fit a curve using Equation 6 ($p_1 = 0.65$; $p_2 = 0.55$) but no distribution (i.e., with $p_1 + p_2 = 1$). A combination of three Beta distributions yields a fairly good distribution fit ($p_1$, $p_2$, $p_3 = 1/3$; $a_1 = 64.08$, $b_1 = 147.9$; $a_2 = 2.596$, $b_2 = 32.33$; $a_3 = 10.7$, $b_3 = 30.45$; $c = 0.025$ with $R^2 = 0.944$, RMSE = 0.0031):

$$f(x)=c\cdot\left[p_1\cdot f_{beta}(x,a_1,b_1)+p_2\cdot f_{beta}(x,a_2,b_2)+p_3\cdot f_{beta}(x,a_3,b_3)\right] \tag{8}$$

However, we can (and therefore must) further simplify the distribution to a combination of two beta distributions and the normal distribution, with which we achieve an almost perfect fit ($R^2 = 0.998$, RMSE = 0.0021 for $a_1 = 68.03$, $b_1 = 132.5$; $a_2 = 4.034$, $b_2 = 54{,}49$; $c = 0.015$, $d = 0.64$; $e = 78.87$, $f = 7.834$, $\mu = 28.65$, $\sigma^2 = 6.489$; x-scores (densities) have been normalized by $x_{norm} = 36$ to [0 : 1] before fitting):

$$f(x)=c\cdot\left(d\cdot f_{beta}(x,a_1,b_1)+(1-d)\cdot f_{beta}(x,a_2,b_2)\right)+(1-c)\cdot\varphi_{\mu,\sigma^2}(e\cdot x+f) \quad (9)$$

The parameters $a_1$, $a_2$, $b_1$, $b_2$ define the skewness and the location of the mode of the Beta distribution, c and d are distribution weights, e and f are normalizing constants.

I chose the Beta distribution for several reasons. First, the Beta distribution is very generic in the sense that it allows a parameterization of the curve's skewness both to the left (a < b) and to the right (a > b), having the mode at $(a-1)\cdot(a+b-2)^{-1}\cdot x_{norm}$. These parameters seem necessary in our case (see the varying token counts for $\varrho' = [1;3]$ and $\varrho' = [10;12]$ respectively). Second, the distribution is continuous, which allows us to describe the probability of tokens covered by a particular density $\varrho$, not only the approximation $\varrho'$. Third, it is already known in Quantitative Linguistics. In particular, Altmann/Burdinski (1982) have derived the discrete negative hypergeometric (or: Beta-binomial) distribution using it, which in turn follows the Menzerath-Altmann Law. Of note, $f_{beta}$ has also been applied to histogram-based image segmentation (Al-Saleh/El-Zaart 2007), which is a remarkable fact because the text density metric and the accompanied BlockFusion algorithm inherit the notion of density as well as the block-merge strategy from Computer Vision, too (cf. Kohlschütter/Nejdl 2008).

I conclude that the distribution of text densities can be divided into two fuzzy classes $C_1$ and $C_2$; the transition from $C_1$ to $C_2$ follows the normal distribution, which means that for blocks with particular densities it is rather undetermined to which class the contained text belongs. Moreover, from the distribution parameter d we see that $C_1$ roughly covers one third of the tokens enclosed in the corpus and $C_2$ covers two thirds. Figure 2 depicts this fit as well as its three parts; we see that for $5 \leq \varrho' \leq 10$ the normal distribution dominates. Notably, these classes are not visible at token level (see Figure 3); the average token length appears to follow the (unimodal) Beta distribution $y = c \cdot f_{beta}(x / x_{norm}, a, b)$ with a = 40.53, b = 2612, c = 0.002458, $x_{norm}$ = 358 ($R^2$ = 0.9876, RMSE = 0.003). This supports the assumption that text density and average token length are measures at different linguistic levels.

Generally, the Beta distribution seems to fit very well to linguistic data. For example, I was able to fit it to the distribution of English sentence lengths in the standard Brown Corpus[2] with $R^2$ = 0.996, RMSE = $8.87 \cdot 10^{-4}$ having a = 1.926, b = 6.356, c = 0.013, $x_{norm}$ = 79.

[2] http://people.scs.fsu.edu/~burkardt/m_src/prob/english_sentence_length_pdf.m

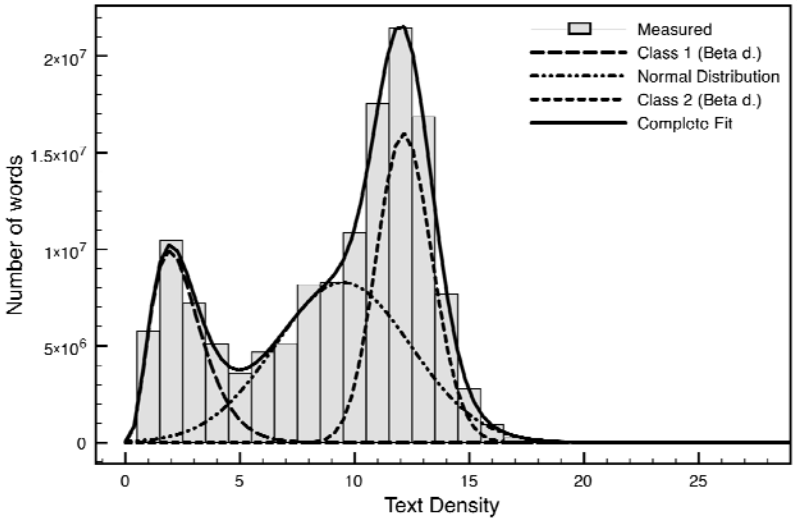



Figure 2: Density Distribution Model

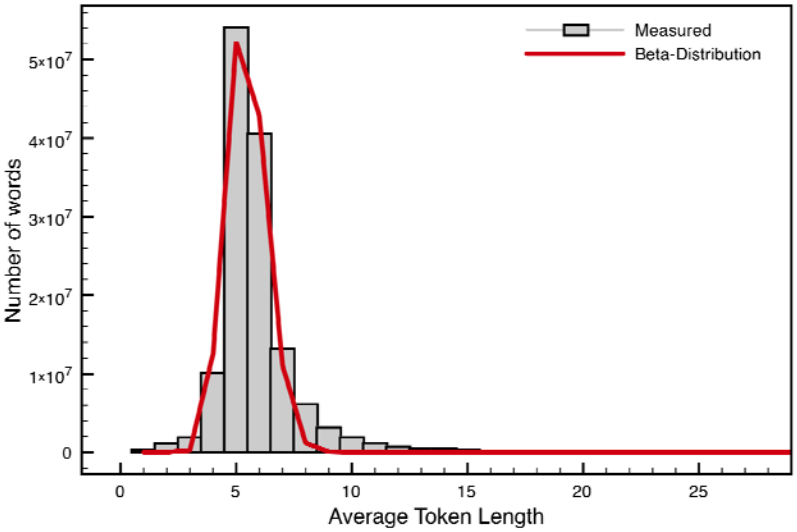



Figure 3: Average Token Length Distribution

*4.4 Term Typicality*

To make a statement on the meaning of the determined two classes, the content of these classes, that is the term vocabulary, needs to be analyzed. If the two classes are different, then the contained token vocabulary should also expose noticeable differences. To prove this, let us first divide the corpus text into two partitions $\pi_1$ and $\pi_2$; $\pi_1$ only contains blocks with densities $\varrho' \leq 8$ and $\pi_2$ with $\varrho' \geq 9$ ($\varrho' = 8$ is the boundary point of the two discrete beta distributions). Second, let us analyze the token distributions of the two partitions. As we want to express the peculiarities of the two classes $C_1$ and $C_2$, which are roughly represented by $\pi_1$ and $\pi_2$, we compare the partition-specific term document frequencies. We expect that terms that are typical for $C_1$ appear much more often in $\pi_1$ than in $\pi_2$, and vice versa. We examine this relationship by computing the corresponding document frequency ratio. The normalized ratio follows a power law distribution of the form $y = c(x/(1 - x))^{-a_1}$ with $a_1 = 0.39$ and $c = 0.01$ ($R^2 = 0.9468$, RMSE = 0.0034, see Figure 4). This type of power law distribution has recently been discussed by Lava-lette (2007) and Popescu (2003) as a generalization of Zipf's law. In our case, we can interpret the ratio $x/(1 - x)$ as the combination of two Zipfian subsets, a top-ranked and a bottom-ranked one, which mutually influence the curve. In fact the document frequencies of the considered terms apparently are Zipfian, too, and for both partitions enough typical terms exist.

To avoid over-interpreting the impact of rarely occurring terms, the analysis is limited to terms with a collection-wide document frequency $w_{1 \cup 2}$ of at least 100. For these terms, we compute the term typicality $\varepsilon(t)$, which I define as the logarithmic ratio of the corresponding document frequencies $w_1$, $w_2$ of the examined term t in the two partitions. The ratio is normalized by the logarithm to base N + 1 with N being the number of documents in the corpus (i.e., the maximum document frequency):

$$\varepsilon(t) = \log_{N+1} \frac{w_2(t)+1}{w_1(t)+1} \tag{10}$$

The resulting values are in the range of [-1;+1]. The absolute score is the degree of typicality, the sign indicates the direction of typicality (-1 means the term clearly belongs to class 1, +1 states that the term clearly belongs to class 2). In our setup, of the 2938 terms with $w_{1 \cup 2} \geq 100$, 589 terms (20%) expose a term typicality $\varepsilon \leq -0.05$

(i.e., $C_1$) and 1255 terms (42.7%) a term typicality of $\varepsilon \geq +0.05$ (i.e., $C_2$). Table 2 shows the top-20 typical terms for C1 and C2 respectively. As one can see, $C_1$ terms are very likely to appear in template blocks, whereas $C_2$ terms are more likely for full-text.

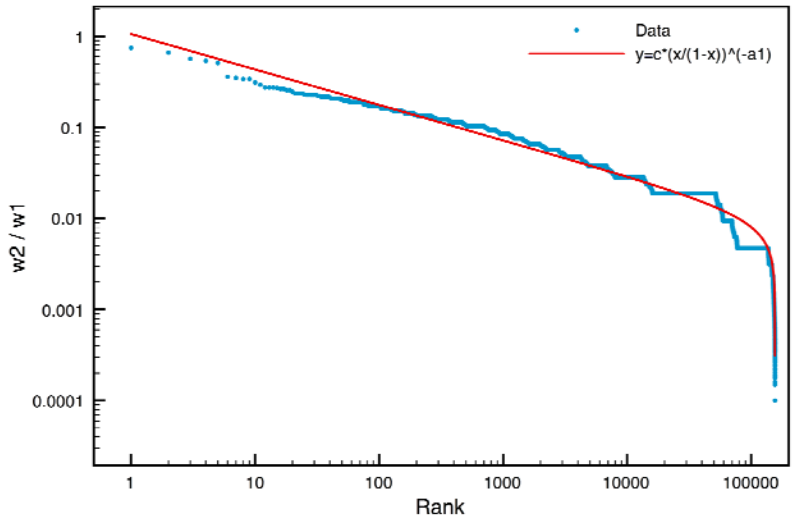



Figure 4: Document Frequency Ratio ($a_1 = 0.39$)

### *4.5 The "Full Stop" Criterion*

I argued that template text usually contained no full stop. This clearly is an observation that needs to be empirically analyzed for the whole corpus. A strong correlation between the feature "segment contains full stop" and the class relationship would support this assumption.

We proceed as follows. Let us partition the corpus' text into segments that contain at least one full stop ($\pi_F$) and those without ($\pi_N$). I simply define "full stop" as a dot character (.) which immediately follows a white-space terminated sequence of at least two letters, except for a few known abbreviations (vs., DC., Inc., Ltd., No., VAT. and Jan. to Dec.). First, let us analyze the histographical distribution of tokens enclosed by segments of particular text densities (as in Section 4.3), for both partitions $\pi$F and $\pi$N separately and compare to the overall distribution. The histograms are depicted in Figure 5. Even though segments with and without full stop exist for all present text densities, the two

strata are clearly visible. Again, we can fit the Beta distribution $y = c \cdot f_{beta}(x,a,b)$ to each of the two partitions with high correlation. For the non-full stop part, we get $R^2 = 0.91$, RMSE = 0.015 with a = 1,394, b = 10,75, c = 0.02749. For the full stop part, we get $R^2 = 0.94$, RMSE = 0.014 with a = 30.15, b = 61.54, c = 0.024. The $R^2$ scores are not as good as for the overall fit (see Figure 2) as we ignored the impact of the inter-class normal distribution in this case. Of note, the modes of the two Beta distributions (1.4 and 11.7) are almost identical to the ones found for the classes $C_1$ (1.99) and $C_2$ (12.15).

Using Weka[3], I computed the expected information gain provided by the "full stop" feature at segment-level. The information gain (Kullback-Leibler-divergence) is defined as the change in information entropy from the prior state (C) to a state that takes some information as given (C|A): IG(C,A) = H(C) -H(C |A). Indeed, text density has a fairly high information gain for predicting the occurrence of a full stop (0.711), which is substantiated by a classification accuracy of 91.4% using a simple linear classifier.

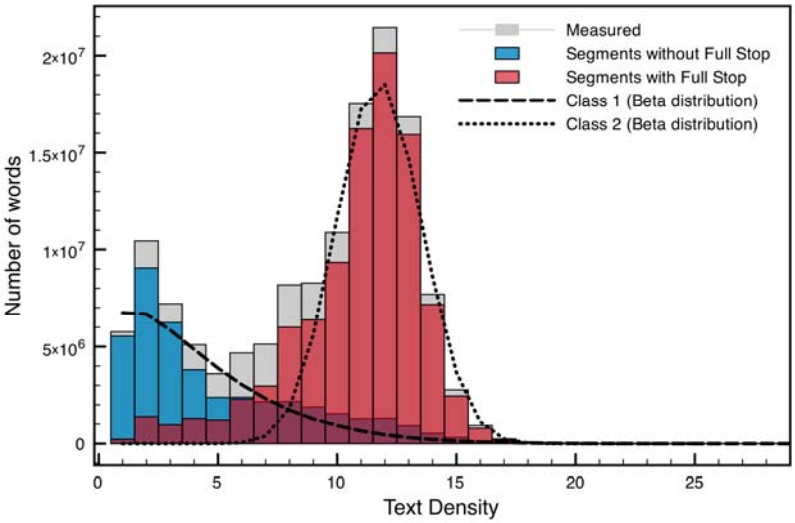



Figure 5: Full Stop as a simple partitioning criterion

Finally, the amount of tokens enclosed in $\pi_N$ (4.18 · 107 or 69%) and in $\pi_F$ (9.50 · 107 or 31%) is in line with the ratio between $C_1$ and $C_2$ determined by the Beta distribution fit described in Section 4.3.

## *5 TEMPLATE REMOVAL*

We now investigate the correlation of the discovered properties of Web text to a baseline strategy for template detection.

### *5.1 Baseline*

As shown in Gibson et al. (2005), the frequency of a segment hash in the collection is a good measure for detecting templates, especially those which are regarded static "boilerplate" text; hashing is not very effective for rarely occurring template segments and for those which contain dynamic, context-sensitive (e.g., time and date) or random text.

For each text segment in the corpus, a hash fingerprint is created as follows: First, the text is converted into lowercase. Month and weekday tokens (January-December, Jan-Dec, Monday-Sunday, Mon-Sun) as well as AM and PM and any URL found in plaintext are removed from the input. Then any characters except Latin letters are replaced by whitespace, which is normalized to a single space between any remaining token. If the original text contained at least one date token, $DATE$ is added as a special indicator token. If the original text contained at least one URL, $URL$ is added; the two indicators are meant to help avoiding hash conflicts of actually different strings. Finally, the SHA1 digest is computed for the remaining text, whose hexadecimal representation is the text's fingerprint.

### *5.2 Evaluation*

The above process results in 1,717,039 distinct fingerprints. Interestingly, the sequence of segment frequencies seems to follow Lavalette's extension to the Zipf law (see Section 4.4) of slope $a_1 = 0.7408$ with $R^2 = 0.9329$, RMSE = 14.2665 (unnormalized). The curve is depicted in Figure 6. This may again be due to the fact that basically two classes of text exist in the corpus, those that are very frequent (boiler-plate templates) and those that are not frequent (non-boilerplate content).

Next, let us investigate the token-level distribution of frequent templates. We will consider segments that have a fingerprint frequency

[3] http://www.cs.waikato.ac.nz/ml/weka/

of at least 10. 38,634 of such segments exist, representing 28% of the tokens in the corpus (3.8 · 107 out of 1.37 · 108) – see Table 4 for the most frequent ones. The corresponding token distribution again can be fitted to a combination of two Beta distribution and the normal distribution (Equation 9) with $R^2 = 0.9966$, RMSE = 0.0026 having $a_1 = 106.6$, $b_1 = 162.6$; $a_2 = 5.348$, $b_2 = 64.35$; $c = 0.0204$, $d = 0.4821$; $e = 81.69$, $f = 2.045$, $\mu = 23.81$, $\sigma^2 = 9.264$; x scores (densities) have been normalized by $x_{norm} = 31$ to [0:1] before fitting). From d we see that the two Beta-distributed parts are almost equally important to the distribution (48.2% vs. 51.8%), this also correlates with the ratio between templates with full stop $\pi_{F,T}$ and without $\pi_{N,T}$ (47% to 53%); see Figure 7.

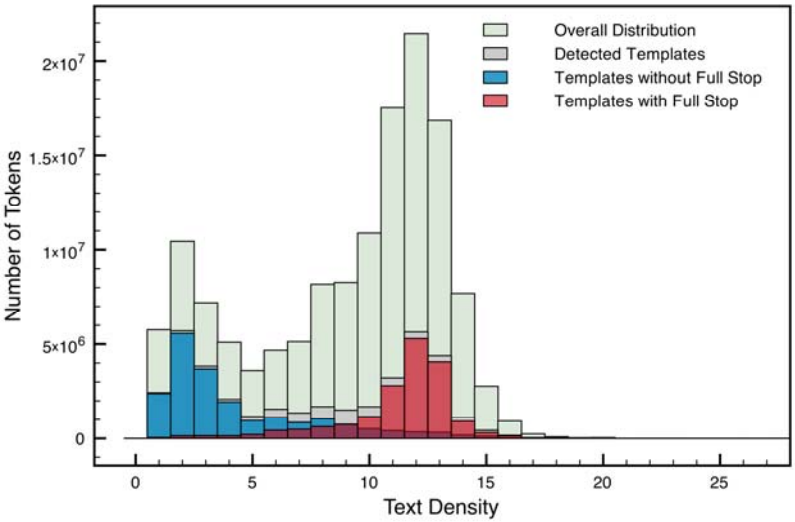



Figure 7: Templates detected by Fingerprinting

As we can see, the relative amount of detected template content in class $C_1$ is much higher than in class $C_2$. $\pi_{N,T}$ represents 63% of the tokens covered by segments with $\varrho'(b) \leq 5$, whereas $\pi_{F,T}$ only represents ca. 17% of the tokens for $\varrho(b) \geq 6$ and 20% for $\varrho(b) \geq 9$. An analysis of the remaining segments with $\varrho'(b) \leq 5$ which are not covered by $\pi_{N,T}$ using the term-typicality measure (Table 2) and a random-sample manual inspection (Table 3) indicates that no significant structural difference exists between the detected boilerplate templates and the remainder of text with a text density of 5 or less. Some segments appear to be part of a headline or closing words of a letter, which may

indicate a sub-optimal segmentation caused by the BlockFusion algorithm (with a segmentation accuracy of 80% this was expectable). I conclude that the part of C1 with $\varrho'(b) \leq 5$ represents template text with a high probability and could simply be removed right after the segmentation without requiring a global (fingerprint frequency-based) strategy. The removed content represents 23% of all tokens in the corpus and 84% of the tokens detected by the baseline strategy.

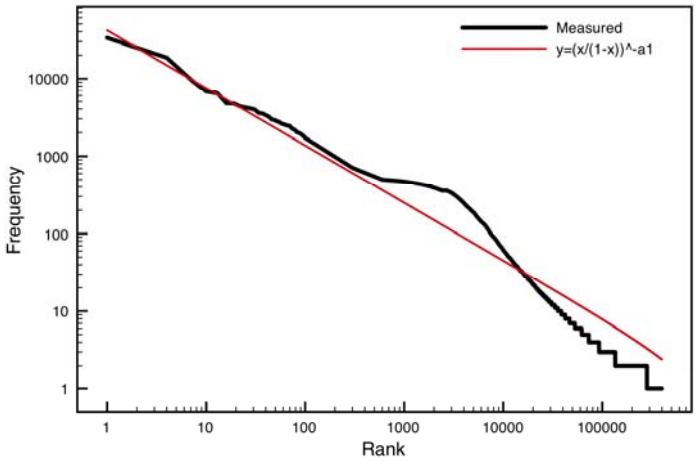



Figure 6: Segment Frequency Distribution ($a_1$=0.74)

## *6 CONCLUSIONS AND FUTURE WORK*

### *6.1 Dichotomy of Web text*

Web text exposes two prominent classes (strata) of content; they can be seen best at segment-level when analyzing the ratio between text density and token count. Each class can be modeled using the Beta distribution with a fuzzy transition between them following the normal distribution. The proportion of the two classes (in tokens) roughly is 1:2. This classification is found when inspecting the text's densities (or possibly sentence lengths), not when comparing lower levels of text (e.g. token lengths).

As we have seen, the textual contents of the two classes significantly differ from each other, in notation (sentences vs. non-sentential text) as well as in terminology. With regard to the linguistic model, we may

interpret the class with a low text density average ($C_1$) as a class that is of navigational nature (i.e., allowing a quick, economic perception of provided or related content), whereas the class with a high text density average ($C_2$) describes content of descriptive nature (i.e., supplying the reader with the subject matter's details at the cost of higher syntactic complexity).

### *6.2 Application to Template Removal*

47% of the tokens which were classified as template content by the baseline strategy are covered by segments with a text density of $\varrho'(b) \leq 5$. The partition represents 23% of the tokens in the corpus and 84% of tokens detected by the baseline. The obvious strategy to obtain a cleaner text collection is to segment each document using the Block-Fusion algorithm and to remove all segments that have a maximum text density of 5. As opposed to the fingerprint baseline strategy, no site-level or global information is necessary for this pruning operation.

### *6.3 Next Steps*

Although this work presents a well-grounded model for template content, it would be interesting to see whether machine-learning techniques could further improve the classification task. By adding more features for this classification, for example the number of links in a segment, an even higher accuracy is expected. Of course, one also has to measure the impact of this template removal strategy to search (in terms of Precision and Recall). Moreover, it is to be investigated how well the presented model matches to machine-generated spam content.

Finally, it must be analyzed to what extent the findings from Web text may be applied to content from other media, for example illustrated magazines, which also contain a good portion of low-density segments (headlines, legends and other captions).

## *ACKNOWLEDGMENTS*

I would like to express my gratitude to Professor Dr. Gabriel Altmann and Dr. Peter Fankhauser for fruitful and enlightening discussions.

## *REFERENCES*


Al-Saleh, A.; A. El-Zaart (2007): Unsupervised Learning Technique for Skin Images Segmentation Using a Mixture of Beta Distributions. In: *Biomed 06, IFMBE Proceedings* 15, 304-307.

Altmann, G. (1992): Das Problem der Datenhomogenität, in: *Glottometrika* 13, 287–298.

Altmann, G., and Violetta Burdinski (1982): Towards a Law of Word Repetitions in Text-Blocks, in: *Glottometrika* 4, 147–167.

Baluja, S. (2006): Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, ACM, 33–42.

Bar-Yossef, Z.; S. Rajagopalan (2002): Template detection via data mining and its applications. In *WWW*, 2002, 580–591.

Cai, Deng, Shipeng Yu, Ji-Rong Wen; Wei-Ying Ma (2003): Extracting content structure for web pages based on visual representation. In: Zhou, X.; Zhang,Y.; Orlowska, M. E. (eds.): APWeb. Volume 2642 of LNCS, Springer, 406–417.

Cai, Deng, Shipeng Yu, Ji-Rong Wen; Wei-Ying Ma (2004): Block-based web search. In: *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York: ACM, 456–463.

Chakrabarti, Deepayan, Ravi Kumar; Kunal, Punera (2007): Page-level template detection via isotonic smoothing. In: *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York: ACM, 61–70.

Chakrabarti, Deepayan, Ravi Kumar; Kunal Punera (2008): A graph-theoretic approach to webpage segmentation. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York: ACM, 377–386.

Chen, Yu, Wei-Ying Ma; Hong-Jiang Zhang (2003): Detecting web page structure for adaptive viewing on small form factor devices. In: *WWW '03: Proceedings of the 12th international conference on World Wide Web*. New York: ACM, 225–233.

Debnath, Sandip, Prasenjit Mitra, Nirmal Pal; C. Lee Giles (2005): Automatic identification of informative sections of web pages. IEEE Transactions on Knowledge and Data Engineering, vol. 17, 9, 1233-1246.

Fernandes, David, Edleno S. de Moura, Berthier Ribeiro-Neto, Altigran S. da Silva; Marcos Andrй Gonзalves (2007): Computing block importance for searching on web sites. In: *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. New York: ACM, 165–174.

Gibson, David, Kunal Punera; Andrew Tomkins (2005): The volume and evolution of web page templates, in: *WWW'05*, 830–839.

Grzybek, Peter (ed.) (2006): *Contributions to the Science of Text and Language.* Heidelberg: Springer.

Kao, Hung-Yu, Jan-Ming Ho; Ming-Syan Chen (2005): Wisdom: Web intrapage informative structure mining based on document object model, in: *Transactions on Knowledge and Data Engineering, IEEE, vol.* 17, 5, 614-627.

Köhler, Reinhard (1990): Elemente der synergetischen Linguistik, in: *Glottometrika* 12, 179–187.

Köhler, Reinhard (2005): Synergetic linguistics. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin u.a.: Walter de Gruyter, 760–774. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Kohlschütter, Christian; Wolfgang Nejdl (2008): A Densitometric Approach to Web Page Segmentation. In: *ACM 17th Conference on Information and Knowledge Management (CIKM 2008)*. Napa Valley, California, USA, 377-386.

Lavalette, Daniel (2007): A general purpose ranking variable with applications to various ranking laws. In: Grzybek, P.; Köhler, R. (eds.): *Exact Methods in the Study of Language and Text*. Berlin: de Gruyter, 371–382.

Naranan, S.; Balasubrahmanyan V. K. (2005): Power laws in statistical linguistics and related systems. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin u.a.: Walter de Gruyter, 716–738. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Popescu, Ioan-Iovitz (2003): On a Zipf's Law Extension to Impact Factors, in: *Glottometrics* 6, 83-93.

Popescu, Ioan-Iovitz; Altmann, G. (2006): Some aspects of word frequencies, in: *Glottometrics* 13, 23–46.

Tuldava, Juhan (2005): In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin u.a.: Walter de Gruyter, 368–387. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Tweedie, Fiona J. 2005. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin u.a.: Walter de Gruyter, 387–397. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Vieira, Karane et al. (2006): A fast and robust method for web page template detection and removal. In: *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management.* New York: ACM, 258–267.

Wimmer, Gejza; Altmann, Gabriel (1999): *Thesaurus of univariate discrete probability distributions.* Stamm Verlag.

Wimmer, Gejza, Altmann, G. (2005): In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin u.a.: Walter de Gruyter, 791–807. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27].

Yi, Lan, Bing Liu; Xiaoli Li (2003): Eliminating noisy information in web pages for data mining. In: *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.* New York: ACM, 296–305.

Zipf, George K. (1949): *Human Behavior and the Principle of Least Effort.* Reading: Addison-Wesley.

*APPENDIX*

Table 1
The top-20 terms for $\pi_1$ and $\pi_2$

| **Rank** | **Term** | $\varepsilon$ | **Term** | $\varepsilon$ |
|---|---|---|---|---|
| 1 | sitemap | -0.33 | spelled | 0.51 |
| 2 | bookmark | -0.29 | thousands | 0.36 |
| 3 | accessibility | -0.29 | temporarily | 0.35 |
| 4 | misc | -0.29 | gave | 0.34 |
| 5 | skip | -0.28 | tried | 0.33 |
| 6 | shipping | -0.28 | aimed | 0.33 |
| 7 | polls | -0.28 | seem | 0.32 |
| 8 | affiliates | -0.27 | eventually | 0.31 |
| 9 | username | -0.27 | unfortunately | 0.31 |
| 10 | thu | -0.27 | obvious | 0.31 |
| 11 | faq | -0.26 | helped | 0.31 |
| 12 | miscellaneous | -0.26 | majority | 0.30 |
| 13 | jun | -0.26 | reached | 0.30 |
| 14 | basket | -0.26 | despite | 0.30 |
| 15 | gmt | -0.26 | incorrectly | 0.30 |
| 16 | wed | -0.26 | hundreds | 0.30 |
| 17 | faqs | -0.25 | themselves | 0.30 |
| 18 | currency | -0.25 | although | 0.30 |
| 19 | homepage | -0.24 | whether | 0.29 |
| 20 | checkout | -0.24 | we'll | 0.29 |

Table 2
The top-20 typical terms in segments with $\varrho' \leq 5$
(frequent vs. infrequent segments)

| Rank | Term | $\varepsilon$ | Term | $\varepsilon$ |
|---|---|---|---|---|
| 1 | memberlist | -0.37 | option | 0.32 |
| 2 | usergroups | -0.34 | van | 0.31 |
| 3 | headcovers | -0.33 | liability | 0.29 |
| 4 | accesskey | -0.33 | rd | 0.29 |
| 5 | changelog | -0.33 | gloucester | 0.28 |
| 6 | thimbles | -0.31 | income | 0.28 |
| 7 | notifications | -0.30 | provider | 0.27 |
| 8 | tuskers | -0.30 | tea | 0.26 |
| 9 | gnomes | -0.30 | settings | 0.25 |
| 10 | qed | -0.30 | cheap | 0.24 |
| 11 | videophone | -0.30 | pension | 0.23 |
| 12 | stocked | -0.30 | creed | 0.22 |
| 13 | brvbar | -0.30 | their | 0.22 |
| 14 | landscaper | -0.29 | these | 0.22 |
| 15 | prater | -0.29 | adverse | 0.21 |
| 16 | upchurch | -0.29 | directory | 0.21 |
| 17 | sge | -0.29 | michael | 0.21 |
| 18 | barebone | -0.29 | sku | 0.21 |
| 19 | dr.who | -0.29 | double | 0.21 |
| 20 | turntables | -0.29 | accident | 0.20 |

Table 3
Examples of infrequent segments with $\varrho' \leq 5$

| Frequency | Segment |
|---|---|
| 1 | What synthetic methods are used? |
| 1 | Home > Gears & Sprockets > Plastic Model Gears |
| 2 | B' Team photo |
| 1 | Brian Stone 11 May 05 Wheatear Bakewell [...] |
| 1 | How do you read Braille? |
| 1 | Electricians in Tyne & Wear |
| 1 | media assistance mapping |
| 3 | Home > IPOD / MP3 > Other MP3 Accessories |
| 1 | Ford Mustang 67 |
| 2 | Summer Barbecues |
| 1 | cheaptickets airline cheapticket |
| 1 | Cheers, Kevan. |

Table 4
The most frequent segments in the corpus

| **Freq.** | **Segment** | **Freq.** | **Segment** | **Freq.** | **Segment** |
|---|---|---|---|---|---|
| 33,349 | Home | 4,231 | What's New? | 2,943 | site search |
| 27,841 | Search | 4,204 | Skip navigation | 2,941 | Quantity |
| 14,897 | Contact Us | 4,190 | Events | 2,933 | Introduction |
| 10,101 | Links | 4,131 | Features | 2,903 | Not Found |
| 9,747 | Back | 4,073 | Publications | 2,892 | Quick Search |
| 8,517 | News | 4,045 | The document has moved here. | 2,874 | Sitemap |
| 7,937 | About Us | 3,949 | Tell A Friend | 2,791 | Services |
| 6,800 | Information | 3,942 | Main Menu | 2,771 | Accessories |
| 6,696 | Site Map | 3,874 | Products | 2,760 | You are not logged in. |
| 6,573 | Login | 3,669 | Price: | 2,717 | Shopping Cart |
| 6,433 | Categories | 3,560 | Terms & Conditions | 2,713 | About |
| 5,400 | Contact | 3,517 | FAQ | 2,634 | Print this page |
| 4,814 | Advanced Search | 3,465 | This object may be found here. | 2,590 | profile |
| 4,775 | Help | 3,417 | Checkout | 2,586 | Jump to: |
| 4,762 | Object Moved | 3,344 | Newsletter | 2,562 | Accessibility |
| 4,760 | Log In | 3,308 | Privacy Policy | 2,551 | Description |
| 4,736 | Back to top | 3,231 | Skip to content | 2,509 | Please try the following: |
| 4,366 | top | 3,190 | home page | 2,500 | Technical Information (for support personnel) |
| 4,360 | Register | 3,132 | Reviews | 2,496 | Quick Find |
| 4,237 | Latest News | 3,041 | Navigation | 2,479 | Contact Details |

# Перлокутивний ефект висловлення-звинувачення

*Антоніна Король (Чернівці, Україна)*

Сучасний етап розвитку лінгвістики характеризується парадигмою новітніх напрямів досліджень, серед яких чільне місце посідає вивчення комунікації. Цілеспрямований процес інформативного обміну між комунікантами дає змогу пізнати аспекти вияву мови та мовлення, зокрема, встановити значущість мови в процесі соціально-виробничих відносин людей, усвідомити роль людини у створенні ситуацій спілкування, здійснення мовленнєвого акту, його інтерпретації з огляду на умови реалізації, мовленнєву стратегію, тактику, ефективність тощо.

Об'єктом дослідження є – реактивні висловлення-відповіді на звинувачення-стимул мовця на матеріалі німецькомовного художнього дискурсу.

Метою статті є встановлення та аналіз основних реактивних стратегій на висловлення-звинувачення адресанта.

Перлокуцію можна розглядати як поняття, яке складається щонайменше з двох компонентів: перлокутивної мети (спричинити певну мовленнєву чи не мовленнєву поведінку адресата за допомогою мовних засобів) та перлокутивного ефекту (наслідок досягнення/недосягнення даної мети). Варто враховувати той факт, що навіть тоді, коли продуцент висловлення-звинувачення не має наміру подіяти на свого співрозмовника, перлокутивний ефект має місце. У висловленні-звинуваченні ми позначили перлокутивну мету як каузацію почуття провини, вибачення/виправдання адресата, щоб запобігти повторенню таких дій у майбутньому, які мовець кваліфікує як неприйнятні в даному соціумі. Якщо об'єкт звинувачення відчуває провину за свої неправомірні вчинки, порушення норм поведінки, моральних цінностей, традицій тощо, вибачається за це або знаходить обґрунтоване виправдання, тоді продуцент висловлення-звинувачення досяг бажаного перлокутивного ефекту. Невідповідність між перлокутивною метою та перлокутивним ефектом визначається в теорії мовленнєвих актів як комунікативна невдача.

Ф. Гундснуршер наголошує на тому, що висловлення-звинувачення містять "агресивні компоненти", які зашкоджують іміджу одного з учасників інтеракції (адресату), так як даним висловленням мовець приписує своєму партнеру по комунікації відповідальність за певні дії/наслідки, які оцінюються ним негативно. Адресат висловлення-звинувачення може відновити своє "обличчя"/імідж лише в тому випадку, якщо він вибачиться за свої вчинки або знайде обґрунтоване виправдання (Hundsnurscher 1997). Таким чином, вибачення (Entschuldigung) та виправдання (Rechtfertigung) вважаються очікуваними реакціями на продуковане висловлення-звинувачення, які допомагають уникнути подальшого розвитку конфліктної ситуації.

Необхідність дослідження перлокутивного ефекту висловлення-звинувачення обґрунтовується нами як розгляд проблеми успішної/неуспішної взаємодії комунікантів у вирішенні конфліктної ситуації. Продуцент висловлення-звинувачення наносить шкоду „обличчю" звинувачуваного, який у свою чергу повинен намагатись відновити позитивну репутацію в суспільстві. Інтерес до проблеми відновлення „обличчя-репутації" (image-restoration) набуває все більшої актуальності на сучасному етапі, коли людина заповзято заявляє про свої етичні права. „Обличчя" (репутація) відіграє велику роль для індивіда в суспільстві. На думку Дж. Даймонд, коли люди спілкуються, вони роблять це не стільки для передачі інформації, скільки для створення та підтримання репутації/іміджу (Diomond 1996). Важливість лінгвістичного дослідження в цій галузі людських взаємин полягає в наступному: якщо когось „критикують, звинувачують, дорікають, він повинен знати, як відповісти, щоб зберегти своє „обличчя", свою репутацію" (Benoit 1995).

При загрозі втрати свого „обличчя" люди вимушені звертатись до так званої „стратегії захисту" або „відновлення обличчя" (image-restoration strategy), тобто вдаватись до певних дій для порятунку свого обличчя (self-defense). Реалізація цих дій відображається у визначених стратегіях. І, не дивлячись на культурні відмінності, репертуар стратегій по „збереженню власного обличчя" багато в чому виявляється схожим (Owen 1983: 161). Деякі з них називаються в роботі П. Браун і С. Левінсон: вибачення, визнання провини або відповідальності за здійснену чи нездійснену дію. Існує інша класифікація, в якій поряд з вибаченням та виправданням виділяються такі стратегії як заперечення провини, ухилення від відповідальності (Evading the Responsibility), заниження шкоди від події, що відбулась (Corrective Action), обіцянка виправитись.

Німецький лінгвіст Г. Грубер пропонує у своїй науковій праці „Streitgespräche: zur Pragmatik einer Diskursform" 6 наступних стратегій реагування адресата на висловлення-звинувачення: 1) уникати будь-якої реакції (Vermeidung einer Reaktion); 2) зволікати/переводити розмову на іншу тему (Hinhalten/Ablenken); 3) захищати власну невинність/непричетність (Verteidigung der Unschuld); 4) за-перечувати відповідальність (Abstreiten der Verantwortlichkeit); 5) виправдовуватись (Rechtfertigung); 6) вибачатись (Entschuldigung) (Gruber 1996: 64).

У цій статті ми пропонуємо власну класифікацію реактивних висловлень-відповідей на звинувачення-стимул мовця, що базується на проведеному аналізі найчастотніших стратегій, до яких вдається адресат висловлення-звинувачення у німецькомовному художньому дискурсі для збереження власного „обличчя". Розглядатимуться реактивні висловлення-відповіді на висловлення-звинувачення адресанта, що відновлюють гармонійні стосунки між комунікантами, та висловлення, що посилюють конфліктну ситуацію між ними.

Реактивні висловлення-відповіді на звинувачення-стимул, що сприяють погіршенню міжособистісних стосунків між комунікантами, провокують подальший розвиток конфлікту. Вивчення природи комунікативних невдач такого типу дозволить забезпечити ефективність порозуміння між людьми та запобігти конфліктно спрямованому спілкуванню.

У процесі дослідження мовного матеріалу проведено кількісний аналіз реактивних реплік секвенції (використовуємо термін „Sprechhandlungssequenz" / „Paarsequenz" – (Diomond 1996; Behagel-Thomsen/Lundquist-Mog/Mog 1993) на висловлення-звинувачення, результати якого графічно зображені на рис. 1.

Встановлено 8 типів реактивних висловлень-відповідей на звинувачення-стимул: найчастіше адресат висловлення-звинувачення застосовує тактику **виправдання** власного вчинку, який мовець розглядає як негативний/ненормативний, – (23,4 %); на другому місці – **зустрічні звинувачення** адресанта (21,7 %), які ми відносимо до категорії висловлень, що посилюють конфліктну ситуацію між комунікантами; в однаковому співвідношенні знаходяться тактичні варіанти **заперечення провини** (13,1 %) та **мовчання** (13,0 %), а також висловлення **ігнорування** *звинувачення* (10,2 %) та **здивування** (10,0 %); **визнання власної провини/вибачення** складають лише 6,5 % від проаналізованої вибірки; найменшу частку серед вищезгаданих типів реактивних висловлень-відповідей

займають висловлення **погрози** (2,1 %), які розглядаються нами як „агресивні" мовленнєві акти, що загрожують розвитку конфліктної ситуації звинувачення до суперечки, сварки тощо, тому застосовуються комунікантами вкрай рідко.

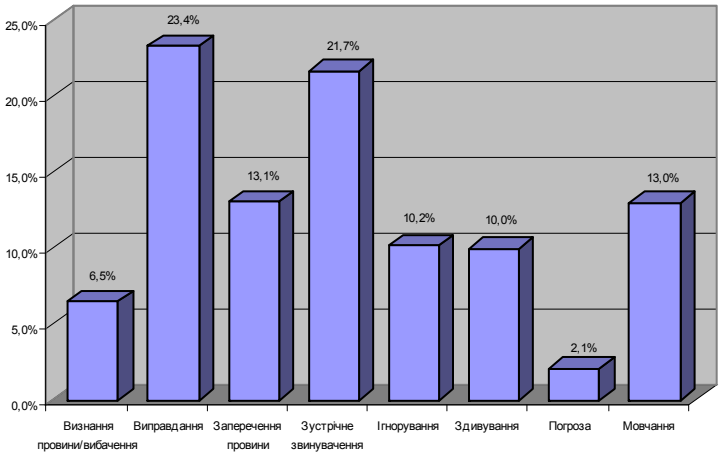



Рис. 1 Типи реактивних висловлень-відповідей на висловлення-звинувачення

Наступним етапом дослідження перлокутивного ефекту висловлення-звинувачення було встановлення залежності різних типів реактивних висловлень від форми реалізації ініціативного висловлення-звинувачення (пряме/"не-пряме" висловлення-звинувачення). Результати аналізу подані в табл. 1.

Як показують дані таблиці 1, тактичний варіант **виправдання** на висловлення-звинувачення мовця відзначається найбільшим показником частоти вживання у проаналізованій вибірці (23,4 %) і продукуються переважно у відповідь на непрямі експліцитні висловлення-звинувачення; незафіксований даний тип реактивного висловлення у випадку імпліцитного висловлення-звинувачення (0 %).

У процесі аналізу секвенційних пар „**звинувачення → зустрічне звинувачення**" встановлено, що показник частоти їх вживання дорівнює 21,7 %. Дані табл. 1 дозволяють зробити висновок, що дана стратегія реагування на висловлення-звинувачення займає

друге місце поряд з домінуючою тактикою виправдання (23,4 %) і реалізується в основному на пряме експліцитне висловлення-звинувачення мовця (12,5 %). Зустрічне звинувачення як реакційна мовленнєва дія не було зафіксоване в діалогічних єдностях з перформативним висловленням-звинуваченням, що пояснюється неможливістю застосування такої стратегії в стандартних інституційно-прив'язаних ситуаціях з обов'язковою регламентацією вербальної та невербальної поведінки мовця, що вказує на особливості німецької ментальності щодо дотримання чіткого "порядку", особливо в офіційно-діловій комунікації.

Таблиця 1
Залежність типу реактивного висловлення
від форми висловлення-звинувачення

| № | Форма звинувачення / Тип реактивного | Перформатив висловлення-звинувачення | Пряме експліцитне звинувачення | Непряме експліцитне звинувачення | Імпліцитне висловлення-звинувачення | Всього (%) |
|---|---|---|---|---|---|---|
| 1 | Визнання провини/ вибачення | 0,3 | 1,3 | 4,7 | 0,2 | 6,5 |
| 2 | Виправдання | 0,3 | 6,5 | 16,6 | 0 | 23,4 |
| 3 | Заперечення провини | 0,2 | 7,1 | 5,3 | 0,5 | 13,1 |
| 4 | Зустрічне звинувачення | 0 | 12,5 | 8,5 | 0,7 | 21,7 |
| 5 | Ігнорування | 0 | 4,9 | 4,2 | 1,1 | 10,2 |
| 6 | Здивування | 0 | 5,8 | 4,0 | 0,2 | 10,0 |
| 7 | Погроза | 0 | 1,3 | 0,8 | 0 | 2,1 |
| 8 | Мовчання | 0,5 | 6,1 | 5,9 | 0,5 | 13,0 |
| Всього (%) | | 1,3 | 45,5 | 50 | 3,2 | 100 |

Аналіз даних таблиці 1 засвідчує, що показник частоти вживання стратегії **заперечення провини** як реакції на висловлення-звинувачення займає третє місце після стратегій виправдання та зустрічного звинувачення серед загальної кількості зафіксованих варіантів реактивних висловлень (13,1 %) і здійснюється переважно у відповідь на пряме експліцитне висловлення-звинувачення (7,1 %).

Кількісний аналіз ситуацій **ігнорування** висловлення-звинувачення показує, що показник частоти їхнього вживання в німецькомовному художньому дискурсі дорівнює 10,2 %, серед яких 4,9 % належать до реплік-реакцій на пряме експліцитне

висловлення-звинувачення, 4,2 % – на непряме експліцитне висловлення-звинувачення 1,1 % – на імпліцитне висловлення-звинувачення, не зафіксовано у випадку перформативного висловлення-звинувачення.

Аналіз секвенційних пар **"звинувачення → здивування"** дозволяє констатувати факт нечастотного застосування даної комунікативної стратегії адресатом висловлення-звинувачення – 10,0%, яка продукується частіше на звинувачення-стимул у прямій експліцитній формі (5,8 %) і не зустрічається взагалі у випадку перформативного висловлення-звинувачення адресата.

З табл. 1 видно, що **визнання власної провини/вибачення** не є поширеним явищем (6,5 %) як реакція на різні форми висловлення-звинувачення адресата в художньому дискурсі. Переважна більшість реактивних висловлень секвенції „звинувачення → визнання провини/вибачення" здійснюється мовцем на непряме експліцитне висловлення-звинувачення – 4,7 %.

У проаналізованій вибірці реактивне висловлення **погрози** характеризується найнижчим показником частоти вживання (2,1 %), як правило, у контекстах прямого висловлення-звинувачення і незафіксований взагалі серед прикладів перформативного та імпліцитного висловлення-звинувачення.

У цій статті ми пропонуємо аналіз найпоширенішої реакції адресата у комунікативній ситуації „звинувачення" – пошук виправдання, яке дозволило б йому зменшити відповідальність за здійснений негативний вчинок і тим самим компенсувати шкоду „обличчю", заподіяну продуцентом звинувачення.

Виправдання та заниження провини у неприйнятних/нелегітимних діях адресата може реалізовуватись через приписування мовцем відповідальності іншій особі чи причинам, заперечення наміру, заперечення контролю над ситуацією. Розглянемо на прикладах застосування даних стратегій. Пояснюючи причину, що викликала необхідність здійснення небажаного для адресанта висловлення-звинувачення вчинку, адресат може вказати на неможливість запобігти йому:

(1) *"Du hast mich betrügen müssen? Mich hintergehen müssen?"*
*"Es ist doch nicht anders möglich, Papa…Sie zwingen mich…"* (Werfel, S. 45)

Вказуючи на причину, звинувачуваний може пояснити здійснення вчинку ненавмисного характеру, зважаючи на певні обставини:

(2) *"Du hast gelauscht."*
*"Ihr habt laut genug gesprochen"* (Heinrich, S. 120)
(3) *"Und dem Wärter hast du einen Stoss versetzt"*
*"Unabsichtlich, nur so beim Aufspringen"* (Loest, S. 9)

У наведених прикладах адресат вдається до тактики заниження власної провини, не заперечуючи повністю відповідальності за здійснені вчинки, що оцінюються продуцентом висловлення-звинувачення як негативні.

У випадку реалізації висловлення виправдання адресат може зменшити ступінь шкоди для власного „обличчя", використовуючи стратегію перенесення відповідальності на іншу особу (shifting the blame) (Benoit 1995). Розглянемо наступну комунікативну ситуацію, учасниками якої є батько (продуцент непрямого експліцитного висловлення-звинувачення) та син (реципієнт висловлення-звинувачення).

Пан Майєр "попросив" своїх синів – Петера та Фріца – приносити йому щодня ранкову газету перед тим, як вони підуть до школи. Батько обґрунтував своє прохання: якщо він буде робити це сам, тоді не встигне на автобус, щоб доїхати на роботу. Петер та Фріц, будучи відповідно добре вихованими синами, сприймають всі прохання батька як накази. Тому зобов'язались по черзі приносити батькові щодня ранкову газету з найближчого кіоску. Проте, одного ранку хлопці побігли до школи, не придбавши для батька бажаної газети. Пан Маєр, як справжній педант, не міг відмовитись від вже звичного для нього порядку: сніданок – ранкова газета – робота, тому був змушений самостійно купувати газету, через що запізнився на роботу. Коли Фріц прийшов зі школи, батько звернувся до нього з наступним висловленням-звинуваченням:

(4) *"Warum hast du heute morgen die Zeitung nicht geholt?"*
*„Heute war doch Peter an der Reihe"* (Pehnt, S. 69)

Мовець виражає за допомогою реалізованого висловлення-звинувачення наступне:

1. Мовець стверджує, що адресат порушив дану ним обіцянку (приносити щодня газету) і тим самим піддав сумніву його домінуючий статус – роль батька, накази якого обов'язково повинні виконуватись.
2. Мовець оцінює дану ситуацію негативно: діти не повинні ігнорувати прохання батьків.

3. Мовець передбачає, що адресат висловлення-звинувачення, вдавшись до альтернативних дій – купити газету або повідомити, що не зможе цього зробити, міг запобігти порушенню їхньої домовленості.
4. Мовець вимагає від адресата вираження власного ставлення до здійсненого вчинку, вибачення чи виправдання своїх дій.

Висловлюючи – *Warum hast du heute morgen die Zeitung nicht geholt? / Чому ти не приніс сьогодні вранці газету?*, мовець реалізує одразу два *звинувачення*: 1) якби ти приніс газету, я б вчасно потрапив на роботу; 2) якби ти приніс газету, ти б виконав моє прохання. З чого слідує: якби адресат поважав батька, він не порушив би обіцянки і купив йому газету.

Адресат висловлення-звинувачення вказує у своєму виправданні на провину Петера (свого брата), який повинен був принести газету, бо прийшла його черга – *„Heute war doch Peter an der Reihe“*. Для батька стає зрозумілим, що об'єктом його звинувачення є відсутній на даний момент Петер. Таким чином, за допомогою прийому перенесення відповідальності на іншу особу Франц компенсував шкоду „обличчю", заподіяну продуцентом висловлення-звинувачення.

Максимальний ступінь компенсації заподіяної шкоди реалізується в стратегії обіцянки звинувачуваного виправитись, не вдаватись до схожих дій в майбутньому:

(5) *“Warum haben Sie die Seite herausgerissen?”*
*“Ich hab mich verschrieben”*
*“Deshalb hatten Sie nicht die ganze Seite herausreißen müssen!”*
*“Ich schreib sie neu”* (Ferolli, S. 61)

У прикладі (5) спостерігаємо ситуацію „звинувачення" адресата з нижчим комунікативним статусом у припущенні професійної помилки в офіційно-діловому спілкуванні, на яку адресат реагує виправданням (*Ich hab mich verschrieben / я помилилась*), підсилюючи його своєю обіцянкою виправитись (*Ich schreib sie neu / я напишу її заново*).

Отже, особливістю висловлення-звинувачення у межах німецькомовного художнього дискурсу є широта його перлокутивного потенціалу: реактивні висловлення-відповіді адресата на звинувачення-стимул мовця варіюють в діапазоні від визнання провини та вибачення до заперечення провини, зустрічного звинувачення, ігнорування, здивування, мовчання й погрози.

*ЛІТЕРАТУРА*

Behagel-Thomsen H.; Lundquist-Mog A.; Mog P. (1993): *Typisch deutsch? Arbeitsbuch zu Aspekten deutscher Mentalität.* – Radolfzell: Verlag für Gespräch-Forschung.

Benoit William L. (1995): *Accounts, Excuses and Apologies: A Theory of Image Restoration Strategies.* – Albany (N.Y.): State University of New York Press.

Diomond J. (1996): *Status and Power in Verbal Interaction: A Study of Discourse in a Closenit Social Network.* – Amsterdam: J. Benjamins Publishing Co.

Gruber H. (1996): *Streitgespräche: zur Pragmatik einer Diskursform.* – Opladen: Westdt. Verl.

Hundsnurscher F. (1997): *Streitspezifische Sprechakte* // Intention – Bedeutung – Kommunikation. Kognitive und handlungstheoretische Grundlagen der Sprechakttheorie. – Opladen: Westdeutscher Verlag.

Owen M. (1983): *Appologies and Remedial Interchanges: A Study of Language Use in Social Interaction.* – Berlin etc.: Mouton.

# Слова-консоциации и побочный смысл: забытые идеи Г. Шпербера и К.О. Эрдмана

*Виктор Левицкий (Черновцы, Украина)*

## *1. ВСТУПИТЕЛЬНЫЕ ЗАМЕЧАНИЯ*

В последние годы наблюдаются случаи, когда авторы диссертаций, следуя неписанным (а, возможно, вообще не существующим) рекомендациям ВАК, стремятся использовать в своих исследованиях работы, опубликованные преимущественно в последнее время. С одной стороны, такое стремление вполне соответствует требованиям современной науки. С другой стороны, однако, на практике это приводит к тому, что диссертант получает информацию не из первых, а из вторых или даже третьих рук, нередко не замечая даже, что автор, на которого делаются ссылки, излагает не свои собственные мысли, а мысли своих предшественников. В результате авторство той или иной идеи, методы дефиниции приписываются вовсе не тому, кто их высказал, а тому, кто последним их упомянул. Но если в подобных случаях речь идет, скорее, о нормах этики, а не об интересах науки, то незнание истории вопроса или такое «знание», глубина которого охватывает в лучшем случае последние 10-15 лет, пренебрежение к «устаревшим» работам «традиционного» европейского и советского языкознания нередко приводят, по выражению М.И. Стеблина-Каменского, к раскладыванию «нового терминологического пасьянса», а не к новым открытиям.

Сказанное в определенной степени относится к идеям Г. Шпербера, высказанным в начале XX столетия и получившим дальнейшее развитие в 60-е годы в советском языкознании. Кроме того, нам приходилось уже писать о том, что некоторые современные трактовки концепта очень близки по своему содержанию к интерпретации лексического значения в трудах немецких семасиологов начала XX столетия (см. Левицкий/Лех 2007).

## *2. МОДЕЛИ ПРЕДСТАВЛЕНИЯ МЕНТАЛЬНЫХ СУЩНОСТЕЙ*

Д.С. Лихачев сопоставляет концепт с основным значением слова; при этом концепт, по мнению Д.С. Лихачева, отличается от основного значения тем, что говорящий по-своему – в зависимости от уровня образования, опыта, профессии, социального статуса и т.п. – интерпретирует и реализует это основное значение слова (см. Лихачев 1997: 213, 214).

С другой стороны, концепт сопоставляется с полным набором значений многозначного слова, включает в себя и денотативное, и коннотативное значение и отражает представления носителей языка, связанные с индивидуальным опытом и определенной культурой (Кравченко 2007: 128, Рябцева 1991: 75).

Бросается в глаза явное сходство интерпретации понятия «концепт» в приведенных выше высказываниях с трактованием значения слова, высказанным К.О. Эрдманом около ста лет назад.

Известно, что К.О. Эрдман различал в значении слова три основных компонента: понятийное ядро (Begriff), побочный смысл (Nebensinn) и эмоциональное значение (Gefühlswert). Под Nebensinn К.О. Эрдман понимает все побочные и сопровождающие представления, которые слово обычно и непроизвольно вызывает в нашем сознании. Под чувственным значением или настроенческим содержанием (Stimmungsgehalt) понимаются все «реактивные чувства и настроения, которые производит слово» (Erdmann 1910: 107). В побочный смысл включаются все индивидуальные представления, которые связаны у человека с тем или иным понятием (ср. с выше приведенными трактовками концепта). По сравнению с понятийным содержанием побочный смысл и чувственное значение слова, полагает К.О. Эрдман, являются весьма неустойчивыми, неопределенными и подвержены временным, пространственным и индивидуальным колебаниям (ср. с приведенным высказыванием Д.С. Лихачева). Побочный смысл и эмоциональное значение слова, продолжает К.О. Эрдман, содержат субъективные «добавки» к представлению или понятию. Переход от объективного к субъективному осуществляется постепенно, и провести четкие границы между понятийным содержанием, с одной стороны, и побочным смыслом и чувственным значением, с другой, не всегда удается (Erdmann 1910: 125). Эти мысли К.О. Эрдмана несомненно перекликаются с некоторыми определениями концепта в современной лингвистике:

1) концепты – это нежестко структурированные единицы без четких границ и внутренней структуры (Попова/Стернин 2000: 23); концепты – в отличие от понятий – не только мыслятся, но и переживаются, они являются предметом эмоций, симпатий и антипатий …, концепты – это сгусток культуры в сознании человека, то, в виде чего культура входит в ментальный мир человека …, тот «пучок» представлений, понятий, знаний, ассоциаций, которые сопровождают слово (Степанов 2001: 40, 85).

Аналогичным образом Л.О. Чернейко и В.А. Долинский полагают, что концепт включает в свое содержание не только наивное понятие, но и множество коннотативных элементов имени, что проявляются в его сочетаемости (Чернейко/Долинский 1996: 22), а А.Н. Приходько, подытоживая проведенное сравнение понятия и концепта, пишет: «Понятие лишь тогда становится концептом, когда оно валоризуется, то есть обрастает определенными ценностными ассоциациями» (см. Приходько 2008: 54).

Таким образом, из приведенных сопоставлений следует, что семасиологи прошлого хорошо понимали различия между понятийным содержанием слова и совокупностью представлений, ассоциаций, индивидуальных эмоциональных оценок, связанных в сознании носителя языка со значением слова. Причем, они не только хорошо понимали такого рода различия, но и умели четко моделировать структуру лексического значения. Сравнение того, что в современной лингвистике понимается под концептом, и того, что в лингвистике XX столетия понимали под семантической структурой, состоящей из денотативного, сигнификативного и коннотативного значений, позволяет сделать следующий вывод.

Ментальные единицы и сущности могут быть представлены в виде двух основных моделей или их разновидностей (что на самом деле происходит в сознании, мы не знаем). Одна модель вычленяет некоторую «логическую» сущность («понятие»), которое дополняется в коллективном или индивидуальном сознании побочными ассоциациями и экспрессивно-эмоциональными компонентами. Разновидностью этой модели является также трехкомпонентная структура, в которой, однако, два последних компонента (ассоциация + экспрессивно-эмоциональный компонент) объединяются в одно целое, именуемое «коннотативное значение», а первый компонент («понятийное ядро») как бы расщепляется на два – объем понятия («предметное содержание», денотат) и содержание понятия («логическое содержание», сигнификат).

Преимущество первой модели заключается, с нашей точки зрения, в том, что четкое размежевание (по крайней мере, в Теории) денотативного, сигнификативного и коннотативного значений (а внутри последнего – экспрессивных, эмоциональных, оценочных компонентов) позволило разработать адекватные методы исследования каждого из этих значений. Денотативное и сигнификативное значения изучались с помощью компонентного, квантативно-компонентного, дистрибутивного, дистрибутивно-статистического анализа; коннотативное (а также частично и денотативно-сигнификативное) значение исследовалось с помощью психолингвистических методов. Что касается концептов, то они вообще могут быть невербализованными (т.е. мыслиться и вычленяться исследователем совершенно произвольно) или вербализируются с помощью слов, словосочетаний, синонимов, антонимов, семантических полей, фразеологических единиц, языковых шаблонов и т.п. На практике это означает, что, если даже исследователь использует какие-либо из перечисленных приемов, то обобщение такого «анализа» не может осуществляться иначе, как интуитивно.

Наконец, коротко заметим, что определение концепта как «знания» («кванта знания» и т.п.) не делает концепт существенно отличным от значения слова. Еще в 1990-1991 годах, когда в отечественной лингвистике еще мало было известно о том, что такое концепт, мы определяли значение слова как «приобретенные посредством опыта в процессе деятельности знания о совокупности речевых и неречевых ситуаций, в которых может быть употреблено слово» (см. Левицкий 1991: 112; Левицький 1990: 124-125).

### *3. ВОЗМОЖНОСТЬ ИЗУЧЕНИЯ КОННОТАТИВНОГО ЗНАЧЕНИЯ С ПОМОЩЬЮ ЛИНГВИСТИЧЕСКИХ МЕТОДОВ*

Если денотативное и сигнификативное значения слова, как показано выше, исследовались с помощью разнообразных лингвистических методов (структурных, контекстологических, статистических), то коннотативное значение (за исключением эмоционально-экспрессивного компонента) изучалось в основном с помощью методики семантического дифференциала, предложенного Ч. Осгудом. Между тем, в лингвистике уже предпринимались попытки исследовать тот компонент, который К.О. Эрдман назвал «побочным смыслом» (т.е. совокупностью дополнительных

представлений и ассоциаций), с помощью контекстологического анализа.

Еще в 1923 году Г. Шпербер ввел в научный оборот понятие «консоциация» (слова, которые часто совместно встречаются в тексте) – см. Sperber 1923: 10-15. Идеи Г. Шпербера нашли разнообразное применение в советском языкознании 60-х годов. Так, А.Я. Шайкевич выдвинул гипотезу о том, что слова, связанные по смыслу, должны часто встречаться в тексте недалеко друг от друга и, наоборот, слова, часто встречающиеся вместе в осмысленном тексте, связанны друг с другом по смыслу (Шайкевич 1963: 15). Эту гипотезу А.Я. Шайкевич использовал для выделения семантических полей путем статистического анализа стихотворного текста. Следует, однако, иметь в виду, что стихотворный текст отличается от прозаического. Если при анализе стихотворного текста за «единицу измерения» можно взять сравнительно небольшой отрезок – одну строку (такую методику использовал А.Я. Шайкевич), то в прозаическом тексте выбор оптимальной длины текста не только затруднителен, но и может оказаться субъективным, в то время, как основной смысл статистической процедуры, предложенной А.Я. Шайкевичем, заключается как раз в том, чтобы выделить семантические поля объективными методами.

Понятно, что чем большую длину текста выберет исследователь в качестве единицы анализа, тем большим может оказаться расстояние между исследуемыми словами и тем больше будет отстоять семантически одно слово от другого. Следует ожидать также, что семантическое сходство исследуемых слов будет зависеть не только от расстояния между ними, но и от грамматического статуса этих слов. Хорошо известно, например, что синтаксические функции и свойства прилагательного существенного отличаются от таковых других частей речи, например, глагола. Центральное положение глагола, обусловливающее его синтаксическую связь почти со всеми членами предложения, позволяет широко и эффективно применять для определения его значения и синтаксических функций методы структурной лингвистики – дистрибутивный и трансформационный анализ. Семантика же имени прилагательного, которое выступает в предложения главным образом в роли определения (реже – предикатива), существенно зависит от семантики определяемого имени существительного. Отсюда можно предположить, что, если в тексте несколько имен прилагательных встречаются в одной атрибутивной синтагме, выступая

определениями одного и того же имени существительного, то между этими прилагательными существует какая-то семантическая связь.

Зная частоту совместна встречаемости различных имен прилагательных в тексте, можно с помощью тех или иных статистических методов установить степень связи между ними. Наиболее простим способом нахождения таких связей может быть использование различимое вариантов формул, в которых учитываются соотношения «общих» и «частных:» долей (см.: Тулдава 1987: 156-157).

Для изучения семантических связей между прилагательными, встречающимися в тексте перед одним и тем же существительным, мы подвергли статистическому анализу произведения немецких и австрийских прозаиков общим объемом около 1,8 млн. словоупотреблений. Если какое-либо прилагательное встречалось совместно с другим более одного раза в одном и том же произведении, считалось, что для данного произведения эта связь не случайна и является своего рода семантическим микрополем данного текста; если какое-либо прилагательное встречалось совместно с другим более одного раза хотя бы в двух не связанных сюжетной линией произведениях одного автора, считалось, что эта связь не случайна и свидетельствует о наличии определенного микрополя, характерного для языка данного автора. Наконец, если прилагательное встречалось совместно с другим в произведениях различных авторов, считалось, что данная связь является не случайной и характерной для языка данной эпохи.

Слово, относительно которого проводился анализ, считалось основным и условно обозначалось термином «доминанта»; слова, которые встречались совместно с доминантой, условно назывались «сопроводителями». Н.А. Шехтман предлагает обозначать такие слова термином «аллоним»: «аллонимы – это контекстуально обусловленные слова-уточнители, которые раскрывают значение предшествующего слова, обнаруживая с ним некоторую семантическую общность» (Шехтман 1965: 8). Хотя, конечно, последующее прилагательное может уточнять семантику предыдущего, это имеет место далеко не во всех случаях; ср. *большой высокий дом* и *большой белый дом.* В первой синтагме *высокий* уточняет семантику слова *большой.* Во второй синтагме два прилагательных характеризуют предмет, выраженный существительным *дом*, с разных сторон. Поэтому мы отдаем предпочтение термину «сопроводитель»,

а не «уточнитель». Кроме того, сопроводителями мы называем такие прилагательные, которые не обязательно встречаются в постпозиции по отношению к слову-доминанте. В зависимости от цели исследования одно и то же слово может рассматриваться либо в функции доминанты, либо в функции сопроводителя. Так, если слово *klar* принималось за доминанту, то его сопроводителями (в порядке убывания частоты совместного появления с ним) оказывались *rein, deutlich, sonnig, ruhig*; если за доминанту принималось *rein*, то его сопроводителями были *klar, hell*; при доминанте *hell* сопроводителями оказывались *heiter*, *klar*. Зная частоту совместного появления каждого анализируемого слова с тем или иным сопроводителем, нетрудно вычислить степень (меру) его связи с другими словами, пользуясь коэффициентами Я. Чекановского и П. Жаккара, которые в советском языкознании (вслед за С.Г. Бережаном) находились чаще всего с помощью формулы

$$M = \frac{n-1}{N}, \quad (1),$$

где (в данном случае) $M$ означает искомую меру, $n$ – число употреблений слова-сопроводителя совместно со словом-доминантой, а $N$ – общее число появления слова-доминанты со словами-сопроводителями.

Выявив на первом этапе семантические связи между отдельными словами, на втором этапе можно подвергнуть качественному анализу полученные группы сопроводителей: слова-сопроводители, более тесно связанные друг с другом по смыслу, объединяются в семантические подгруппы (подклассы). Анализируя, например, аморфный в семантическом отношении ряд, в котором слова расположены в порядке убывания частоты совместного употребления со словом groß, а именно: groß – klein, weiß, dunkel, hager, dick, blau, deutsch, gelb, hoch, rot, rund, schön, schwer, riesig, braun, blond, blaß, französisch, grün, hell, schlank, mager, stolz, kräftig и т.д., можно выделить следующие семантические подклассы, расположив их в порядке убывания средней частоты подкласса: groß – klein; groß – weiß, dunkel, blau, gelb, rot, braun, blond, blaß, grün, hell; groß – hager, mager; groß – schlank, noch; groß – dick; groß – deutsch, französisch; groß – schwer.

Весьма интересными оказываются результаты, если в семантические подклассы объединить слова-доминанты. В этом случае

может быть дана количественная оценка взаимосвязи между целыми лексико-семантическими подклассами. Так, выясняется, что в современном немецком языке прилагательные размера чаще всего сочетаются с прилагательными цвета, затем – друг с другом; прилагательные цвета чаще всего встречаются с прилагательными, обозначающими размер и форму предметов, затем – совместно друг с другом, с оценочными прилагательными (*schön, hübsch, fein, gut, häßlich*), прочими прилагательными. Прилагательные температурной группы (*warm, heiß* и др.) довольно часто встречаются с прилагательными *feucht*, *weich*; частым сопроводителем *kalt* является, с одной стороны, *heiß*, а с другой – *frisch*.

В таблице 1 представлены величины коэффициента M (см. выше формулу 1), т.е. мера связи прилагательных размера с другими семантическими подклассами.

Таблица 1
Мера связи прилагательных размера
с другими семантическими подклассами прилагательных

| **Подклассы слов-доминант** | **Подклассы слов-сопроводителей** | **Мера связи** |
|---|---|---|
| размер | цвет | 0,22 |
| размер | размер | 0,10 |
| размер | оценка | 0,079 |
| размер | форма | 0,031 |
| цвет | размер | 0,31 |

Для иллюстрации связей между прилагательным-доминантой и прилагательными-сопроводителями воспользуемся словами-доминантами, составляющими семантический подкласс «размер» (см. табл. 2).

Средние коэффициенты связей позволили сделать вывод, что имена прилагательные, обозначающие размер, чаще всего встречаются в тексте перед одним и тем же существительным с прилагательными-антонимами: *groß-klein*, *lang-kurz, hoch-niedrig* (средний коэффициент связи $M = 0{,}34$); затем – с прилагательными, семантика которых включает семы, обозначающие разные, расположенные под углом друг к другу, направления: *lang-schmal, lang-dünn, hoch-breit,* и т.п. ($M = 0{,}17$). Интересно, что применение иной статистической и дистрибутивной процедуры – корреляционного анализа, основанного на частоте встречаемости тех же имен

прилагательных с различными именами существительными в тексте, привело к аналогичным результатам:

$r_1$ *(hoch-niedrig)* = 0,42
$r_2$ *(weit-eng)* = 0,38
$r_3$ *(breit-schmal)* = 0,35
$r_4$ *(lang-kurz)* = 0,23
$r_5$ *(breit-hoch)* = 0,15
$r_6$ *(kurz-weit)* = 0,06
$r_7$ *(breit-niedrig)* = 0,03
$r_8$ *(schmal-eng)* = 0,01

Таблица 2
Мера связи между прилагательными размера

| Слово-доминанта | Слово-сопроводитель | Мера связи |
|---|---|---|
| groß | klein | 0,48 |
| groß | dick | 0,12 |
| groß | breit, weit, hoch, riesig | 0,04 |
| klein | groß | 0,48 |
| klein | dick | 0,17 |
| klein | schmal | 0,18 |
| lang | dünn | 0,2 |
| | schmal | 0,27 |
| | kurz | 0,13 |
| | breit, niedrig | 0,11 |
| kurz | dick | 0,4 |
| | lang | 0,13 |
| hoch | niedrig | 0,4 |
| | schmal | 0,09 |
| | riesig | 0,08 |
| | breit | 0,11 |
| niedrig | hoch | 0,4 |
| | breit | 0,2 |
| | lang | 0,11 |
| breit | niedrig | 0,2 |
| | dick, noch, lang | 0,11 |
| schmal | lang | 0,27 |
| | klein | 0,18 |
| | hoch | 0,11 |

Первые четыре коэффициента характеризуют силу связей между прилагательными-антонимами; коэффициенты $r_5$-$r_7$ – силу связей между прилагательными, обозначающими различные направления; последний коэффициент характеризует связь между псевдосинонимами (*eng* и *schmal н*е заменяют друг друга в тексте, так как *schmal* означает «узкий в одном направлении», а *eng* – «узкий в двух и трех направлениях», т.е. «тесный»).

Совместное употребление нескольких прилагательных в пределах одной синтагмы может быть обусловлено, как показано выше, различными причинами. Так, в одних случаях перечисляются различные признаки одного предмета или объединенные денотативными связями (а потому и сходные) признаки разных предметов, например: *ein großer runder Tisch или gelbe, rote und weiße Blumen*. В других случаях говорящий, употребляя несколько имен прилагательных подряд, с помощью последующих слов стремится глубже и точнее выразить тот смысл, который обозначен семантикой предыдущего слова. Вследствие этого, как указывает целый ряд авторов (см. Behagel 1928, Анциферова 1962, Шехтман 1965), одно и то же слово может служить уточнителем другого.

Выявление подобных случаев совместного употребления прилагательных может явиться важным вспомогательным средством обнаружения ассоциативных связей, существующих между понятиями, выражаемыми этими именами прилагательными, а отсюда – вспомогательным средством в этимологических исследованиях, когда требуется подтвердить предполагаемую линию семантического развития или восстановить недостающие звенья в семантической цепи, а также для установления семантических законов. Небезынтересно отметить в этой связи, что ассоциация понятий «малый» и «тонкий, изящный, нежный», о наличии которой свидетельствует частое совместное употребление имен прилагательных *klein* и *zierlich, fein* и *klein, klein* и *zart*, *klein* и *schwach,* существовала на более ранних ступенях развития языка и подтверждает линию семантического развития немецкого *klein* («тонкий, изящный», «маленький»), а также подобную линию семантического развития англ. *small.* Ассоциация понятий «легкий» и «нетвердый», а также «легкий» и «малый», о чем свидетельствует частое совместное употребление имен прилагательных *leicht* и *mühelos, leicht* и *klein,* лежала в основе изменения значений немецкого *gering;* частое совместное употребление имен прилагательных *schwer* и *massig, groß* и *massig, groß* и *stark* отражает существующую связь между

понятиями «тяжелый», «массивный», «крепкий», «сильный», «большой». Интересно, что подобные связи в синхронии могут «развертываться» в диахронии, образуя, как показано выше, цепочку семантических изменений того или иного слова. Так, англ. *big* прошло путь семантического развития, в основе которого лежала ассоциация понятий «тяжелый, массивный, большой». Во многих языках существуют ассоциации, наличие которых подтверждает совместное употребление таких слов, как *fein* – «тонкий» *schlau* – «хитрый» (ср. семантику др.-рус. хитрый)*, dunkel* – «темный», «мрачный» и *grausam* – «жестокий», *weiß* – «белый» и *sauber* – «чистый», *kalt* – «холодный» и *ruhig* – «спокойный», *nieder* – «низкий» и *gemein* – «подлый» и т.д. Среди других пар прилагательных, совместно употребленных в одной синтагме и отмеченных довольно высокой частотой встречаемости, следует назвать*: jung-blond, klien-schwach, breit-flach, schlank-hübsch, jung-gesund, jung-stark, klein-jung, sicher-ruhig.*

Однако применение формулы (1) для определения степени связи между лексемами не позволяет судить о том, обладают ли найденные коэффициенты статистической значимостью. Более точным и более чувствительным инструментом анализа в этом случае может служить критерий $\chi^2$ и коэффициент сопряженности Чупрова.

Применение этих статистических приемов для нахождения «стандартных» синтагматических связей имен прилагательных со значениями «сильный» и «слабый» (см.: Капатрук 1981) в немецком языке позволило обнаружить следующее. Наибольшим числом стабильных синтагматических связей, т.е. связей, имеющих статистическую значимость, обладают имена прилагательные *stark* и *kräftig.* Стабильными партнерами имени прилагательного *stark* (в рамках одной атрибутивной или предикативной синтагмы) являются имена прилагательные *groß* ($\chi^2$ = 14,5), *lang* ($\chi^2$ = 8,6), *klug* ($\chi^2$ *= 11,4), glücklich* ($\chi^2$ = 10); у имени прилагательного *kräftig* стабильными партнерами являются: *jung* ($\chi^2$ = 10,2), *schön* ($\chi^2$ *= 7,9), düster* ($\chi^2$ *= 5,2), untersetzt* ($\chi^2$ = 8,6); у *mächtig – reich* ($\chi^2$ = 181); у *heftig – kurz* ($\chi^2$ = 15,7); у *stämmig – klein* ($\chi^2$ = 42,4); у *schwächlich – klein* ($\chi^2$ = 28,1). Прилагательное *jung* встречается совместно как с прилагательными, имеющими значение «сильный», так и с прилагательными, имеющими значение «слабый»; тем не менее с первой группой прилагательное *jung* связано теснее. Объединение данных о частоте сочетаемости отдельных слов, обозначающих

«сильный» и «слабый», в два ряда, соответствующих двум семантическим подклассам группе прилагательных со значением «сильный» и группе прилагательные значением «слабый», дает возможность исследовать лексическую сочетаемость лексико-семантической микросистемы. Статистический анализ показывает, что подгруппа «сильный» теснее всего связана с прилагательными *groß* ($\chi^2 = 14,8$), *breit* ($\chi^2 = 6,1$), *schön* ($\chi^2 = 6,1$); подгруппа «слабый» – с прилагательными *alt* $\chi^2 = 7,7$), *klein* ($\chi^2 = 4,5$), *krank* ($\chi^2 = 4,1$).

Таким образом, имена прилагательные со значением «сильный» теснее всего связаны в тексте с прилагательными со значением общей положительной оценки («умный», «красивый», «большой», «счастливый» и т.п.), а прилагательные со значением «слабый» – с прилагательными условной отрицательной оценки – «старый», «маленький», «больной». Если учесть, что аналогичные связи (и аналогичными методами) установлены на материале английского языка (см.: Быстрова 1978: 47), следует сделать вывод, что установленная закономерность носит межъязыковой характер и, следовательно, обусловлена внеязыковыми факторами.

Изучение совместной встречаемости двух и более прилагательных перед одним и тем же существительным может оказаться плодотворным не только для этимологического или историко-семасиологического анализа слова, но и для исследования возникновения парных словосочетаний. Так, например, в средневерхненемецких поэтических текстах прилагательные grõz и michel, klein и lützel часто употребляются в одной и той же или в соседних строках, нередко с одним и тем же существительным, образуя своего рода парное сочетание; например:

was kurzwîle sô michel unde grôz (Nib., 1819, 1)
des heldes sterke … diu was michel unde grõz (Nib., 1492, 2).

Аналогичные примеры, иллюстрирующие совместное употребление muchel и gret в среднеанглийском, приводит К.О. Кох (см. Koch 1906: 67).

Указывая на особенность употребления прилагательных michel и grôz поэтами 13 ст., М.М. Гухман в первой части своей монографии «От языка немецкой народности к немецкому национальному языку» рассматривает сочетания типа michel unde grôz как сочетания территориальных дублетов. Что касается более позднего периода, то здесь М.М. Гухман придерживается иной точки зрения,

считая, что употребление прилагательных michel и groß в библиях 15 ст. не было связано с локальными расхождениями, а объяснялось факторами хронологического характера (устаревшее michel вытеснялось из литературного языка словом groß). – см. Гухман, ч. 2, с. 12. Мы рассматриваем сочетание michel и grôz в приведенных выше свн. текстах как сочетание темпоральных дублетов: одно слово как бы передает свои функции другому, это второе, как справедливо указывают О. Бехагель (Behagel 1928: 368) и Г. Пауль (Paul 1960: 107), уточняет значение первого, повторяет его смысл. Неслучайно это сочетание встречается именно в тот период, когда функции выражения понятия «большой» переходят от слова michel к слову grôz. О функции слов-уточнителей в парных словосочетаниях писали в свое время Г. Анциферова (см. Анциферова 1962) и И.Г. Ольшанский (см. Ольшанский 1963).

## *4. СЛОВА-СОПРОВОДИТЕЛИ В КОНЦЕПТОЛОГИИ*

В последние годы появились работы, в которых совместная встречаемость слов в определенном отрезке текста (в пределах одного предложения) используется для экспликации концептов. Следует, безусловно, приветствовать стремление современных концептологов оточнить и усовершенствовать методику концептуального анализа путем использования более точных и более объективных процедурных приемов. Так, в работе А. Андрусь (Андрусь 2008) выполненной под руководством М.М. Полюжина и посвященной концептуальному анализу английских прилагательных big и large, предлагается использовать контекстуальный анализ для выявления фреймов, ролей, модуса и других процедурных компонентов концептуального анализа.

Так, например, анализируя предложение A **large man** in every way, he was *tall* and *broad-shouldered*; bellicose when fondest, hearty when not, автор делает вывод, что прил. *large* «в случае характеристики внешности мужчины объективирует представление о мужчине высокого роста с широкими плечами, что на вербальном уровне подтверждается с помощью лексем *tall* и *broad-shouldered*». Нам представляется, что, во-первых, при экспликации концептов с помощью слов-сопроводителей следует четко различать случаи употребления таких слов в одной синтагме с одним и тем же существительным и случаи употребления тех или иных прилагательных

(или иных слов) перед другими существительными в рамках одного предложения. Например, рассматривая предложение She saw me, and the big eyes stared with mild curiosity, but without fear, исследователь выделяет концептуальный признак «широко открытые глаза», что, с его точки зрения, подтверждается глаголом to stare, обозначающим действие, при осуществлении которого человек широко раскрывает глаза. «Кроме того, – продолжает А. Андрусь, – с помощью выражения with mild curiosity, but without fear осуществляется выявление представлений о том, что человек ситуативно открывает глаза от позитивных эмоций, т.е. от заинтересованности, а не от страха. Опираясь на смыслы фрагментов, можно сделать вывод, что фрейм прилагательного «большие глаза» можно моделировать таким образом: **роли**: глаза женщины; **действия** (человека): широко раскрывать, вытаращить, пристально всматриваться; **статические признаки**: широко раскрытые от позитивных эмоций, выпуклые» (Андрусь 2008: 257).

Понятно, что степень семантической соотнесенности слова-сопроводителя со словом-доминантой в первом из указанных выше случаев (в одной синтагме) и во втором (при употреблении прилагательного с иным существительным в рамках одного предложения) будет неодинаковой. Если исследователь полагает, что в обоих случаях уточняется одна и та же ментальная единица («концепт») то, как минимум, можно предложить выделять при этом различные зоны ассоциаций, состоящих из слов-сопроводителей, – «ближнюю», «дальнюю» и т.п. Во-вторых, должен быть составлен полный, исчерпывающий, список слов-сопроводителей к исследуемому слову-доминанте с точными данными о частоте встречаемости каждой лексической единицы этого списка. Только в случае довольно регулярной повторяемости того или иного слова в этом списке (например, не менее 5 раз) исследователь имеет право делать вывод о неслучайности наблюдаемого явления и приписывать соответствующие свойства («признаки») эксплицируемому концепту. Лишь при строгом сочетании контекстологического, компонентного, статистического и других видов анализа можно надеяться на то, что концептуальный анализ сможет приблизиться к тому уровню процедур, которые были разработаны в лингвистике в «докогнитивный» период.

Однако, пока в современной концептологии будет оставаться нерешенным главный вопрос: что следует понимать под концептом – приблизительно то, что в традиционной семасиологии принято

считать отдельным значением слова или всю совокупность значений многозначного слова (ср. приведенные выше определения концепта) или что-либо еще (например, еще более крупную «невербализованную» ментальную единицу) – даже использование указанных выше процедур не позволит концептологии достичь хотя бы той точности, какой сумела достичь семасиология, ибо при понимании концепта как некоторой «невербализованной ментальной сущности» исследователь может объявить концептом все, что ему заблагорассудится, – в зависимости от того, если следовать диалектическому учению Ф. Энгельса, в какую пятку укусила его ночью проголодавшаяся блоха.

Не следует путать, как нам представляется, крупные ментальные сущности, репрезентирующие основные этнокультурные ценности того или иного народа, с теми семантическими единицами, которые являются предметом лингвистического анализа в многочисленных диссертациях последнего времени. Путаница возникает потому, что термином «концепт» обозначаются, по крайней мере, три различные сущности: 1) упомянутые выше ментальные репрезентанты этнокультуры; 2) семантические единицы, равные отдельному значению слова; 3) семантические единицы, охватывающие все значения многозначного слова. Нельзя сказать, что в семасиологии не различают две последние единицы (2 и 3): третья обозначается иногда термином «семантема», вторая – «семема», а составные части семемы (семантические компоненты) или «признаки» термином «сема». Развиваемый в семасиологии, особенно в советской семасиологии, под влиянием идей структурализма системный подход к изучению лексики привел к тому, что предметом лингвистического анализа оказывалось лишь одно – основное (главное) значение слова, т.к. слова объединялись в лексические микросистемы (семантические поля, ЛСГ и т.п.) по своему основному значению, которое и подвергалось анализу в рамках всей микросистемы. Второстепенные значения слова становились предметом анализа в рамках иной микросистемы, где они входили в периферийные зоны соответствующего поля. Поэтому нам представляется, что наиболее перспективным для концептологии было бы рассмотрение семантики слова как всей совокупности его прямых и переносных значений, включая побочные ассоциации и эмоциональные значения. В этом случае «концептологический» подход, хотя и проигрывает «семасиологическому» в том, что касается системного рассмотрения явления, но выигрывает – по сравнению

с семасиологией – в том, что касается изучения всей совокупности смыслов, выражаемых словом в данном языке. Понятно, что при таком подходе под концептом следует понимать то, что выше обозначено пунктом 3. Возможен и другой вариант лингвистического анализа: семасиология, не отказываясь от системного подхода к лексике, дополняет рассмотрение – по крайней мере, центрального, ядерного, основного – состава лексической микросистемы изучением второстепенных значений соответствующих слов, их побочных смыслов и эмоциональных значений.

Именно для выполнения подобной задачи могут оказаться наиболее полезными и эффективными предложенные выше процедуры контекстологического анализа с опорой на слова-сопроводители.

Только четкое соотнесение концепта с определенными языковыми формами позволит концептологии превратиться из искусства, уровень которого подчас не отличается от примитивной художественной самодеятельности, в подлинную науку.

*ЛИТЕРАТУРА*

Андрусь А. (2008): Деякі аспекти концептуального аналізу мовних одиниць // *Буковинський журнал*, № 2, 254-262.

Анциферова Г. (1962): Beiträge zur Analyse der semantischen Struktur der Wortpaare // *Сборник научных работ студентов-членов НСО 1-го МГПИИЯ*, № 1 (8). – Москва.

Быстрова Л.В. (1978): Вивчення синтагматичних зв'язків слів за допомогою статистичних методів // *Мовознавство,* № 4.

Гухман М.М. (1955): *От языка немецкой народности к немецкому национальному языку*. – Москва.

Капатрук Н.Д. (1981): *Прилагательные со значениями «сильный» и «слабый» в современном немецком языке*: Автореф. дис. … канд. филол. наук. – Минск.

Кравченко А.В. (2001): *Знак, значение, знание*. – Иркутск.

Левицкий В.В. (1965): Некоторые аспекты применения статистики в семасиологии // *Научная конференция аспирантов, посвященная проблемам Романо-германской филологии. Тезисы докладов*. – Москва, 49-51.

Левицкий В.В. (1989): Статистическое изучение лексической семантики. – Киев: УМК ВО, 55-60.

Левицький В.В. (1990): Так що ж таке значення слова? // *Семантика мови і тексту: Матеріали наукової конференції*. – Івано-Франківськ, частина 1, 124-125.

Левицкий В.В. (1991): Значение слова: образ или отношение? // Лексика и лексикография. – Москва, 108-114.

Левицкий В.В.; Лех О.С. (2008): Концептуальное поле размера в немецком языке и методы его исследования // *Studia germanica et romanica*, том 5, № 2 (14), 40-57.

Лихачев Д.С. (1997): Концептосфера русского языка // *Русская словесность: От теории словесности к структуре текста. Антология*. – Москва: Academia, 280-287.

Ольшанский И.Г. (1963): Парные сочетания слов в современном немецком языке // *Уч. записки 1-го МГПИИЯ*, том 28, ч. 2. – Москва.

Попова З.Д.; Стернин И.А. (2000): *Понятие „концепт" в лингвистических исследованиях*. – Воронеж: ВГУ.

Приходько А.М. (2008): *Концепти і концепто-системи в когнітивно-дискурсивній парадигмі лінгвістики*. – Запоріжжя: Прем'єр.

Рябцева Н.К. (1991): „Вопрос": прототипическое значение концепта // *Логический анализ языка: Культурные концепты*. – Москва: Наука, 73-78.

Степанов Ю.С. (2001): *Константы. Словарь русской культуры*. – Москва: Академ. Проект.

Тулдава Ю. (1987): *Проблемы и методы квантитативно-системного исследования лексики*. – Таллинн: Валгус.

Чернейко Л.О.; Долинский В.А. (1996): Имя СУДЬБА как объект концептуального и ассоциативного анализа // *Вестник Московского университета*. – Сер. 9. – Филология. – № 6, 20-41.

Шайкевич А.Я. (1963): Распределение слов в тексте и выделение семантических полей // *Иностранные языки в высшей школе*. – Вып. 2. – Москва.

Шехтман Н.А. (1965): *Сочетаемость прилагательных и система их значений в современном английском языке*. – Автореферат канд. дисс. – Ленинград.

Bahagel O. (1928): *Deutsche Syntax*. Bd. III, Heidelberg.

Erdmann K.O. (1910): *Die Bedeutung des Wortes*. – Leipzig.

Koch C.O. (1906): Contributions to an historical study of the adjectives of size in english. – *Göteborgs Högskolas arsskrift*. Bd. XII. – Göteburg.

Paul H. (1960): *Deutsches Wörterbuch*. – Halle (Saale).

Sperber H. (1923): *Einführung in die Bedeutungslehre*. – Bonn, Leipzig.

# Arc length development and the highest word frequency

*Ján Mačutek (Graz, Austria)*

## 1. INTRODUCTION AND MOTIVATION

Popescu, Mačutek and Altmann (2008) introduced several indices characterising word rank-frequency distributions in texts. All those indices are based on arc length.

Arc length $L$ (of a rank-frequency distribution) is the sum of Euclidean distances between neighbouring frequencies, i.e.,

$$L = \sum_{r=1}^{V-1} \left[ \left( f_r - f_{r+1} \right)^2 + 1 \right]^{\frac{1}{2}},$$

$r$ being the rank, $f_r$ the frequency at the rank $r$, $V$ the text vocabulary.

The value of $L$ depends on several factors (text length, type of language, etc.). We investigate the influence of the highest word frequency $f_1$ on the development of $L$. Both characteristics are measured in cumulative steps of 500 words (i.e., we measure $f_1$ and $L$ after the first 500 words in a text, after the first 1000 words, etc.). The step of 500 words is just an ad hoc decision and its change should not have any impact on obtained results. Values obtained from four texts in four different languages (Latin, German, English, Hawaiian) are presented in the following tables.

The first frequencies and the respective arc lengths from Tables 1 and 2 can be seen in Figure 1.

## 2. TEST FOR PARALLELISM

The aim of the paper is to test whether lengths of two arcs develop differently or not. For the purpose of testing, arc length $L$ will be considered a linear function of $f_1$. Although some kind of a power function is probably more realistic, here we do not have the ambition to explain

the model or to interpret its parameters. If we limit ourselves to testing differences between arc developments, linear regression seems to be satisfactory – it yields excellent fits (the values of $R^2$ are 0.9918 for English, 0.9892 for German, 0.9903 for Hawaiian and 0.9950 for Latin) and there are several tests for parallelism or identity of regression lines (analogous tests for nonlinear models would be far more complicated to derive, and even a nonlinear model itself, given that we have data from four texts only, would be more intuitively than theoretically based).

Table 1
$L$ and $f_1$ in Latin and English

| **Latin** (Ovidius, Ars Amatoria) | | | **English** (Pearl Buck, Nobel lecture) | | |
|---|---|---|---|---|---|
| $N$ | $f_1$ | $L$ | $N$ | $f_1$ | $L$ |
| 500 | 10 | 381.52 | 500 | 32 | 229.04 |
| 1000 | 17 | 710.63 | 1000 | 53 | 407.28 |
| 1500 | 30 | 1029.58 | 1500 | 90 | 567.47 |
| 2000 | 40 | 1319.00 | 2000 | 143 | 734.44 |
| 2500 | 54 | 1580.69 | 2500 | 172 | 874.66 |
| 3000 | 60 | 1819.92 | 3000 | 205 | 1011.78 |
| 3500 | 73 | 2079.69 | 3500 | 236 | 1125.15 |
| 4000 | 85 | 2321.33 | 4000 | 256 | 1233.71 |
| 4500 | 95 | 2546.45 | 4500 | 284 | 1356.32 |
| 5000 | 102 | 2763.09 | 5000 | 321 | 1501.34 |
| 5084 | 102 | 2798.46 | 5500 | 364 | 1633.96 |
| | | | 6000 | 394 | 1736.46 |
| | | | 6500 | 440 | 1859.72 |
| | | | 7000 | 478 | 1973.66 |
| | | | 7500 | 518 | 2092.21 |
| | | | 8000 | 550 | 2189.50 |
| | | | 8500 | 586 | 2302.71 |
| | | | 9071 | 616 | 2398.82 |

In general, normality of the residuals of the linear models is dubious ($p$-values for the Shapiro-Wilk normality test are 0.8294 for English, 0.0101 for German, 0.0139 for Hawaiian and 0.7966 for Latin). Therefore we use the nonparametric Sen-Adichie test for parallelism of regression lines (cf. Hollander and Wolfe 1999, pp. 429-435). The chapter discusses also other nonparametric approaches to the problem

(Hollander 1970, Potthoff 1974). We remind that the Sen-Adichie test, as presented by Hollander and Wolfe (1999), allows to test parallelism of several regression lines, hence here we use its special case (for two lines) only.

Table 2
$L$ and $f_1$ in German and Hawaiian

| **German** (Goethe, Reinecke Fuchs 10,11,12) | | | **Hawaiian** (Moolelo Kawelo IV) | | |
|---|---|---|---|---|---|
| $N$ | $f_1$ | $L$ | $N$ | $f_1$ | $L$ |
| 500 | 20 | 309.50 | 500 | 37 | 178.21 |
| 1000 | 50 | 544.49 | 1000 | 69 | 276.85 |
| 1500 | 76 | 765.61 | 1500 | 98 | 388.36 |
| 2000 | 106 | 968.75 | 2000 | 131 | 498.33 |
| 2500 | 123 | 1163.91 | 2500 | 167 | 585.06 |
| 3000 | 136 | 1310.11 | 3000 | 201 | 666.65 |
| 3500 | 160 | 1445.78 | 3500 | 232 | 733.71 |
| 4000 | 187 | 1599.99 | 4000 | 270 | 783.40 |
| 4500 | 216 | 1744.91 | 4500 | 327 | 870.54 |
| 5000 | 247 | 1876.93 | 5000 | 355 | 930.83 |
| 5500 | 274 | 2014.61 | 5500 | 397 | 996.15 |
| 6000 | 299 | 2158.50 | 6000 | 424 | 1054.29 |
| 6500 | 318 | 2286.80 | 6500 | 466 | 1131.42 |
| 7000 | 336 | 2400.59 | 7000 | 497 | 1194.11 |
| 7500 | 350 | 2504.75 | 7500 | 539 | 1265.85 |
| 8000 | 369 | 2600.08 | 8000 | 568 | 1310.99 |
| 8500 | 397 | 2729.66 | 8500 | 598 | 1355.43 |
| 9000 | 422 | 2865.42 | 9000 | 638 | 1411.16 |
| 9500 | 452 | 2996.64 | 9500 | 680 | 1475.22 |
| 10000 | 466 | 3093.50 | 10000 | 724 | 1534.59 |
| 10500 | 496 | 3209.25 | 10500 | 766 | 1603.63 |
| 11000 | 518 | 3317.00 | 11000 | 804 | 1680.38 |
| 11500 | 541 | 3410.66 | 11500 | 845 | 1758.61 |
| 12000 | 577 | 3533.73 | 12000 | 884 | 1818.95 |
| 12333 | 596 | 3609.93 | 12542 | 920 | 1869.07 |

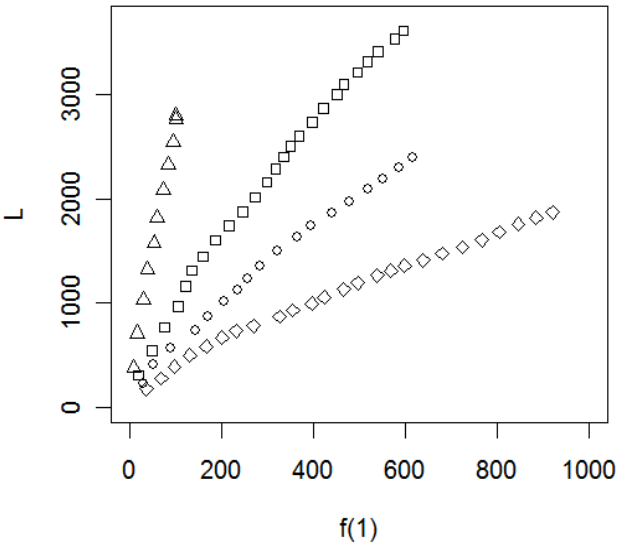



Figure 1
Arc length development in terms of $f_1$ in Latin (triangles), German (squares), English (circles) and Hawaiian (diamonds)

Consider two regression lines

$$Y_1 = \alpha_1 + \beta_1 x_1,$$
$$Y_2 = \alpha_2 + \beta_2 x_2.$$

The lines are parallel if they have a common (but not specified) slope, i.e., the null hypothesis is

$$H_0 : \beta_1 = \beta_2.$$

It will be tested against the alternative hypothesis

$$H_1 : \beta_1 \neq \beta_2 .$$

The test statistic will be constructed in several steps. As an example, we compare arc length developments of the Latin and the English text (cf. Table 1 and Figure 1).

a) Denote $x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}$ and $x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}$ values of $f_1$ in the Latin and in the English text, $Y_{1,1}, Y_{1,2}, \ldots, Y_{1,n_1}$ and $Y_{2,1}, Y_{2,2}, \ldots, Y_{2,n_2}$ values of $L$ in the Latin and in the English text, both characteristics measured in cumulative steps of 500 words.

For the Latin text we have

$$x_{1,1} = 10, x_{1,2} = 17, \ldots, x_{1,11} = 102$$

(cf. Table 1, Latin, column $f_1$) and

$$Y_{1,1} = 381.52, Y_{1,2} = 710.63, \ldots, Y_{1,11} = 2798.46$$

(cf. Table 1, Latin, column $L$); for the English text

$$x_{2,1} = 32, x_{2,2} = 53, \ldots, x_{2,18} = 616$$

(cf. Table 1, English, column $f_1$) and

$$Y_{2,1} = 229.04, Y_{2,2} = 407.28, \ldots, Y_{2,18} = 2398.82$$

(cf. Table 1, English, column $L$).

Evaluate the means

$$\bar{x}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1,j}, \quad \bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2,j} .$$

For the considered texts we have

$$\bar{x}_1 = \frac{1}{11}(10 + 17 + \ldots + 102) = 60.73$$

for the Latin text and

$$\bar{x}_2 = \frac{1}{18}(32 + 53 + \ldots + 616) = 318.78$$

for the English one.

b) Find the common slope $\bar{\beta}$ under the null hypothesis, given by

$$\bar{\beta} = \frac{\sum_{j=1}^{n_1}(x_{1,j} - \bar{x}_1)Y_{1,j} + \sum_{j=1}^{n_2}(x_{2,j} - \bar{x}_2)Y_{2,j}}{\sum_{j=1}^{n_1}(x_{1,j} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2,j} - \bar{x}_2)^2}.$$

For our data we obtain $\bar{\beta} = 4.007$.

c) For all values compute

$$Y^*_{i,j} = Y_{i,j} - \bar{\beta}x_{i,j}, \quad i = 1,2;\ j = 1,2,\ldots,n_i$$

(note that $Y^*_{i,j}$ are not residuals in the usual sense of the word, intercepts are not used). Denote $r^*_{i,j}$ the ranks of $Y^*_{i,j}$ separately within each of the two samples (i.e., we obtain two sets of ranks).

E.g., for the English text we have the following $Y^*_{2,j}$, $j = 1,2,\ldots,18$ : 100.816, 194.909, 206.840, 161.439, 185.456, 190.345, 179.498, 207.918, 218.332, 215.093, 175.412, 157.702, 96.640, 58.314, 16.584, -14.350, -45.392, -69.492 with the respective ranks 7, 14, 15, 9, 12, 13, 11, 16, 18, 17, 10, 8, 6, 5, 4, 3, 2, 1.

d) Compute

$$T_1 = \frac{1}{n_1 + 1} \sum_{j=1}^{n_1} \left( x_{1,j} - \bar{x}_1 \right) r^*_{1,j},$$

$$T_2 = \frac{1}{n_2 + 1} \sum_{j=1}^{n_2} \left( x_{2,j} - \bar{x}_2 \right) r^*_{2,j}$$

and

$$C_1^2 = \sum_{j=1}^{n_1} \left( x_{1,j} - \bar{x}_1 \right)^2,$$

$$C_2^2 = \sum_{j=1}^{n_2} \left( x_{2,j} - \bar{x}_2 \right)^2.$$

We obtain $T_1 = 91.99$, $T_2 = -609.81$, $C_1^2 = 11226.18$ and $C_2^2 = 583256.1$.

e) Finally, we construct the Sen-Adichie statistic

$$V = 12 \left( \frac{T_1^2}{C_1^2} + \frac{T_2^2}{C_2^2} \right).$$

$V$ has the $\chi^2$-distribution with 1 degree of freedom. For our data we obtain $V = 16.696$, with the $p$-value 0.00004. Hence the null hypothesis is rejected and the slopes are significantly different.

The approach was applied here to language typology, but it can perhaps also show differences between genres, authors, historical periods, etc.

*ACKNOWLEDGEMENT*

Supported by FWF, Austria (Lise-Meitner-Programm).

*REFERENCES*

Hollander, M. (1970): A distribution-free test for parallelism. *Journal of the American Statistical Association 65, 387-394.*

Hollander, M.; Wolfe, D.A. (1999): *Nonparametric Statistical Methods.* New York: Wiley.

Popescu, I.-I.; Mačutek, J.; Altmann, G. (2008): Word frequency and arc length. *Glottometrics 17, 18-42.*

Potthoff, R.F. (1974): A non-parametric test of whether two simple regression lines are parallel. *The Annals of Statistics 2, 295-310.*

# Narrative psychological study of self and object representation with young deviant people

*Bernadette Péley, János László[1] (Pécs, Hungary)*

## *1. DEVELOPMENTAL PSYCHOPATHOLOGY*

A number of developmental theories accept the assumption of interpersonal origin of self, however they stress different aspects of this assumption (Balint, 1939, Klein, 1975, Mahler et al., 1975, Bion, 1984, Winnicott, 1990, Bowlby, 1982). It is supposed that infants internalize many aspects of early interactive experiences with caregivers. Interaction involves integration of self- and interpersonal regulation (Beebe et al., 1997). Representations of these early experiences have strong influence on the organization of later interpersonal relationships.

From the point of view of later development, the „affective nucleus" of the early representations is centrally important (Emde, 1999). It is formed by recurrent emotional experience when interacting with significant others at very early age (Emde, 1999, Bucci, 1997). Affects not only influence perception and experiences, but they also shape the representations of "Being-with-Another-in-a–Certain-Way" (Stern, 1995). Importance of meaning is stressed because these theories search for the sense of continuity not on the behavioral level, but on the level of meaning (Sroufe, 1990).

These theories also suggest that self-object representations of the parents influence significantly the self-object representations of their children, hence their children's self-development (Byng-Hall & Stevenson-Hinde, 1991, Stern, 1995, Fonagy et al. 1995). Mental structures, which develop through early interactions, have long term impact on emotional and behavioral responses in social context. Neglected or maltreated children face difficulties in emotional and behavioral regulation. Their social relations are conflict laden. These difficulties are interpreted as

[1] Research was supported by Bólyai Scholarship 2002-2004 and OTKA Research Grant No. T-018306 to the first author.

consequences of mental structures having been formed in early interaction. Maybe these structures have been adaptive in early development, later they become maladaptive.

The process of forming generalized early mental structures is based on recurrent interactions between the child and the parents. These *prototypical* mental structures involve behavioral and emotional elements which are connected by temporal, physical, and causal relations (Stern, 1995, Beebe, Lachmann and Jaffe, 1997). Repetition of affective experiences gives rise to an invariant *affective nucleus*, which has been termed as *prerepresentational self* by Emde (1981). These early "episodic representations" can be conceived as event structures with characters, goals, and states. In this sense, they have the attributes of the later full fledged narratives.

With the emergence of speech, most of our life events become consciously accountable. However, perception of events and reporting on them is influenced by non-conscious experiences and emotions, which are unattended. These non-conscious experiences contribute to the interpretation of the situation, to constructing meanings in the situation. Assignment of meaning in a situation is always based on previous experiences, eventually on early self-representations (Stern, 1985; 2002).

One of the forerunners of narrative psychology, McAdams (1985) claims that a person becomes a biographer of his/her own self when the person reaches the level of formal-operational thinking at the begin-ning of his/her puberty. In this sense puberty is a crucial phase of development. However, in the light of the object-relations theory, it is very likely that our stories have their roots in a much earlier develop-mental phase and, more importantly, originate in early interpersonal relationships.

Narrative analysis of life stories supports this assumption. For instance Shields, Ryan & Cicchetti (2001) confirmed the hypothesis that maladaptive representations are related to continuity in relationship disturbances across the family and peer domains. Maltreated children's narrative representations were more negative/constricted and less positive/coherent than those of non-maltreated children. Furthermore, children's representations mediated maltreatment's influences on peer rejection. In the study below we want to gain further evidence that perception of people in life episodes and behavior towards them is highly influenced by early representations. In other words, maladaptive behavior through narrative analysis of life stories can be traced back to disturbed early self development.

### *1.1 Narrative psychological analysis of characters and their functions*

As in every story, different characters do different actions in the interviewee's episodes. Analyzing a specific class of stories, the Russian magic tales, Vladimir Propp revealed that stories consist of combinations of limited number of actions, and that combinations are led by rules. Story-actions generalize characters' actions that are diverging in their manifest form. For example action-units of causing damage can be realized in 20 different ways, among others the enemy kidnaps someone, the enemy mutilates his/her victim, the enemy makes someone disappear, the enemy cast a spell upon someone, etc. (Propp, 1999, 38-41.). As Propp studies action-units according to their function in building plot of a tale, and naturally the actions are executed by characters of the tale, he recognizes them as function of characters. He reveals altogether 33 functions, for instance *A member of the family leaves home, The hero receives some command of prohibition, The enemy tries to explore the field*, etc. (33-65 pp.). Propp notices that a lot of functions generate certain logical circles, and functions or roles correspond to characters. He found only seven such roles in magic tales (enemy, bestowed, associate, etc.). Let me illustrate the above by four short examples after Propp:

1. The czar gives an eagle to the warrior. The eagle makes the warrior fly away over the mountains.
2. Grandpa buys a horse for Suchenko. The horse gallops to the dragon's cave riding Suchenko.
3. The wizard conjures a boat for Ivan. The boat takes Ivan to another realm.
4. The czarina gives a young serf a ring as a present. Warriors who serve the ring, get the young serf to the place of the combat.

It is obvious that there are no two characters or actions being the same in the four examples, but there are two functions in common: a tendency of *possessing the vehicle of magic by the hero*, of *getting the hero to the right place*, and the roles of *hero, bestowed,* and *associate* in all the stories.

Though Propp's analysis is not without psychological implications: behind the functions there are motivations, personality characters of actors, and functions arranged in cause-result schemata (that was exploited by story-grammar in the 1970's), the analyses unambiguously is directed

to the structuralist description of a definite corpus of short stories, of the magic tales. However, with some modifications the narrative features revealed by Propp afford the opportunity of an analysis of intra-psychic events.

Characters of the biographical episodes can also be classified from psychological point of view. Naturally, it is not an exclusive feature of narration: Mérei (1984) for instance draws conclusions about the order of signification in attachment from the repertoire of actors appearing in dream-contents. But narration, and especially biographical stories have a distinctive feature, namely that characters not only push the plot forward by their actions (so they have not only action-functions) but represent significant intra-psychic, psychological functions, which are related to personality development and personality states. Function of action of *help* and *protection* can be well understood by psychological function of *defence* and *security* and that makes a marked difference whether this function is executed by parent against child or by child against parent.

Thus, psychological content-analysis of characters and their functions does not take an arbitrary choice of textual components into account but is based on previously defined narrative features of texts by providing psychological understandings to these features. It is a kind of narrative psychological content analysis (cf. László et al., 2007, László, 2008) .

### *1.2 Codes*

Codes were generated for identifying characters' psychological functions in stories. These codes were based on characters' activity and attributes in story-situations. The following codes were identified:

- *anti-model*: The character possesses features or deeds that becomes a negative example for the narrator and the narrator explicitly refers to these.
- *traitor*: according to the narrator, the character informed someone about an intimate thing, a 'common secret' without the aware and approval of the narrator.
- *drug-friend*: relation to the character is connected to drug.
- *leaver*: the character left the narrator because of divorce, removal, breaking-up.
- *enemy*: according to the narrator, the character is against him/her, quarrelled with him/her, or antecedents made the narrator dislike the character.
- *lost*: the character died.

- *adult associate*: the character reinforces the narrator in his/her adult identity, this function is more concrete than general supporting function.
- *threatening person*: the character does or says something that risks the narrator's psychical and/or physical existence.
- *restricting person*: the character inhibits the narrator in his/her physical or mental achievements.
- *model:* the character possesses features or deeds that appear as a model for the narrator.
- *non-caring*: according to the narrator, the character does not care for someone, does not assure the requirements of security, of existence.
- *non-supportive*: the character does not help the narrator achieve his/her aims, does not support him/her in difficulties, but at the same time does not inhibits him/her either, is not against him/her.
- *partner*: the character appears as a lover, or in intimate relationship.
- *helper*: according to the narrator, the character supports the narrator to achieve physical or mental aims, is an active participant and is attentive.
- *fellow*: the character and the narrator suffers the same thing.
- *anguishing person*: the character appears for the narrator as someone threatening, this can be caused by some concrete deeds and / or by a 'state' of the character.
- *supportive*: the character is actively present, paves the way for someone instead of supporting him/her, but does not helps him/her directly.
- *associate*: the character and the narrator do something together, it is not a passive way of being together, and some common (symmetric) experience appear .
- *competitor:* the character appears for the narrator as someone rival: their aims are the same and they inhibit each other.
- *protector*: according to the narrator, the character provides security, protects from danger.
- *protégé*: the character requires the narrator's protection, and has the function of being someone who is protected by the narrator.

### *1.3 Method*

We worked out a structured life-interview-methodology to analyze the structure and function of self-object representations and mechanisms.

Subjects were asked to tell ten biographical stories. These biographical episodes were the following: 1. first good memory, 2. first bad memory, 3. a memory with parents, 4. a situation in which the subject were treated as an adult for the first time, 5. memory of the first partner-relationship, 6. a story about a hero to whom you wanted to bear a likeness to, 7. an achievement 8. a fearful situation the subject was unable to cope with, 9. a fearful situation the subject successfully coped with, 10. the memory of the first loiter. Time, scene, and characters were freely chosen by the subjects unless one episode, in which the characters (parents) were mandatory. Each story touches upon aspects of early self-object representations and significant components of identity development:

Episodes and *related representational aspects* are the following:

- First good memory, first bad memory, and a memory with parents: *representations and attainability of representations based on early experience*
- Fearful situation he/she could control or was unable to control: *defence and security*
- Memory of the first loiter: *separation, demand for separation*
- A hero you wanted/want to bear a likeness to: *models of identification and their attainability*
- Achievement: *self-esteem*
- Situation in which he/she was treated as an adult for the first time: *states and desires belonging to adult identity, and perception of the changing attitude of the environment*
- Memory of the first partner-relationship, significant relationship outside the family: *intimacy, engagement*

### *1.4 Research sample*

We employed two socio-demographically strictly matched samples: 20 drug-addict and 20 normal conduct youth between age 18 and 22. Matching covered sex (equal proportion), education (secondary school, university), education of parents (elementary school, secondary school, university) family background (wealthy, ordinary, poor), and family status (divorced or non-divorced parents).

### *1.5 Research procedure*

Interviews were made in the autumn of 1997 and in 1998. Drug-addicts were interviewed in the Drug Centre of Pécs and the control group in

the Psychological Institute of PTE BTK.[2] Steps of the analysis were the following:

1. When compiling the text of each subject, the interviewer's parts and 'social texts' belonging to the interview situation were omitted.
2. Each text was spell-checked.
3. All characters in all stories were coded in Atlas.ti. As a result, the „hero" and the „first partner relation" stories were omitted from the further analysis. „Hero" stories involved mainly characters from movies and novels and they were hard to analyze. „Partner" stories were crowded with ellipses and anaphoras what worked against safe pairing of characters and their functions.
4. Characters were categorized into six categories: *mother* (all the nicknames, and other versions and derivatives of the word 'mother') *father* (all the nicknames, other versions and derivatives of the word 'father'), *parent* (all the nicknames, other versions and derivatives of the word 'parent' *narrow family* (grandmother, grandfather, brother, sister, and all the nicknames, other versions and derivatives of these words), *broad family* (godmother and godfather, nephew, niece, brother-in-law, mother-in-law, uncle, aunt, fiancé, foster-father, foster-mother, step-parents, partner, great-granny, relatives, and all the derivatives of these words), *non-relatives* (girl, boy, János, director, Péter, lad, trainer, Adrien, teacher, acquaintance, driver, Hugo, old woman, boyfriend, girl-friend, friend, guy, chap, bloke, married-couple, neighbour, Kornél, Laci, Eszter, discipline, patron, representative, community and all the derivatives of these words).
5. 8-8 hyper-texts were built in Atlas.ti for both groups by collap-sing subjects' stories in each story type. It means that we conducted the analysis not on individual, but on group level. These texts were coded according to the psychological functions of the characters.

### *1.6 Coding*

Coding was performed in Atlas.ti (Muhr, 1991) by two independent coders. Their results showed a 92% agreement. Rules for coding were the following:

[2] We would like to thank Zsuzsa Laky, and János Szemelyácz of the Drug Centre of Pécs for bringing together the drug-addict group, and Melinda Pohárnok for bringing together the control group.

a. in one scene a function of a character was allowed to be coded only once.
b. in one scene more than one function of a character was allowed to be coded.
c. the same function was allowed to be coded for a character in different scenes.

### *1.7 Hypotheses*

Following the above-introduced theoretical framework, our hypo-theses are the following:

- drug abuse expresses developmental disorders at manifest behavioral level.
- drug-addict carrier can be traced back to defects of self-object representations, which are based on early interactional experiences.
- these representations can be examined in autobiographical narratives.

### *1.8 Results*

Table 1 shows the occurrences of the characters in the stories of the deviant (D) and the normal (N) groups.

The total number of characters is 1069. It is divided proportionally between the two groups: 559 in the drug-addicts' group and 510 in the normal group.

As the number of characters and functions were practically equal in both groups, probability of incidence of an character and of a function also could be regarded equal which allowed to make binomial trials on total results to find significant differences (Cohen, 1977)[3]. **Bold print** shows $p < 0.05$ significant results, whereas *italics* shows $p < 0.06$ strong tendency results.

Considering total characters by categories (mother, father, etc.) independently from episodes, we can see that there is no significant difference comparing the two groups. However, considering episodes one by one, we can see significant difference in the number of parents as characters both in the 'controlled' and 'no-control' situations: parents appear in significantly higher number in the drug-addicts' group. Broken down the total number of parents according to the three original

[3] We thank Mihály Sipos for helping in statistical processing.

categories, we find that 'mother' appears significantly more times in the 'controlled' situation while in 'no-control' situation categories of 'parent' and 'father' appear significantly more times in the drug-addicts' group. In the normal group members of the narrow and broad family appear significantly more times in 'no-control' situations.

Table 1
Distribution of characters by episodes

| CHARACTERS EPISODS | MO-THER | | FATHER | | PARENTS | | NARROW FAMILY | | BROAD FAMILY | | NON-RE-LATIVES | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | D | N | D | N | D | N | D | N | D | N | D | N |
| Parent | 43 | 43 | 40 | 45 | 5 | 2 | 19 | 23 | 2 | 2 | 15 | 10 |
| Good memory | 11 | 7 | 10 | 5 | 1 | 3 | 13 | 7 | 2 | 4 | 17 | 10 |
| Bad memory | *22* | *11* | 19 | 13 | 11 | 11 | 16 | 17 | 2 | 2 | 21 | 20 |
| Control | **13** | **4** | 11 | 6 | 0 | 1 | 2 | 7 | 1 | 1 | 31 | 21 |
| No control | 12 | 11 | *12* | *5* | **6** | **0** | 5 | 11 | 3 | 7 | 29 | 22 |
| Achievement | 6 | 9 | 3 | 5 | 0 | 2 | 3 | 4 | 0 | 1 | 21 | 12 |
| Loiter | 10 | 10 | 16 | 15 | 3 | 4 | 3 | 9 | 4 | 0 | 31 | 30 |
| Adult | 18 | 17 | 9 | 13 | 4 | 7 | 12 | 8 | 0 | 1 | 22 | 32 |
| Total | 135 | 112 | 120 | 107 | 30 | 30 | 73 | 86 | 14 | 18 | 187 | 157 |

In the 'bad memory' episode there are again significantly higher number of parents and especially more 'mother' characters in the drug-addicts' group.

In the 'loitering' episode, the numbers of the parent characters are equal both in total number (29:29) and by categories (10:10, 16:15, 3:4). However, the number of members of the narrow family is significantly higher in the same episode considering the normal group.

It is the 'efficiency' episode where the ratio of parents significantly changes, thus, this not only means that there is a significant difference between the groups, but this is the only episode in which the number of parents is higher in the normal group. The in-group ratio is remarkable in this episode as well: while the ratio of parents and non-relatives is 16:12 for the normal group, this ratio is 21:9 for the drug-addicts' group.

*Distribution of functions by episodes and in total number*

Table 2 shows the distribution of functions by episodes.

Table 2
Distribution of functions by episodes

| EPISODES | | Good memory | Bad memory | Parent | Adult | Achiev | Control | No control | Loiter | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CODES | | | | | | | | | | | Total |
| Antimodel | D | 0 | 2 | 2 | 1 | 0 | 2 | 0 | 0 | **7** | 8 |
| | N | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **1** | |
| Traitor | D | 0 | 1 | 0 | 0 | 0 | 2 | 2 | 0 | **5** | 5 |
| | N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | |
| Drug-friend | D | 3 | 0 | 2 | 0 | 2 | 2 | 6 | 0 | **15** | 15 |
| | N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | |
| Leaver | D | 3 | 12 | 9 | **9** | 0 | 1 | 2 | 2 | **38** | 59 |
| | N | 1 | 7 | 7 | **0** | 3 | 2 | 0 | 1 | **21** | |
| Enemy | D | 1 | 0 | 2 | 0 | 1 | 1 | 3 | 1 | 9 | 10 |
| | N | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | |
| Lost | D | 4 | 6 | 3 | 1 | 1 | 0 | 6 | 0 | 21 | 31 |
| | N | 2 | 2 | 1 | 2 | 0 | 2 | 1 | 0 | 10 | |
| Adult associate | D | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 11 | 18 |
| | N | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 7 | |
| Threatening | D | 1 | 6 | 3 | 0 | 0 | 5 | 3 | 7 | 25 | 48 |
| | N | 0 | 7 | 2 | 1 | 0 | 4 | 3 | 6 | 23 | |
| Restricting | D | 2 | 3 | 19 | 7 | **0** | 0 | 7 | **7** | **45** | 111 |
| | N | 4 | 7 | 21 | 9 | **5** | 1 | 3 | **16** | **66** | |
| Model | D | 2 | 6 | **4** | 1 | 4 | 0 | 2 | 0 | 19 | 48 |
| | N | 0 | 2 | **15** | 3 | 6 | 0 | 1 | 2 | 29 | |

| EPISODES | | Good memory | Bad memory | Parent | Adult | Achiev | Control | No control | Loiter | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CODES | | | | | | | | | | | Total |
| Non-caring | D | 1 | 7 | 2 | 3 | 6 | 3 | 1 | **5** | **28** | 41 |
| | N | 1 | 3 | 3 | 4 | 0 | 1 | 1 | **0** | **13** | |
| Non-support. | D | **7** | **0** | 11 | 0 | 0 | 1 | 6 | 0 | 25 | 54 |
| | N | **1** | **6** | 10 | 1 | 1 | 1 | 8 | 1 | 29 | |
| Partner | D | 3 | 0 | 1 | 2 | 0 | 0 | 2 | 3 | **11** | 15 |
| | N | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | **4** | |
| Helper | D | 1 | 1 | 0 | **0** | 5 | 1 | 0 | 1 | **9** | 30 |
| | N | 2 | 3 | 3 | **6** | 3 | 3 | 1 | 0 | **21** | |
| Fellow | D | 1 | 4 | 4 | 4 | 1 | 2 | 3 | 5 | 24 | 43 |
| | N | 0 | 0 | 9 | 2 | 0 | 3 | 2 | 2 | 18 | |
| Anguishing | D | 3 | **16** | 14 | 3 | 0 | 7 | 9 | **4** | **56** | 77 |
| | N | 0 | **4** | 8 | 2 | 0 | 2 | 5 | **0** | **21** | |
| Supportive | D | 17 | 21 | 26 | 16 | **3** | **19** | **9** | 17 | 128 | 261 |
| | N | 17 | 14 | 28 | 27 | **11** | **5** | **17** | 11 | 130 | |
| Associate | D | 3 | 4 | 11 | **1** | 2 | 4 | 1 | 14 | **40** | 106 |
| | N | 6 | 8 | 8 | **6** | 0 | 6 | 3 | 29 | **66** | |
| Competitor | D | 2 | **0** | 8 | 1 | 2 | 1 | 1 | 1 | 16 | 34 |
| | N | 2 | **5** | 4 | 1 | 1 | 4 | 1 | 0 | 18 | |
| Protector | D | 1 | 0 | 2 | 0 | 0 | **0** | 1 | 1 | **5** | 20 |
| | N | 0 | 3 | 2 | 1 | 1 | **4** | 4 | 0 | **15** | |
| Protégé | D | 1 | 2 | 2 | 4 | 4 | **8** | 2 | 0 | 23 | 39 |
| | N | 0 | 2 | 2 | 4 | 2 | **1** | 5 | 0 | 16 | |

There are significant differences between the two groups in the following functions: anti-model, traitor, drug-friend, leaver, enemy, restricting person, non-caring, helper, anguishing, associate, protector, lost and partner.

There are significantly more characters in the anti-model, traitor, drug-friend, leaver, enemy, non-caring, and anguishing functions in the drug-addicts' group. There are significantly more characters in 'restricting person', 'helper', 'associate', and 'protector' functions in the normal group.

When divided the functions into two groups according to negative versus positive features, we received the following results (positive functions are: model, helper, supportive, and protector; negative functions are: anti-model, traitor, leaver, enemy, lost, threatening person, non-caring, and anguishing):

Table 3
Distribution of positive and negative functions

| FUNCTION | DRUG | NORMAL |
|---|---|---|
| Positive | 160 | 198 |
| Negative | 214 | 119 |

The difference is significant both with positive and with negative functions ($p < 0.05$).

There is significant difference in the distribution of functions by episodes in the 'no-control' situation and in the 'drug-friend' function (see Table 2). Naturally, this function is not exist in the normal group, but this is the only episode where this function is present in great number in the drug-addicts' group. Considering the same episode in the normal group, we find significantly more supportive function. The leaver function appears significantly more times in the 'adult' episode in the drug-addicts' group. The restricting person function has a significant difference in the 'efficiency' and in the 'loiter' episodes, both occur more in the normal group. The *model* function appears significantly more times in the 'parent' episode in the normal group. The non-caring function occurs significantly more times in the drug-addicts' group in the 'efficiency' and in the 'loiter' episodes. The non-supportive function has an interesting occurrence. Considering the total number of episodes, we find only a small difference between the groups (25:29). But there are significant differences in two cases: in the 'good memory' and in the 'bad memory' episodes. The non-supportive function

occurs significantly more times in the 'good memory' considering the normal group and in the 'bad memory' considering the drug-addicts' group. The helper function in the 'adult' episode has a significant difference in the normal group.

The anguishing function appears significantly more times in the 'bad memory' and in the 'loiter' episodes in the drug-addicts' group. The associate function has a significant occurrence in the normal group in general and in the 'adult' episode in particular. In the 'controlled' situation the protector function appears significantly more times in the normal group and the protégé function in the drug-addicts' group.

It seems to be useful to categorize some functions from content based aspects (Table 4) distrust: traitor, enemy defencelessness (uncontrollability): lost, leaver, threatening person, anguishing active supply of security: helper, protector.

Table 4
Distribution of function clusters

| FUNCTION CLUSTERS | DRUG | NORMAL |
|---|---|---|
| Distrust | 14 | 1 |
| Defenselessness | 140 | 75 |
| Active supply of security | 13 | 36 |

The difference is significant in all three cases ($p < 0.05$).
Table 5 shows distribution of functions summed by characters

- The *mother* appears significantly more times in the anguishing, non-caring, and leaver functions in the drug-addicts' group.
- The *father* is in greater number in the anguishing, lost, and anti-model functions. The normal group exhibits significantly higher number in the protector function.
- The *parents* (summary of codes of mother, father, parent) appear significantly more times in the anti-model, lost, leaver, and anguishing functions in the drug-addicts' group. In the normal group the helper and protector functions have significant occurrence.
- Members of the *narrow family* appear significantly more times in the associate and restricting person functions in the normal group.
- *Non-relatives* occur significantly more times in the traitor, drug-friend, and partner functions in the drug-addicts' group and in restricting and supportive functions in the normal group.

Table 5
Distribution of functions by characters

| | MOTHER | | FATHER | | PARENT | | NARROW FAMILY | | BROAD FAMILY | | NON-RELATIVE | | □ PARENTS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | D | N | D | N | D | N | D | N | D | N | D | N | D |
| Antimodel | 0 | 1 | **0** | **4** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | **0** | **5** |
| Traiter | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | **5** | 0 | 0 |
| Drug-friend | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | **15** | 0 | 0 |
| Leaver | **6** | **13** | 8 | 9 | 2 | 6 | 3 | 4 | 0 | 0 | 2 | 6 | **16** | **28** |
| Enemy | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 1 | 5 | 0 | 1 |
| Lost | 2 | 5 | **0** | **6** | 0 | 0 | 5 | 7 | 2 | 2 | 1 | 1 | **2** | **11** |
| Adult assoc. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 11 | 0 | 0 |
| Threatening | 3 | 2 | 12 | 16 | 0 | 0 | 1 | 1 | 1 | 0 | 6 | 6 | 15 | 18 |
| Restricting | 26 | 19 | 19 | 15 | 6 | 8 | **4** | **0** | 1 | 0 | **10** | **3** | 51 | 42 |
| Model | 12 | 5 | 9 | 8 | 0 | 0 | 6 | 5 | 1 | 0 | 1 | 1 | 21 | 13 |
| Non-caring | **6** | **13** | 5 | 8 | 2 | 1 | 0 | 3 | 0 | 0 | 0 | 3 | 13 | 22 |
| Non-support. | 9 | 7 | 5 | 10 | 3 | 1 | 1 | 0 | 0 | 0 | 11 | 6 | 17 | 18 |
| Partner | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **4** | **11** | 0 | 0 |
| Helper | 5 | 2 | 6 | 1 | 0 | 0 | 2 | 0 | 1 | 1 | 7 | 5 | **11** | **3** |
| Fellow | 2 | 0 | 0 | 0 | 0 | 0 | 10 | 3 | 0 | 0 | 6 | 11 | 2 | 0 |
| Anguishing | **7** | **17** | **10** | **22** | 1 | 5 | 0 | 3 | 1 | 0 | 2 | 9 | **18** | **44** |
| Supportive | 28 | 38 | 26 | 16 | 13 | 9 | 22 | 7 | 0 | 7 | 31 | 41 | 67 | 63 |
| Associate | 0 | 3 | 0 | 0 | 0 | 0 | **17** | **8** | 0 | 1 | **49** | **28** | 0 | 3 |
| Competitor | 0 | 0 | 1 | 0 | 0 | 0 | 7 | 7 | 0 | 2 | 10 | 7 | 1 | 0 |
| Protector | 1 | 3 | **5** | **0** | 3 | 0 | 3 | 2 | 0 | 0 | 3 | 0 | **9** | **3** |
| Protégé | 5 | 7 | 1 | 4 | 0 | 0 | 4 | 1 | 0 | 0 | 6 | 11 | 6 | 11 |

## *2. DISCUSSION*

Whereas we have not found significant difference in the total number of parents, normal subjects mention their parents significantly less time in the 'controlled', 'no-control', and 'bad-memory' episodes. These episodes refer to early experiences and to their attainability and 'value', to defence, to security, and to the experience of defencelessness. Interpretation of this result in terms of object relations theory is that normal subjects have an internalized 'good object' and it makes the explicit occurrence of parents in these stories less necessary. This interpretation is also supported by the fact that it was both difficult to separate the 'good memory'-'bad memory' and the 'controlled'-no-control' episodes and to make the interviewees see these episodes as separate in the drug-addicts' group.

Another converging evidence is that it is only in the 'efficiency situation' where the proportion of parents is reversed. Drug addicts, as opposed to normal subjects, rather mention non-relatives. Appearance of their parents is very limited. Appreciation and the value of one's own results are connected to whether parents regard it valuable. Stories refer not only to proudness but to the value of the self. Efficiency and being proud satisfy narcissistic needs both for the parents and for the child. It is crucial that the parent has to be able to be receptive and to support the child (and to see her own narcissistic needs independently). This result is underscored by the distribution of the characters' functions. In efficiency episodes parents' supportive function appear in much greater number in the normal group, whereas mother's and parents' non-caring function is typical in the drug-addicts' group.

To summarize, the distribution of characters across episodes supports the assumption about parents' involvement in self development as well as about the relation between later maladaptive behavior and disturbances in early self development.

Similar conclusion can be drawn with characters' functions. Comparing to the normal group, there are much more negative and much less positive functions in the drug-addicts' group.

This picture becomes even sharper when we combine most frequent characters with their most typical functions in both groups, as we did it in Table 6. We considered only those functions that occur significantly more times comparing to the same character of the other group.

Table 6
Characters' typical functions

| | DRUG | NORMAL |
|---|---|---|
| MOTHER | anguishing, non-caring, leaver | |
| FATHER | anguishing, lost, anti-model | Protective |
| PARENT | anguishing, leaver, lost, anti-model | protective, helper |
| NARROW F. | | restricting, associate |
| NON-RELAT. | traitor, drug-friend, partner | restricting. supportive |

Frequencies of parents' functions reflect typical features of the representation of parents. For drug addicts, parents' most typical functions are *anguishing,-non-caring-leaver-lost-anti-model*. In the normal group *protective-helper functions dominate*. Considering other characters (narrow family, non-relative) with significant differences between the two groups, (*drug-friend-traitor- partner* for the drug-addicts and *restrictive-supportive* for the normal subjects), we see that parental representations somehow expand to other interpersonal relations.

We have started with the assumption that self-development is embedded in interactive experience. Self structures are based on interaction structures. Giving meaning from the perspective of the self and organization of events in narratives is continuous from early developmental period on. We have suggested that narrative meaning construction is highly dependent on representation of parents. Adaptive and maladaptive states of the self can be traced back by analyzing psychological functions of parents and other characters in life narratives. Our results gained by quantitative narrative text analysis provided evidence for the assumption that maladaptive behavior is a consequence of disturbed early self-development.

## *REFERENCES*

Balint, M. (1939): Early States of Ego Development: Primary Object – Love, in: *International Journal of Psychoanalysis,* 30, 265-273.

Beebe, B., Lachmann, F. M., Jaffe, J. (1997): Mother- infant interaction structures and presymbolic self and object representations, in: *Psychoanalytic Dialogues*, 7, 133-182.

Bion, W. (1984): *Learning from Experience*. London: Karnac Books.

Blos, P., (1962): *On Adolescence, A Psychoanalytic Interpretation*. New York: Free Press of Glencoe.

Bowlby, J. (1982): Attachment and Loss: Retrospect and Prospect, in: *American Journal of Orthopsychiatry*, 52, 664-678.

Bucci, W. (1997): *Psychoanalysis and Cognitive Science; A Multiple Code Theory.* New York: Guilford Press.

Byng-Hall, J., Stevenson-Hinde, J. (1991): Attachment relationships within a family system, in: *Infant Mental Health Journal*, 12, 2, 3-13.

Cohen, J. (1977): *Statistical power analysis for the behavioral sciences*. NY, San Francisco, London.

Emde, R.N. (1999): *Moving ahead: integrating influences of affective processes for development and for psychoanalysis*. Paper presented at the 41st Congress of the International Psychoanalytical Association in Santiago, in: International Journal of Psycho-Analysis, 80, 317-339.

Erikson, E. H., (I974): *Identity, Youth und crisis*. London: Faber and Faber.

Fonagy, P., Steele, M., Steele, H., Leigh, T., Kennedy R., Mattoon, G., Target, M. (1995): Attachment, the reflective self, and borderline states: The predictive specificity of the Adult Interview and pathological emotional development. In: S. Goldberg, R. Muir, J. Kerr (eds.): *Attachment theory: Social, developmental and clinical perspectives*. New York: Analytic Press, 233-278.

Gergely, G. (1992): Developmental reconstructions: Infancy from the point of view of psychoanalysis and developmental psychology, in: *Psychoanalysis and Contemporary Thought,* 15, 1, 3-55.

Hámori E., Péley B. (1999): A pszichoanalitikus rekonstrukció problémája - Mahler fejlődési modellje és ennek kortárs csecsemőkutatás szempontú kritikái a gyermekpszichoterápiás gyakorlat szemszögéből. (The problem of psychoanalytical reconstruction), in: *Pszichoterápia,* 8, 1, 5-13.

Josselson, R., (1988): The Embedded Self: We and Thou Revisited. In: Lapsley, D. K., Power, F. C. (eds.), *Self, Ego and Identity, Integrative* Approaches, Springer Verlag, 91-106.

Klein, M. (1975): *Collected Works*. London: Hogarth Press and Institute of Psychoanalysis.

Mahler, M., Pine, B., Bergman, A. (1975): *The Psychological Birth of the Human Infant: Symbiosis and Individuation*. New York: Basic Books.

McAdams, D.P. (1988): *Power, Intimacy, and the Life Story. Personological Inquiries into Identity.* The Guilford Press.

Mérei F. (1984): *Lélektani napló III. Az implikált tudás az álomban (Implied knowledge in dreams).* Budapest: Művelődéskutató Intézet.

Muhr, T, (1991): ATLA/ti- A prototype for the support of text interpretation, in: *Qualitative Sociology, 14,* 349-371.

Propp, V.J. (1968): *Morphology of the Folk Tales*. Austin, London: Texas University Press.

Schafer, R. (1980): Narration in the psychoanalytic dialogue, in: *Critical Inquiry*, 7, 29-53.

Shields, A., Ryan, R.M. & Cicchetti, D. (2001): Narrative Representations of Caregivers and Emotion Dysregulation as Predictors of Maltreated Children's Rejection by Peers, in: *Developmental Psychology* Vol. 37, No. 3, 321-337.

Snow, C.E. (1990): Building Memories: The Ontogeny of Autobiography. In.: D. Cicchetti and M. Beeghly (eds.): *The Self in Transition*. *Infancy to childhood.* University of Chicago Press. 213-242.

Spence, D. P. (1982): *Narrative Truth and Historical Truth. Meaning and Interpretation in Psychoanalysis*. New York: Norton.

Sroufe, L. A. (1990): An organizational perspective on the self. In: D. Cicchetti and M. Beeghly (eds.): *The self in transition: Infancy to childhood.* Chicago: University of Chicago Press, 281-307.

Stern, D. N. (1985): *The Interpersonal World of the Infant*. New York: Basic Books.

Stern, D.N. (1995): *The Motherhood Constellation.* New York: Basic Books.

Winnicott, D. W., (1985): *Playing and Reality*, Pelican Books.

Winnicott, D.W. (1990): *The Maturational Processes and the Facilitating Environment.* London: Karnac Books.

Wolf, D. P. (1990): Being of Several Minds: Voices and Versions of the self in Early Childhood. In: D. Cicchetti and M. Beeghly (eds.): *The Self in Transition*. *Infancy to childhood.* University of Chicago Press, 183-212.

# A modified text indicator

*Ioan-Iovitz Popescu (Bucharest, Romania),*
*Gabriel Altmann (Lüdenscheid, Germany)*

In several previous works on ranked frequencies of word-forms, the *h*-point has been introduced as a fuzzy boundary dividing two strata of words, namely autosemantics and synsemantics. The *h*-point has been used to establish measures of thematic concentration, auto-semantic compactness and other characteristic properties of texts, and could be used also in language typology (cf. Popescu et al. 2009). At last, the developers settled *h* at that point at which the bisector $y = x$ intersects the straight line joining two neighbouring rank-frequencies. Hence, the simplest way to compute the *h*-point is

$$h = \begin{cases} r & \text{if there is an } r = f(r) \\ \dfrac{f(r_i)r_j - f(r_j)r_i}{r_j - r_i + f_i - f_j} & \text{if there is no } r = f(r) \end{cases} \qquad (1)$$

In the second case one takes two neighbouring ranks for which $r_i < f(r_i)$ and $r_j > f(r_j)$. In case that $r_{max} < f(r_{max})$ one transforms the frequency sequence in $f^*(r) = f(r) - f(r_{max}) + 1$. The *h*-point and its further uses have been demonstrated in several publications (cf. e.g. Popescu, Altmann 2008, Popescu, Mačutek, Altmann 2009; Laufer, Nemcová 2009, Kelih in this volume). The *h*-point increases with increasing *N* (text length). Hirsch (2005) observed that the relationship between *N* and *h* can be expressed as

$$N = ah^2 \qquad (2)$$

and it has been shown that the indicator *a* is a characteristic of analyticity/syntheticity of language (cf. Popescu et al. 2009 where also a test for differences between two *a*-indicators has been presented).

However, it seems more appropriate if one redefines (2) because *N* is the area under the rank-frequency sequence defined by the points

(1, $f(1)$), (1, 1), and ($R$, 1) whose zero-point is (1,1) and not (0,0), where $R$ is the vocabulary (= highest rank $r_{max}$). Also the natural $h$-square unit is defined by the points (1, 1), (1, h), ($h$, $h$), and ($h$, 1), and has the area $(h - 1)^2$. Hence we propose in the present paper a new version for the relationship (2), namely

$$N = b(h - 1)^2 , \qquad (3)$$

where the old indicator $a$ and the new one $b$ are related by

$$a/b = (1 - 1/h)^2. \qquad (4)$$

Obviously, the difference between relations (2) and (3), respectively between $a$ and $b$, for very great $h$ can be neglected.

Now, in order to characterize the rank order sequence of word frequencies, the indicator

$$p = \frac{L_{\max} - L}{h - 1} \qquad (5)$$

has been introduced (cf. Popescu, Mačutek, Altmann 2009; Nemcová, Popescu, Altmann 2009) and tested on different language phenomena. The maximum arc length is defined as

$$L_{max} = R - 1 + f(1) - 1, \qquad (6)$$

and the arc length between <1, $f(1)$> and <$R$, 1> as the sum of Euclidean distances between the ranked frequencies, i.e.

$$L = \sum_{r=1}^{R-1} [(f(r) - f(r+1))^2 + 1]^{1/2} . \qquad (7)$$

Considering expression (3), another indicator can be established, namely

$$q = \frac{L_{max} - L}{N^{1/2}} , \qquad (8)$$

quite similar to $p$, however $N^{1/2}$ has been taken into account instead of $(h-1)$. Since both $L_{max}$ and $L$ increase with increasing $N$, the increase is relativized. Finally, the three indicators $b$, $p$, $q$ are interrelated by the simple relation

$$b = \frac{N}{(h-1)^2} = \left(\frac{p}{q}\right)^2 . \qquad (9)$$

Unlike $p$, the new $q$-indicator is able to express both differences between individual texts and in form of averages also differences between languages. In order to demonstrate this fact, let us consider first the individual values in 100 texts from 20 languages (cf. Table 1). The values of $p$, $q$, $b$ have been computed according to (5), (8), (9) respectively.

Table 1
Indicators (5), (8) and (9) in 100 texts from 20 languages
(B = Bulgarian, Cz = Czech, E = English, G = German, H = Hungarian, Hw = Hawaiian, I = Italian, In = Indonesian, Kn = Kannada, Lk = Lakota, Lt = Latin, M = Maori, Mq = Marquesan, Mr = Marathi, R = Romanian, Rt = Rarotongan, Ru = Russian, Sl =Slovenian, Sm = Samoan, T = Tagalog)

| ID | $N$ | $V$ | $f(1)$ | $h$ | $L$ | $L_{max}$ | $p$ | $q$ | $b$ |
|---|---|---|---|---|---|---|---|---|---|
| B 01 | 761 | 400 | 40 | 10 | 428 | 438 | 1.111 | 0.362 | 9.395 |
| B 02 | 352 | 201 | 13 | 8 | 205 | 212 | 1.000 | 0.373 | 7.184 |
| B 03 | 515 | 285 | 15 | 9 | 290 | 298 | 1.000 | 0.353 | 8.047 |
| B 04 | 483 | 286 | 21 | 8 | 297 | 305 | 1.143 | 0.364 | 9.857 |
| B 05 | 406 | 238 | 19 | 7 | 247 | 255 | 1.333 | 0.397 | 11.278 |
| Cz 01 | 1044 | 638 | 58 | 9 | 684 | 694 | 1.250 | 0.309 | 16.313 |
| Cz 02 | 984 | 543 | 56 | 11 | 586 | 597 | 1.100 | 0.351 | 9.840 |
| Cz 03 | 2858 | 1274 | 182 | 19 | 1432 | 1454 | 1.222 | 0.412 | 8.821 |
| Cz 04 | 522 | 323 | 27 | 7 | 342 | 348 | 1.000 | 0.263 | 14.500 |
| Cz 05 | 999 | 556 | 84 | 9 | 627 | 638 | 1.375 | 0.348 | 15.609 |
| E 01 | 2330 | 939 | 126 | 16 | 1043 | 1063 | 1.333 | 0.414 | 10.356 |
| E 02 | 2971 | 1017 | 168 | 22 | 1157 | 1183 | 1.238 | 0.477 | 6.737 |
| E 03 | 3247 | 1001 | 229 | 19 | 1205 | 1228 | 1.278 | 0.404 | 10.022 |
| E 04 | 4622 | 1232 | 366 | 23 | 1567 | 1596 | 1.318 | 0.427 | 9.550 |
| E 05 | 4760 | 1495 | 297 | 26 | 1761 | 1790 | 1.160 | 0.420 | 7.616 |
| E 07 | 5004 | 1597 | 237 | 25 | 1801 | 1832 | 1.292 | 0.438 | 8.688 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| E 13 | 11265 | 1659 | 780 | 41 | 2388 | 2437 | 1.225 | 0.462 | 7.041 |
| G 05 | 559 | 332 | 30 | 8 | 351 | 360 | 1.286 | 0.381 | 11.408 |
| G 09 | 653 | 379 | 30 | 9 | 398 | 407 | 1.125 | 0.352 | 10.203 |
| G 10 | 480 | 301 | 18 | 7 | 310 | 317 | 1.167 | 0.320 | 13.333 |
| G 11 | 468 | 297 | 18 | 7 | 307 | 313 | 1.000 | 0.277 | 13.000 |
| G 12 | 251 | 169 | 14 | 6 | 175 | 181 | 1.200 | 0.379 | 10.040 |
| G 14 | 184 | 129 | 10 | 5 | 133 | 137 | 1.000 | 0.295 | 11.500 |
| G 17 | 225 | 124 | 11 | 6 | 128 | 133 | 1.000 | 0.333 | 9.000 |
| H 01 | 2044 | 1079 | 225 | 12 | 1289 | 1302 | 1.182 | 0.288 | 16.893 |
| H 02 | 1288 | 789 | 130 | 8 | 907 | 917 | 1.429 | 0.279 | 26.286 |
| H 03 | 403 | 291 | 48 | 4 | 332 | 337 | 1.667 | 0.249 | 44.778 |
| H 04 | 936 | 609 | 76 | 7 | 674 | 683 | 1.500 | 0.294 | 26.000 |
| H 05 | 413 | 290 | 32 | 6 | 314 | 320 | 1.200 | 0.295 | 16.520 |
| Hw 03 | 3507 | 521 | 277 | 26 | 764 | 796 | 1.280 | 0.540 | 5.611 |
| Hw 04 | 7892 | 744 | 535 | 38 | 1229 | 1277 | 1.297 | 0.540 | 5.765 |
| Hw 05 | 7620 | 680 | 416 | 38 | 1047 | 1094 | 1.270 | 0.538 | 5.566 |
| Hw 06 | 12356 | 1039 | 901 | 44 | 1877 | 1938 | 1.419 | 0.549 | 6.683 |
| I 01 | 11760 | 3667 | 388 | 37 | 4007 | 4053 | 1.278 | 0.424 | 9.074 |
| I 02 | 6064 | 2203 | 257 | 25 | 2426 | 2458 | 1.333 | 0.411 | 10.528 |
| I 03 | 854 | 483 | 64 | 10 | 534 | 545 | 1.222 | 0.376 | 10.543 |
| I 04 | 3258 | 1237 | 118 | 21 | 1330 | 1353 | 1.150 | 0.403 | 8.145 |
| I 05 | 1129 | 512 | 42 | 12 | 537 | 552 | 1.364 | 0.446 | 9.331 |
| In 01 | 376 | 221 | 16 | 6 | 228 | 235 | 1.400 | 0.361 | 15.040 |
| In 02 | 373 | 209 | 18 | 7 | 219 | 225 | 1.000 | 0.311 | 10.361 |
| In 03 | 347 | 194 | 14 | 6 | 200 | 206 | 1.200 | 0.322 | 13.880 |
| In 04 | 343 | 213 | 11 | 5 | 217 | 222 | 1.250 | 0.270 | 21.438 |
| In 05 | 414 | 188 | 16 | 8 | 196 | 202 | 0.857 | 0.295 | 8.449 |
| Kn 003 | 3188 | 1833 | 74 | 13 | 1891 | 1905 | 1.167 | 0.248 | 22.139 |
| Kn 004 | 1050 | 720 | 23 | 7 | 733 | 741 | 1.333 | 0.247 | 29.167 |
| Kn 005 | 4869 | 2477 | 101 | 16 | 2558 | 2576 | 1.200 | 0.258 | 21.640 |
| Kn 006 | 5231 | 2433 | 74 | 20 | 2481 | 2505 | 1.263 | 0.332 | 14.490 |
| Kn 011 | 4541 | 2516 | 63 | 17 | 2558 | 2577 | 1.188 | 0.282 | 17.738 |
| Lk 01 | 345 | 174 | 20 | 8 | 185 | 192 | 1.000 | 0.377 | 7.041 |
| Lk 02 | 1633 | 479 | 124 | 17 | 580 | 601 | 1.313 | 0.520 | 6.379 |
| Lk 03 | 809 | 272 | 62 | 12 | 318 | 332 | 1.273 | 0.492 | 6.686 |
| Lk 04 | 219 | 116 | 18 | 6 | 126 | 132 | 1.200 | 0.405 | 8.760 |
| Lt 01 | 3311 | 2211 | 133 | 12 | 2328 | 2342 | 1.273 | 0.243 | 27.364 |
| Lt 02 | 4010 | 2334 | 190 | 18 | 2502 | 2522 | 1.176 | 0.316 | 13.875 |
| Lt 03 | 4931 | 2703 | 103 | 19 | 2783 | 2804 | 1.167 | 0.299 | 15.219 |
| Lt 04 | 4285 | 1910 | 99 | 20 | 1983 | 2007 | 1.263 | 0.367 | 11.870 |
| Lt 05 | 1354 | 909 | 33 | 8 | 930 | 940 | 1.429 | 0.272 | 27.633 |
| Lt 06 | 829 | 609 | 19 | 7 | 621 | 626 | 0.833 | 0.174 | 23.028 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| M 01 | 2062 | 398 | 152 | 18 | 527 | 548 | 1.235 | 0.462 | 7.135 |
| M 02 | 1175 | 277 | 127 | 15 | 386 | 402 | 1.143 | 0.467 | 5.995 |
| M 03 | 1434 | 277 | 128 | 17 | 385 | 403 | 1.125 | 0.475 | 5.602 |
| M 04 | 1289 | 326 | 137 | 15 | 444 | 461 | 1.214 | 0.474 | 6.577 |
| M 05 | 3620 | 514 | 234 | 26 | 715 | 746 | 1.240 | 0.515 | 5.792 |
| Mq 01 | 2330 | 289 | 247 | 22 | 507 | 534 | 1.286 | 0.559 | 5.283 |
| Mq 02 | 457 | 150 | 42 | 10 | 179 | 190 | 1.222 | 0.515 | 5.642 |
| Mq 03 | 1509 | 301 | 218 | 14 | 500 | 517 | 1.308 | 0.438 | 8.929 |
| Mr 001 | 2998 | 1555 | 75 | 14 | 1612 | 1628 | 1.231 | 0.292 | 17.740 |
| Mr 018 | 4062 | 1788 | 126 | 20 | 1890 | 1912 | 1.158 | 0.345 | 11.252 |
| Mr 026 | 4146 | 2038 | 84 | 19 | 2099 | 2120 | 1.167 | 0.326 | 12.796 |
| Mr 027 | 4128 | 1400 | 92 | 21 | 1468 | 1490 | 1.100 | 0.342 | 10.320 |
| Mr 288 | 4060 | 2079 | 84 | 17 | 2141 | 2161 | 1.250 | 0.314 | 15.859 |
| R 01 | 1738 | 843 | 62 | 14 | 886 | 903 | 1.308 | 0.408 | 10.284 |
| R 02 | 2279 | 1179 | 110 | 16 | 1269 | 1287 | 1.200 | 0.377 | 10.129 |
| R 03 | 1264 | 719 | 65 | 12 | 770 | 782 | 1.091 | 0.338 | 10.446 |
| R 04 | 1284 | 729 | 49 | 10 | 764 | 776 | 1.333 | 0.335 | 15.852 |
| R 05 | 1032 | 567 | 46 | 11 | 599 | 611 | 1.200 | 0.374 | 10.320 |
| R 06 | 695 | 432 | 30 | 10 | 452 | 460 | 0.889 | 0.303 | 8.580 |
| Rt 01 | 968 | 223 | 111 | 14 | 316 | 332 | 1.231 | 0.514 | 5.728 |
| Rt 02 | 845 | 214 | 69 | 13 | 265 | 281 | 1.333 | 0.550 | 5.868 |
| Rt 03 | 892 | 207 | 66 | 13 | 256 | 271 | 1.250 | 0.502 | 6.194 |
| Rt 04 | 625 | 181 | 49 | 11 | 216 | 228 | 1.200 | 0.480 | 6.250 |
| Rt 05 | 1059 | 197 | 74 | 15 | 251 | 269 | 1.286 | 0.553 | 5.403 |
| Ru 01 | 753 | 422 | 31 | 8 | 441 | 451 | 1.429 | 0.364 | 15.367 |
| Ru 02 | 2595 | 1240 | 138 | 16 | 1357 | 1376 | 1.267 | 0.373 | 11.533 |
| Ru 03 | 3853 | 1792 | 144 | 21 | 1909 | 1934 | 1.250 | 0.403 | 9.633 |
| Ru 04 | 6025 | 2536 | 228 | 25 | 2732 | 2762 | 1.250 | 0.386 | 10.460 |
| Ru 05 | 17205 | 6073 | 701 | 41 | 6722 | 6772 | 1.250 | 0.381 | 10.753 |
| Sl 01 | 756 | 457 | 47 | 9 | 494 | 502 | 1.000 | 0.291 | 11.813 |
| Sl 02 | 1371 | 603 | 66 | 13 | 651 | 667 | 1.333 | 0.432 | 9.521 |
| Sl 03 | 1966 | 907 | 102 | 13 | 991 | 1007 | 1.333 | 0.361 | 13.653 |
| Sl 04 | 3491 | 1102 | 328 | 21 | 1404 | 1428 | 1.200 | 0.406 | 8.728 |
| Sl 05 | 5588 | 2223 | 193 | 25 | 2385 | 2414 | 1.208 | 0.388 | 9.701 |
| Sm 01 | 1487 | 267 | 159 | 17 | 403 | 424 | 1.313 | 0.545 | 5.809 |
| Sm 02 | 1171 | 222 | 103 | 15 | 304 | 323 | 1.357 | 0.555 | 5.974 |
| Sm 03 | 617 | 140 | 45 | 13 | 168 | 183 | 1.250 | 0.604 | 4.285 |
| Sm 04 | 736 | 153 | 78 | 12 | 214 | 229 | 1.364 | 0.553 | 6.083 |
| Sm 05 | 447 | 124 | 39 | 11 | 149 | 161 | 1.200 | 0.568 | 4.470 |
| T 01 | 1551 | 611 | 89 | 14 | 681 | 698 | 1.308 | 0.432 | 9.178 |
| T 02 | 1827 | 720 | 107 | 15 | 807 | 825 | 1.286 | 0.421 | 9.321 |
| T 03 | 2054 | 645 | 128 | 19 | 749 | 771 | 1.222 | 0.485 | 6.340 |

The variability and independence of $p$ and $q$ in terms of the text size $N$ is illustrated in Figure 1.

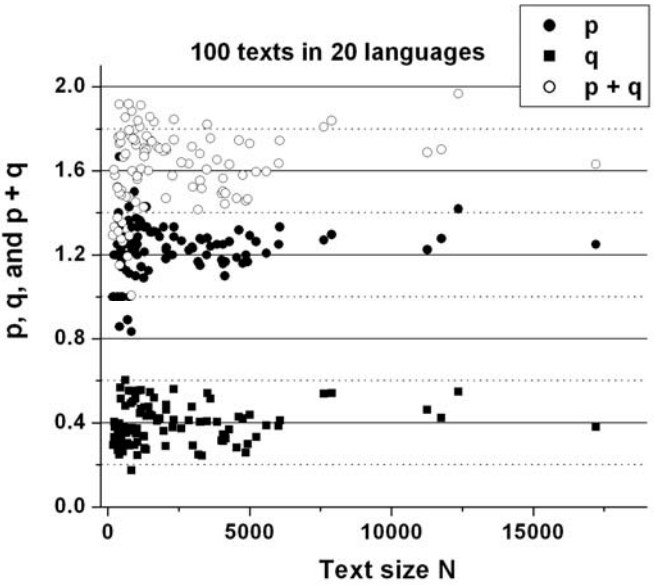



Figure 1. Showing $p$, $q$, and their sum in terms of the text size $N$ for the data of Table 1

The differentiation with one language is presented in Table 2 showing the individual $p$, $q$, $b$ values of 253 texts of 26 German writers covering the period 1729 – 2001.

Table 2
Indicators $p$, $q$, $b$ in 253 texts of 26 German writers

| ID | $N$ | $V$ | $f(1)$ | $h$ | $L$ | $L_{max}$ | $p$ | $q$ | $b$ |
|---|---|---|---|---|---|---|---|---|---|
| Arnim 01 | 7846 | 2221 | 271 | 33 | 2448 | 2490 | 1.313 | 0.474 | 7.662 |
| Arnim 02 | 1201 | 564 | 46 | 13 | 595 | 608 | 1.123 | 0.389 | 8.340 |
| Arnim 03 | 4167 | 1429 | 189 | 26 | 1588 | 1616 | 1.120 | 0.434 | 6.667 |
| Busch 01 | 15820 | 4642 | 527 | 44 | 5112 | 5167 | 1.279 | 0.437 | 8.556 |
| Chamisso 01 | 2210 | 884 | 82 | 18 | 944 | 964 | 1.176 | 0.425 | 7.647 |
| Chamisso 02 | 1847 | 808 | 84 | 16 | 872 | 890 | 1.200 | 0.419 | 8.209 |
| Chamisso 03 | 1428 | 630 | 70 | 14 | 684 | 698 | 1.077 | 0.370 | 8.450 |
| Chamisso 04 | 3205 | 1209 | 123 | 20 | 1305 | 1330 | 1.316 | 0.442 | 8.878 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Chamisso 05 | 2108 | 853 | 79 | 18 | 911 | 930 | 1.118 | 0.414 | 7.294 |
| Chamisso 06 | 1948 | 801 | 75 | 17 | 853 | 874 | 1.313 | 0.476 | 7.609 |
| Chamisso 07 | 1362 | 670 | 44 | 13 | 698 | 712 | 1.167 | 0.379 | 9.458 |
| Chamisso 08 | 1870 | 788 | 80 | 16 | 848 | 866 | 1.200 | 0.416 | 8.311 |
| Chamisso 09 | 1320 | 593 | 96 | 14 | 673 | 687 | 1.077 | 0.385 | 7.811 |
| Chamisso 10 | 1012 | 536 | 52 | 11 | 575 | 586 | 1.100 | 0.346 | 10.120 |
| Chamisso 11 | 1386 | 656 | 66 | 14 | 705 | 720 | 1.154 | 0.403 | 8.201 |
| Droste 01 | 16172 | 4064 | 525 | 49 | 4528 | 4587 | 1.229 | 0.464 | 7.019 |
| Droste 02 | 884 | 492 | 48 | 10 | 527 | 538 | 1.275 | 0.370 | 11.897 |
| Droste 03 | 700 | 425 | 31 | 9 | 444 | 454 | 1.240 | 0.375 | 10.938 |
| Droste 04 | 786 | 408 | 34 | 11 | 430 | 440 | 1.084 | 0.367 | 8.709 |
| Droste 05 | 1274 | 657 | 51 | 13 | 692 | 706 | 1.216 | 0.392 | 9.633 |
| Droste 08 | 965 | 509 | 39 | 11 | 535 | 546 | 1.145 | 0.369 | 9.650 |
| Eichendorff 01 | 3080 | 1079 | 177 | 21 | 1228 | 1254 | 1.300 | 0.468 | 7.700 |
| Eichendorff 02 | 4100 | 1287 | 210 | 25 | 1466 | 1495 | 1.208 | 0.453 | 7.118 |
| Eichendorff 03 | 4342 | 1334 | 182 | 28 | 1482 | 1514 | 1.185 | 0.486 | 5.956 |
| Eichendorff 04 | 1781 | 739 | 79 | 16 | 799 | 816 | 1.133 | 0.403 | 7.916 |
| Eichendorff 05 | 1680 | 699 | 70 | 16 | 750 | 767 | 1.133 | 0.415 | 7.467 |
| Eichendorff 06 | 3223 | 1059 | 130 | 22 | 1163 | 1187 | 1.143 | 0.423 | 7.308 |
| Eichendorff 07 | 2594 | 932 | 121 | 20 | 1031 | 1051 | 1.053 | 0.393 | 7.186 |
| Eichendorff 08 | 3987 | 1320 | 159 | 25 | 1447 | 1477 | 1.250 | 0.475 | 6.922 |
| Eichendorff 09 | 3285 | 1185 | 155 | 22 | 1315 | 1338 | 1.095 | 0.401 | 7.449 |
| Eichendorff 10 | 3052 | 1073 | 131 | 22 | 1178 | 1202 | 1.143 | 0.434 | 6.921 |
| Goethe 01 | 7554 | 2222 | 318 | 33 | 2502 | 2538 | 1.125 | 0.414 | 7.377 |
| Goethe 05 | 559 | 332 | 30 | 8 | 351 | 360 | 1.286 | 0.381 | 11.408 |
| Goethe 09 | 653 | 379 | 30 | 9 | 398 | 407 | 1.125 | 0.352 | 10.203 |
| Goethe 10 | 480 | 301 | 18 | 7 | 310 | 317 | 1.167 | 0.320 | 13.333 |
| Goethe 11 | 468 | 297 | 18 | 7 | 307 | 313 | 1.000 | 0.277 | 13.000 |
| Goethe 12 | 251 | 169 | 14 | 6 | 175 | 181 | 1.200 | 0.379 | 10.040 |
| Goethe 14 | 184 | 129 | 10 | 5 | 133 | 137 | 1.000 | 0.295 | 11.500 |
| Goethe 17 | 225 | 124 | 11 | 6 | 128 | 133 | 1.000 | 0.333 | 9.000 |
| Heine 01 | 19522 | 5769 | 939 | 47 | 6648 | 6706 | 1.275 | 0.415 | 9.430 |
| Heine 02 | 603 | 361 | 50 | 9 | 400 | 409 | 1.171 | 0.358 | 10.720 |
| Heine 03 | 394 | 211 | 21 | 7 | 222 | 230 | 1.302 | 0.393 | 10.944 |
| Heine 04 | 20107 | 5305 | 946 | 47 | 6192 | 6249 | 1.253 | 0.402 | 9.712 |
| Heine 07 | 263 | 169 | 17 | 5 | 179 | 184 | 1.320 | 0.326 | 16.438 |
| Hoffmann 01 | 2974 | 1176 | 95 | 22 | 1247 | 1269 | 1.048 | 0.403 | 6.744 |
| Hoffmann 02 | 1076 | 534 | 29 | 11 | 549 | 561 | 1.200 | 0.366 | 10.760 |
| Hoffmann 03 | 8163 | 2511 | 290 | 34 | 2759 | 2799 | 1.212 | 0.443 | 7.496 |
| Immermann 01 | 28943 | 6397 | 918 | 63 | 7234 | 7313 | 1.274 | 0.464 | 7.529 |
| Kafka 01 | 10256 | 2321 | 448 | 41 | 2717 | 2767 | 1.250 | 0.494 | 6.410 |
| Kafka 02 | 3181 | 1210 | 159 | 23 | 1343 | 1367 | 1.116 | 0.426 | 6.882 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Kafka 03 | 1072 | 513 | 34 | 12 | 532 | 545 | 1.123 | 0.388 | 8.351 |
| Kafka 04 | 625 | 321 | 23 | 10 | 332 | 342 | 1.121 | 0.381 | 8.651 |
| Kafka 05 | 247 | 166 | 14 | 5 | 173 | 178 | 1.333 | 0.339 | 15.438 |
| Kafka 06 | 178 | 137 | 6 | 4 | 138 | 141 | 0.977 | 0.220 | 19.778 |
| Kafka 07 | 132 | 89 | 9 | 4 | 93 | 96 | 1.056 | 0.245 | 18.656 |
| Kafka 08 | 139 | 102 | 9 | 4 | 106 | 109 | 1.288 | 0.273 | 22.240 |
| Kafka 09 | 596 | 343 | 25 | 9 | 358 | 366 | 1.025 | 0.336 | 9.313 |
| Kafka 10 | 86 | 62 | 4 | 4 | 62 | 64 | 0.587 | 0.190 | 9.556 |
| Kafka 11 | 151 | 104 | 9 | 5 | 107 | 111 | 1.106 | 0.315 | 12.327 |
| Kafka 12 | 160 | 101 | 9 | 5 | 104 | 108 | 1.070 | 0.338 | 10.000 |
| Kafka 13 | 232 | 150 | 9 | 6 | 153 | 157 | 0.856 | 0.281 | 9.280 |
| Kafka 14 | 142 | 104 | 11 | 3 | 111 | 113 | 1.055 | 0.177 | 35.500 |
| Kafka 15 | 189 | 136 | 7 | 5 | 138 | 141 | 0.889 | 0.226 | 15.429 |
| Kafka 16 | 255 | 177 | 10 | 6 | 181 | 185 | 0.892 | 0.279 | 10.200 |
| Kafka 17 | 111 | 80 | 11 | 3 | 86 | 89 | 1.425 | 0.271 | 27.750 |
| Kafka 18 | 61 | 48 | 3 | 3 | 48 | 49 | 0.780 | 0.150 | 27.111 |
| Kafka 19 | 41 | 33 | 3 | 2 | 33 | 34 | 1.170 | 0.183 | 41.000 |
| Kafka 20 | 1402 | 539 | 74 | 15 | 596 | 611 | 1.065 | 0.391 | 7.416 |
| Kafka 21 | 610 | 364 | 18 | 10 | 371 | 380 | 1.015 | 0.349 | 8.443 |
| Kafka 22 | 2129 | 887 | 89 | 18 | 956 | 974 | 1.012 | 0.380 | 7.089 |
| Kafka 23 | 255 | 153 | 13 | 6 | 159 | 164 | 1.058 | 0.331 | 10.200 |
| Kafka 24 | 584 | 276 | 25 | 9 | 290 | 299 | 1.204 | 0.374 | 10.382 |
| Kafka 25 | 3414 | 1214 | 104 | 23 | 1290 | 1316 | 1.182 | 0.445 | 7.054 |
| Kafka 26 | 134 | 98 | 7 | 4 | 100 | 103 | 1.040 | 0.225 | 21.440 |
| Kafka 27 | 428 | 240 | 14 | 8 | 246 | 252 | 0.899 | 0.304 | 8.735 |
| Kafka 28 | 470 | 272 | 13 | 8 | 277 | 283 | 0.873 | 0.282 | 9.592 |
| Keller 01 | 25625 | 5516 | 1399 | 59 | 6840 | 6913 | 1.259 | 0.456 | 7.617 |
| Keller 02 | 301 | 196 | 20 | 5 | 209 | 214 | 1.370 | 0.316 | 18.813 |
| Keller 03 | 13149 | 3512 | 724 | 43 | 4181 | 4234 | 1.262 | 0.462 | 7.454 |
| Keller 04 | 1896 | 897 | 103 | 15 | 980 | 998 | 1.273 | 0.409 | 9.673 |
| Lessing 01 | 114 | 78 | 7 | 4 | 80 | 83 | 1.037 | 0.291 | 12.667 |
| Lessing 02 | 208 | 141 | 13 | 4 | 148 | 152 | 1.170 | 0.243 | 23.111 |
| Lessing 03 | 61 | 48 | 4 | 3 | 48 | 50 | 1.173 | 0.225 | 27.111 |
| Lessing 04 | 47 | 41 | 2 | 2 | 40 | 41 | 0.590 | 0.086 | 47.000 |
| Lessing 05 | 182 | 120 | 7 | 5 | 121 | 125 | 1.003 | 0.260 | 14.857 |
| Lessing 06 | 362 | 227 | 13 | 7 | 232 | 238 | 1.035 | 0.326 | 10.056 |
| Lessing 07 | 231 | 161 | 9 | 4 | 165 | 168 | 1.120 | 0.221 | 25.667 |
| Lessing 08 | 74 | 64 | 4 | 2 | 65 | 66 | 1.350 | 0.157 | 74.000 |
| Lessing 09 | 327 | 193 | 24 | 6 | 210 | 215 | 1.050 | 0.290 | 13.080 |
| Lessing 10 | 254 | 154 | 12 | 6 | 159 | 164 | 1.024 | 0.321 | 10.160 |
| Löns 01 | 1672 | 706 | 95 | 15 | 782 | 799 | 1.214 | 0.416 | 8.531 |
| Löns 02 | 2988 | 928 | 141 | 23 | 1042 | 1067 | 1.136 | 0.457 | 6.174 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Löns 03 | 4063 | 1162 | 172 | 26 | 1303 | 1332 | 1.160 | 0.455 | 6.501 |
| Löns 04 | 3713 | 1081 | 167 | 24 | 1218 | 1246 | 1.217 | 0.460 | 7.019 |
| Löns 05 | 4676 | 1235 | 254 | 28 | 1457 | 1487 | 1.111 | 0.439 | 6.414 |
| Löns 06 | 4833 | 1364 | 244 | 29 | 1573 | 1606 | 1.179 | 0.475 | 6.165 |
| Löns 07 | 7743 | 1862 | 414 | 36 | 2232 | 2274 | 1.200 | 0.477 | 6.321 |
| Löns 08 | 6093 | 1724 | 328 | 31 | 2015 | 2050 | 1.167 | 0.448 | 6.770 |
| Löns 09 | 9252 | 2126 | 453 | 39 | 2531 | 2577 | 1.211 | 0.478 | 6.407 |
| Löns 10 | 6546 | 1736 | 274 | 35 | 1968 | 2008 | 1.176 | 0.494 | 5.663 |
| Löns 11 | 4102 | 1294 | 217 | 27 | 1481 | 1509 | 1.077 | 0.437 | 6.068 |
| Löns 12 | 4432 | 1318 | 221 | 26 | 1507 | 1537 | 1.200 | 0.451 | 7.091 |
| Löns 13 | 1361 | 556 | 60 | 14 | 600 | 614 | 1.077 | 0.379 | 8.053 |
| Meyer 01 | 1523 | 801 | 56 | 14 | 840 | 855 | 1.154 | 0.384 | 9.012 |
| Meyer 02 | 573 | 331 | 26 | 8 | 347 | 355 | 1.143 | 0.334 | 11.694 |
| Meyer 03 | 1052 | 551 | 46 | 11 | 583 | 595 | 1.200 | 0.370 | 10.520 |
| Meyer 04 | 2550 | 1142 | 79 | 18 | 1197 | 1219 | 1.294 | 0.436 | 8.824 |
| Meyer 05 | 1249 | 658 | 47 | 12 | 690 | 703 | 1.182 | 0.368 | 10.322 |
| Meyer 06 | 833 | 471 | 34 | 10 | 492 | 503 | 1.222 | 0.381 | 10.284 |
| Meyer 07 | 1229 | 652 | 47 | 13 | 683 | 697 | 1.167 | 0.399 | 8.535 |
| Meyer 08 | 1028 | 556 | 43 | 11 | 585 | 597 | 1.200 | 0.374 | 10.280 |
| Meyer 09 | 776 | 441 | 40 | 9 | 471 | 479 | 1.000 | 0.287 | 12.125 |
| Meyer 10 | 940 | 493 | 41 | 11 | 520 | 532 | 1.200 | 0.391 | 9.400 |
| Meyer 11 | 2398 | 1079 | 88 | 17 | 1146 | 1165 | 1.188 | 0.388 | 9.367 |
| Novalis 01 | 2894 | 1129 | 139 | 21 | 1243 | 1266 | 1.150 | 0.428 | 7.235 |
| Novalis 02 | 3719 | 1487 | 208 | 22 | 1669 | 1693 | 1.143 | 0.394 | 8.433 |
| Novalis 03 | 5321 | 1819 | 233 | 25 | 2018 | 2050 | 1.333 | 0.439 | 9.238 |
| Novalis 04 | 2777 | 1282 | 130 | 18 | 1389 | 1410 | 1.235 | 0.399 | 9.609 |
| Novalis 05 | 8866 | 2769 | 473 | 35 | 3198 | 3240 | 1.235 | 0.446 | 7.670 |
| Novalis 06 | 4030 | 1467 | 178 | 23 | 1617 | 1643 | 1.182 | 0.410 | 8.326 |
| Novalis 07 | 1744 | 792 | 77 | 16 | 851 | 867 | 1.067 | 0.383 | 7.751 |
| Novalis 08 | 2111 | 816 | 75 | 17 | 869 | 889 | 1.250 | 0.435 | 8.246 |
| Novalis 09 | 8945 | 2681 | 442 | 32 | 3082 | 3121 | 1.258 | 0.412 | 9.308 |
| Novalis 10 | 5367 | 1939 | 238 | 26 | 2144 | 2175 | 1.240 | 0.423 | 8.587 |
| Novalis 11 | 1358 | 646 | 83 | 12 | 714 | 727 | 1.235 | 0.357 | 11.950 |
| Novalis 12 | 4430 | 1697 | 195 | 24 | 1861 | 1890 | 1.264 | 0.437 | 8.374 |
| Novalis 13 | 1080 | 514 | 58 | 12 | 557 | 570 | 1.171 | 0.404 | 8.413 |
| Paul 01 | 854 | 487 | 37 | 10 | 512 | 522 | 1.111 | 0.342 | 10.543 |
| Paul 02 | 383 | 255 | 14 | 6 | 260 | 267 | 1.400 | 0.358 | 15.320 |
| Paul 03 | 520 | 311 | 26 | 8 | 326 | 335 | 1.286 | 0.395 | 10.612 |
| Paul 04 | 580 | 354 | 21 | 8 | 365 | 373 | 1.143 | 0.332 | 11.837 |
| Paul 05 | 1331 | 677 | 44 | 12 | 705 | 719 | 1.273 | 0.384 | 11.000 |
| Paul 06 | 526 | 305 | 16 | 8 | 313 | 319 | 0.857 | 0.262 | 10.735 |
| Paul 07 | 508 | 316 | 15 | 7 | 323 | 329 | 1.000 | 0.266 | 14.111 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Paul 08 | 402 | 248 | 22 | 6 | 262 | 268 | 1.200 | 0.299 | 16.080 |
| Paul 09 | 1068 | 547 | 37 | 10 | 570 | 582 | 1.333 | 0.367 | 13.185 |
| Paul 10 | 1558 | 778 | 53 | 13 | 814 | 829 | 1.250 | 0.380 | 10.819 |
| Paul 11 | 2232 | 1027 | 84 | 15 | 1092 | 1109 | 1.214 | 0.360 | 11.388 |
| Paul 12 | 620 | 365 | 25 | 8 | 380 | 388 | 1.143 | 0.321 | 12.653 |
| Paul 13 | 1392 | 652 | 40 | 13 | 676 | 690 | 1.167 | 0.375 | 9.667 |
| Paul 14 | 1400 | 714 | 49 | 14 | 746 | 761 | 1.154 | 0.401 | 8.284 |
| Paul 15 | 1648 | 793 | 65 | 15 | 840 | 856 | 1.143 | 0.394 | 8.408 |
| Paul 16 | 320 | 223 | 12 | 5 | 227 | 233 | 1.500 | 0.335 | 20.000 |
| Paul 17 | 1844 | 897 | 73 | 15 | 952 | 968 | 1.143 | 0.373 | 9.408 |
| Paul 18 | 870 | 489 | 42 | 11 | 520 | 529 | 0.900 | 0.305 | 8.700 |
| Paul 19 | 1236 | 676 | 38 | 13 | 699 | 712 | 1.083 | 0.370 | 8.583 |
| Paul 20 | 2059 | 1011 | 78 | 16 | 1068 | 1087 | 1.267 | 0.419 | 9.151 |
| Paul 21 | 3955 | 1513 | 172 | 24 | 1659 | 1683 | 1.043 | 0.382 | 7.476 |
| Paul 22 | 478 | 302 | 15 | 7 | 309 | 315 | 1.000 | 0.274 | 13.278 |
| Paul 23 | 656 | 386 | 26 | 9 | 401 | 410 | 1.125 | 0.351 | 10.250 |
| Paul 24 | 1465 | 730 | 80 | 13 | 795 | 808 | 1.083 | 0.340 | 10.174 |
| Paul 25 | 588 | 361 | 18 | 8 | 370 | 377 | 1.000 | 0.289 | 12.000 |
| Paul 26 | 1896 | 887 | 61 | 15 | 930 | 946 | 1.143 | 0.367 | 9.673 |
| Paul 27 | 749 | 410 | 26 | 9 | 426 | 434 | 1.000 | 0.292 | 11.703 |
| Paul 28 | 241 | 172 | 8 | 5 | 174 | 178 | 1.000 | 0.258 | 15.063 |
| Paul 29 | 1825 | 872 | 68 | 14 | 921 | 938 | 1.308 | 0.398 | 10.799 |
| Paul 30 | 388 | 238 | 17 | 6 | 248 | 253 | 1.000 | 0.254 | 15.520 |
| Paul 31 | 1630 | 753 | 72 | 14 | 810 | 823 | 1.000 | 0.322 | 9.645 |
| Paul 32 | 163 | 119 | 6 | 4 | 120 | 123 | 1.000 | 0.235 | 18.111 |
| Paul 33 | 596 | 355 | 23 | 8 | 369 | 376 | 1.000 | 0.287 | 12.163 |
| Paul 35 | 1947 | 897 | 82 | 17 | 960 | 977 | 1.063 | 0.385 | 7.605 |
| Paul 36 | 425 | 253 | 15 | 7 | 259 | 266 | 1.167 | 0.340 | 11.806 |
| Paul 37 | 368 | 239 | 12 | 6 | 243 | 249 | 1.200 | 0.313 | 14.720 |
| Paul 38 | 1218 | 636 | 40 | 12 | 660 | 674 | 1.273 | 0.401 | 10.066 |
| Paul 39 | 388 | 248 | 13 | 7 | 253 | 259 | 1.000 | 0.305 | 10.778 |
| Paul 40 | 1370 | 655 | 53 | 14 | 694 | 706 | 0.923 | 0.324 | 8.107 |
| Paul 41 | 1032 | 546 | 43 | 11 | 575 | 587 | 1.200 | 0.374 | 10.320 |
| Paul 42 | 1546 | 731 | 50 | 13 | 764 | 779 | 1.250 | 0.381 | 10.736 |
| Paul 43 | 4148 | 1591 | 152 | 26 | 1714 | 1741 | 1.080 | 0.419 | 6.637 |
| Paul 44 | 1881 | 896 | 66 | 15 | 943 | 960 | 1.214 | 0.392 | 9.597 |
| Paul 45 | 2723 | 1102 | 155 | 18 | 1236 | 1255 | 1.118 | 0.364 | 9.422 |
| Paul 46 | 3095 | 1276 | 99 | 21 | 1351 | 1373 | 1.100 | 0.395 | 7.738 |
| Paul 47 | 516 | 319 | 19 | 8 | 330 | 336 | 0.857 | 0.264 | 10.531 |
| Paul 48 | 1200 | 604 | 50 | 13 | 638 | 652 | 1.167 | 0.404 | 8.333 |
| Paul 49 | 562 | 336 | 19 | 8 | 346 | 353 | 1.000 | 0.295 | 11.469 |
| Paul 50 | 430 | 255 | 23 | 7 | 269 | 276 | 1.167 | 0.338 | 11.944 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Paul 51 | 3222 | 1323 | 116 | 20 | 1413 | 1437 | 1.263 | 0.423 | 8.925 |
| Paul 52 | 1731 | 815 | 71 | 15 | 870 | 884 | 1.000 | 0.336 | 8.832 |
| Paul 53 | 1839 | 864 | 75 | 14 | 922 | 937 | 1.154 | 0.350 | 10.882 |
| Paul 54 | 6644 | 2417 | 245 | 30 | 2625 | 2660 | 1.207 | 0.429 | 7.900 |
| Paul 55 | 7854 | 2680 | 321 | 33 | 2961 | 2999 | 1.188 | 0.429 | 7.670 |
| Paul 56 | 963 | 482 | 47 | 10 | 516 | 527 | 1.222 | 0.354 | 11.889 |
| Pseudonym 01 | 728 | 363 | 30 | 10 | 381 | 391 | 1.111 | 0.371 | 8.988 |
| Pseudonym 02 | 612 | 326 | 23 | 9 | 339 | 347 | 1.000 | 0.323 | 9.563 |
| Raabe 01 | 13045 | 3003 | 691 | 45 | 3638 | 3692 | 1.227 | 0.473 | 6.738 |
| Raabe 02 | 3173 | 962 | 134 | 23 | 1070 | 1094 | 1.091 | 0.426 | 6.556 |
| Raabe 03 | 2690 | 950 | 135 | 21 | 1060 | 1083 | 1.150 | 0.443 | 6.725 |
| Raabe 04 | 6253 | 2110 | 282 | 30 | 2355 | 2390 | 1.207 | 0.443 | 7.435 |
| Raabe 05 | 5087 | 1801 | 196 | 26 | 1964 | 1995 | 1.240 | 0.435 | 8.139 |
| Rieder 01 | 1161 | 510 | 36 | 12 | 532 | 544 | 1.091 | 0.352 | 9.595 |
| Rieder 02 | 1231 | 472 | 55 | 13 | 511 | 525 | 1.167 | 0.399 | 8.549 |
| Rückert 01 | 141 | 97 | 10 | 4 | 102 | 105 | 1.133 | 0.286 | 15.667 |
| Rückert 02 | 327 | 202 | 9 | 7 | 205 | 209 | 0.713 | 0.237 | 9.083 |
| Rückert 03 | 152 | 107 | 8 | 4 | 110 | 113 | 1.097 | 0.267 | 16.889 |
| Rückert 04 | 721 | 412 | 22 | 9 | 423 | 432 | 1.138 | 0.339 | 11.266 |
| Rückert 05 | 212 | 145 | 10 | 5 | 149 | 153 | 0.953 | 0.262 | 13.250 |
| Schnitzler 01 | 2793 | 961 | 109 | 20 | 1044 | 1068 | 1.297 | 0.454 | 8.161 |
| Schnitzler 02 | 1936 | 825 | 59 | 17 | 864 | 882 | 1.105 | 0.402 | 7.563 |
| Schnitzler 03 | 801 | 410 | 28 | 11 | 425 | 436 | 1.057 | 0.373 | 8.010 |
| Schnitzler 04 | 2489 | 870 | 135 | 21 | 982 | 1003 | 1.066 | 0.420 | 6.433 |
| Schnitzler 05 | 2123 | 822 | 110 | 18 | 910 | 930 | 1.215 | 0.439 | 7.640 |
| Schnitzler 06 | 1539 | 668 | 50 | 15 | 701 | 716 | 1.143 | 0.393 | 8.444 |
| Schnitzler 07 | 5652 | 1451 | 259 | 31 | 1673 | 1708 | 1.157 | 0.466 | 6.177 |
| Schnitzler 08 | 1711 | 666 | 63 | 15 | 711 | 727 | 1.210 | 0.398 | 9.224 |
| Schnitzler 09 | 6552 | 1993 | 207 | 32 | 2161 | 2198 | 1.204 | 0.457 | 6.938 |
| Schnitzler 10 | 1349 | 629 | 49 | 15 | 661 | 676 | 1.122 | 0.412 | 7.402 |
| Schnitzler 11 | 1595 | 723 | 97 | 15 | 803 | 818 | 1.086 | 0.381 | 8.138 |
| Schnitzler 12 | 6173 | 1476 | 400 | 31 | 1835 | 1874 | 1.300 | 0.496 | 6.859 |
| Schnitzler 13 | 1184 | 544 | 44 | 13 | 573 | 586 | 1.111 | 0.387 | 8.222 |
| Schnitzler 14 | 3900 | 1309 | 139 | 26 | 1415 | 1446 | 1.265 | 0.496 | 6.497 |
| Sealsfield 01 | 1352 | 600 | 45 | 13 | 629 | 643 | 1.167 | 0.381 | 9.389 |
| Sealsfield 02 | 4663 | 1825 | 142 | 27 | 1936 | 1965 | 1.115 | 0.425 | 6.898 |
| Sealsfield 03 | 3238 | 1197 | 114 | 21 | 1284 | 1309 | 1.250 | 0.439 | 8.095 |
| Sealsfield 04 | 3954 | 1399 | 161 | 24 | 1530 | 1558 | 1.217 | 0.445 | 7.474 |
| Sealsfield 05 | 3187 | 1079 | 96 | 22 | 1149 | 1173 | 1.143 | 0.425 | 7.227 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Sealsfield 06 | 2586 | 1010 | 67 | 20 | 1053 | 1075 | 1.158 | 0.433 | 7.163 |
| Sealsfield 07 | 2939 | 1035 | 75 | 20 | 1086 | 1108 | 1.158 | 0.406 | 8.141 |
| Sealsfield 08 | 4865 | 1333 | 138 | 27 | 1435 | 1469 | 1.308 | 0.487 | 7.197 |
| Sealsfield 09 | 7259 | 2295 | 263 | 31 | 2519 | 2556 | 1.233 | 0.434 | 8.066 |
| Sealsfield 10 | 4838 | 1620 | 138 | 26 | 1726 | 1756 | 1.200 | 0.431 | 7.741 |
| Sealsfield 11 | 3785 | 1265 | 98 | 26 | 1333 | 1361 | 1.120 | 0.455 | 6.056 |
| Sealsfield 12 | 3019 | 1191 | 95 | 20 | 1262 | 1284 | 1.158 | 0.400 | 8.363 |
| Sealsfield 13 | 2370 | 1071 | 89 | 17 | 1139 | 1158 | 1.188 | 0.390 | 9.258 |
| Sealsfield 14 | 2744 | 1198 | 82 | 19 | 1257 | 1278 | 1.167 | 0.401 | 8.469 |
| Sealsfield 15 | 4786 | 1545 | 164 | 27 | 1676 | 1707 | 1.192 | 0.448 | 7.080 |
| Sealsfield 16 | 4497 | 1602 | 137 | 26 | 1707 | 1737 | 1.200 | 0.447 | 7.195 |
| Sealsfield 17 | 6705 | 2273 | 192 | 30 | 2429 | 2463 | 1.172 | 0.415 | 7.973 |
| Sealsfield 18 | 4162 | 1252 | 285 | 24 | 1508 | 1535 | 1.174 | 0.419 | 7.868 |
| Sealsfield 19 | 5626 | 1653 | 171 | 29 | 1789 | 1822 | 1.179 | 0.440 | 7.176 |
| Sealsfield 20 | 8423 | 2735 | 273 | 35 | 2966 | 3006 | 1.176 | 0.436 | 7.286 |
| Sealsfield 21 | 6041 | 2040 | 220 | 29 | 2224 | 2258 | 1.214 | 0.437 | 7.705 |
| Sealsfield 22 | 5748 | 1655 | 157 | 29 | 1776 | 1810 | 1.214 | 0.448 | 7.332 |
| Sealsfield 23 | 1752 | 799 | 80 | 14 | 861 | 877 | 1.231 | 0.382 | 10.367 |
| Sealsfield 24 | 1696 | 753 | 68 | 14 | 803 | 819 | 1.231 | 0.389 | 10.036 |
| Sealsfield 25 | 1368 | 704 | 40 | 12 | 730 | 742 | 1.091 | 0.324 | 11.306 |
| Sealsfield 26 | 1517 | 679 | 44 | 15 | 706 | 721 | 1.071 | 0.385 | 7.740 |
| Sealsfield 27 | 4195 | 1516 | 179 | 24 | 1665 | 1693 | 1.217 | 0.432 | 7.930 |
| Sealsfield 28 | 1515 | 586 | 70 | 15 | 636 | 654 | 1.286 | 0.462 | 7.730 |
| Storm 01 | 38306 | 6233 | 1292 | 76 | 7427 | 7523 | 1.280 | 0.490 | 6.810 |
| Sudermann 01 | 11437 | 2427 | 507 | 43 | 2879 | 2932 | 1.262 | 0.496 | 6.484 |
| Tucholsky 01 | 8544 | 2449 | 351 | 35 | 2757 | 2798 | 1.206 | 0.444 | 7.391 |
| Tucholsky 02 | 7106 | 1935 | 207 | 35 | 2100 | 2140 | 1.176 | 0.475 | 6.147 |
| Tucholsky 03 | 9699 | 2502 | 336 | 38 | 2790 | 2836 | 1.243 | 0.467 | 7.085 |
| Tucholsky 04 | 7415 | 1968 | 214 | 35 | 2139 | 2180 | 1.206 | 0.476 | 6.414 |
| Tucholsky 05 | 4823 | 1399 | 174 | 28 | 1537 | 1571 | 1.259 | 0.490 | 6.616 |
| Wedekind 01 | 4035 | 1336 | 122 | 26 | 1428 | 1456 | 1.120 | 0.441 | 6.456 |
| Wedekind 02 | 6040 | 1731 | 179 | 31 | 1872 | 1908 | 1.200 | 0.463 | 6.711 |
| Wedekind 03 | 7402 | 1934 | 276 | 34 | 2168 | 2208 | 1.212 | 0.465 | 6.797 |
| Wedekind 04 | 1297 | 646 | 44 | 13 | 676 | 688 | 1.000 | 0.333 | 9.007 |
| Wedekind 05 | 1935 | 580 | 89 | 19 | 645 | 667 | 1.208 | 0.494 | 5.972 |
| Wedekind 06 | 5955 | 1689 | 249 | 34 | 1901 | 1936 | 1.052 | 0.450 | 5.468 |
| Wedekind 07 | 605 | 341 | 22 | 9 | 352 | 361 | 1.101 | 0.358 | 9.453 |
| Wedekind 08 | 2033 | 855 | 87 | 17 | 921 | 940 | 1.197 | 0.425 | 7.941 |

The variability and independence of $p$ and $q$ in terms of the text size $N$ for the data of Table 2 are shown in Figure 2.

It should be remarked that, coincidentally or not, the sum of $p$ and $q$ appears to converge towards the golden number 1.618… with increasing $N$. More specifically, for the data of Table 1 the mean of $(p + q) = 1.615$ and the standard deviation is 0.183, whereas for the data of Table 2 the mean of $(p + q) = 1.526$ and the standard deviation of 0.176.

Since the variance of $L$ is known (cf. Popescu, Mačutek, Altmann 2009), tests for differences between $p$ or $q$ values are possible.

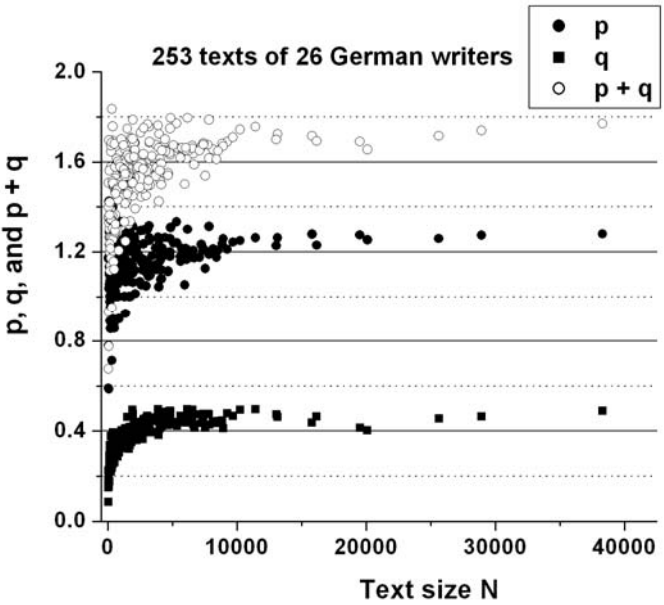



Figure 2. Indicators $p$, $q$, and their sum in terms of the text size $N$ for the data of Table 2

If we consider the means of $p$ we see that this indicator is not very adequate for typological purposes because the differences between languages are not very evident. However, if we use $q$ or $b$, we obtain a very clear picture. The smaller $q$ or the greater $b$, the more synthetic is the given language. The results of averaging are presented in Table 3 where the languages are ordered according to $\overline{q}$.

Finally, Table 4 represents an attempt at ranking 26 German writers by mean $\overline{q}$. Since the differences are conspicuous, the indicator can

be used even within one language to estimate the "morphological" way of writing of an author, i.e. his use of more or less synthetic word forms. According to this scale, the most "synthetic" German authors have a mean $\overline{q}$ of about 0.25 and the most "analytic" ones of about 0.5. A more thorough investigation could, perhaps, illuminate the development of German.

Table 3

The mean indicators $\overline{p}$, $\overline{q}$, and $\overline{b}$ (ranked by increasing $\overline{q}$)

| Language | $\overline{p}$ | $\overline{q}$ | $\overline{b}$ |
|---|---|---|---|
| Kannada | 1.230 | 0.273 | 21.035 |
| Latin | 1.190 | 0.278 | 19.831 |
| Hungarian | 1.395 | 0.281 | 26.095 |
| Indonesian | 1.141 | 0.312 | 13.834 |
| Marathi | 1.181 | 0.324 | 13.593 |
| German | 1.111 | 0.334 | 11.212 |
| Czech | 1.189 | 0.336 | 13.017 |
| Romanian | 1.170 | 0.356 | 10.935 |
| Bulgarian | 1.117 | 0.370 | 9.152 |
| Slovenian | 1.215 | 0.376 | 10.683 |
| Russian | 1.289 | 0.382 | 11.549 |
| Italian | 1.269 | 0.412 | 9.524 |
| English | 1.263 | 0.435 | 8.573 |
| Tagalog | 1.272 | 0.446 | 8.279 |
| Lakota | 1.196 | 0.449 | 7.216 |
| Maori | 1.191 | 0.479 | 6.220 |
| Marquesan | 1.272 | 0.504 | 6.618 |
| Rarotongan | 1.260 | 0.520 | 5.889 |
| Hawaiian | 1.317 | 0.542 | 5.906 |
| Samoan | 1.297 | 0.565 | 5.324 |

Table 4

The mean indicators $\overline{q}$ and $\overline{b}$ for 26 German writers (ranked by increasing $\overline{q}$)

| mid life year | Author | $\overline{q}$ | stdev $q$ | $\overline{b}$ | stdev $b$ | number of texts |
|---|---|---|---|---|---|---|
| 1755 | Lessing | 0.242 | 0.075 | 25.771 | 20.353 | 10 |
| 1827 | Rückert | 0.278 | 0.038 | 13.231 | 3.177 | 5 |
| 1904 | Kafka | 0.307 | 0.087 | 14.436 | 9.109 | 28 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1791 | Goethe | 0.344 | 0.046 | 10.733 | 1.999 | 8 |
| 2001 | pseudonym | 0.347 | 0.033 | 9.275 | 0.406 | 2 |
| 1794 | Paul | 0.347 | 0.051 | 10.950 | 2.717 | 56 |
| 1862 | Meyer | 0.374 | 0.038 | 10.033 | 1.148 | 11 |
| 2001 | Rieder | 0.376 | 0.033 | 9.072 | 0.740 | 2 |
| 1827 | Heine | 0.379 | 0.037 | 11.449 | 2.862 | 5 |
| 1823 | Droste | 0.389 | 0.038 | 9.641 | 1.704 | 6 |
| 1799 | Hoffmann | 0.404 | 0.038 | 8.333 | 2.135 | 3 |
| 1810 | Chamisso | 0.407 | 0.036 | 8.363 | 0.845 | 11 |
| 1855 | Keller | 0.411 | 0.068 | 10.889 | 5.378 | 4 |
| 1787 | Novalis | 0.413 | 0.025 | 8.703 | 1.182 | 13 |
| 1829 | Sealsfield | 0.422 | 0.033 | 8.009 | 1.142 | 28 |
| 1897 | Schnitzler | 0.427 | 0.041 | 7.550 | 0.882 | 14 |
| 1891 | Wedekind | 0.429 | 0.055 | 7.226 | 1.432 | 8 |
| 1806 | Arnim | 0.432 | 0.043 | 7.557 | 0.842 | 3 |
| 1823 | Eichendorff | 0.435 | 0.034 | 7.194 | 0.539 | 10 |
| 1870 | Busch | 0.437 | - | 8.556 | - | 1 |
| 1871 | Raabe | 0.444 | 0.018 | 7.119 | 0.663 | 5 |
| 1890 | Löns | 0.451 | 0.030 | 6.706 | 0.808 | 13 |
| 1818 | Immermann | 0.464 | - | 7.529 | - | 1 |
| 1913 | Tucholsky | 0.470 | 0.017 | 6.731 | 0.504 | 5 |
| 1853 | Storm | 0.490 | - | 6.810 | - | 1 |
| 1893 | Sudermann | 0.496 | - | 6.484 | - | 1 |

## *REFERENCES*

Hirsch, J.E. (2005): An index to quantify an individuals scientific research output. [http://arxiv.org/PS_caxhe/physics/pdf/0508/0508025.pdf and http://en.wikipedia.org/wiki/Hirsch_number].

Laufer, J.; Nemcová, E. (2009): Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 13-25.*

Nemcová, E.; Popescu, I.-I.; Altmann, G. (2009): Word associations in French. (submitted: Festschrift für R. Grotjahn).

Popescu, I.-I. et al. (2009): *Word frequency studies.* Berlin-New York: Mouton de Gruyter.

Popescu, I.-I.; Altmann, G. (2008): On the regularity of diversification in language. *Glottometrics 17, 94-108.*

Popescu, I.-I.; Mačutek, J.; Altmann, G. (2009): *Aspects of word frequencies*. (submitted).

## Appendix

Authors and text titles used in Table 2

| ID | mid life year | Text |
|---|---|---|
| Arnim 01 | 1806 | Der tolle Invalide auf dem Fort Ratonneau |
| Arnim 02 | 1806 | Des ersten Bergmanns ewige Jugend |
| Arnim 03 | 1806 | Frau von Saverne |
| Busch 01 | 1870 | Eduards Traum |
| Chamisso 01 | 1810 | Peter Schlemihls wundersame Geschichte I |
| Chamisso 02 | 1810 | Peter Schlemihls wundersame Geschichte II |
| Chamisso 03 | 1810 | Peter Schlemihls wundersame Geschichte III |
| Chamisso 04 | 1810 | Peter Schlemihls wundersame Geschichte IV |
| Chamisso 05 | 1810 | Peter Schlemihls wundersame Geschichte V |
| Chamisso 06 | 1810 | Peter Schlemihls wundersame Geschichte VI |
| Chamisso 07 | 1810 | Peter Schlemihls wundersame Geschichte VII |
| Chamisso 08 | 1810 | Peter Schlemihls wundersame Geschichte VIII |
| Chamisso 09 | 1810 | Peter Schlemihls wundersame Geschichte IX |
| Chamisso 10 | 1810 | Peter Schlemihls wundersame Geschichte X |
| Chamisso 11 | 1810 | Peter Schlemihls wundersame Geschichte XI |
| Droste 01 | 1823 | Die Judenbuche |
| Droste 02 | 1823 | Der Tod des Erzbischofs Engelbert |
| Droste 03 | 1823 | Das Fegefeuer |
| Droste 04 | 1823 | Der Fundator |
| Droste 05 | 1823 | Die Schwestern |
| Droste 08 | 1823 | Der Geierpfiff |
| Eichendorff 01 | 1823 | Aus dem Leben eines Taugenichts 1 |
| Eichendorff 02 | 1823 | Aus dem Leben eines Taugenichts 2 |
| Eichendorff 03 | 1823 | Aus dem Leben eines Taugenichts 3 |
| Eichendorff 04 | 1823 | Aus dem Leben eines Taugenichts 4 |
| Eichendorff 05 | 1823 | Aus dem Leben eines Taugenichts 5 |
| Eichendorff 06 | 1823 | Aus dem Leben eines Taugenichts 6 |
| Eichendorff 07 | 1823 | Aus dem Leben eines Taugenichts 7 |
| Eichendorff 08 | 1823 | Aus dem Leben eines Taugenichts 8 |
| Eichendorff 09 | 1823 | Aus dem Leben eines Taugenichts 9 |
| Eichendorff 10 | 1823 | Aus dem Leben eines Taugenichts 10 |

| | | |
|---|---|---|
| Goethe 01 | 1791 | Die neue Melusine |
| Goethe 05 | 1791 | Der Gott und die Bajadere |
| Goethe 09 | 1791 | Elegie 19 |
| Goethe 10 | 1791 | Elegie 13 |
| Goethe 11 | 1791 | Elegie 15 |
| Goethe 12 | 1791 | Elegie 2 |
| Goethe 14 | 1791 | Elegie 5 |
| Goethe 17 | 1791 | Der Erlkönig |
| Heine 01 | 1827 | Die Harzreise |
| Heine 02 | 1827 | Die Heimkehr - Götterdämmerung |
| Heine 03 | 1827 | Die Heimkehr - Die Wallfahrt nach Kevlaar |
| Heine 04 | 1827 | Ideen. Das Buch Le Grand |
| Heine 07 | 1827 | Belsazar |
| Hoffmann 01 | 1799 | Der Sandmann - Nathanael an Lothar eliminated |
| Hoffmann 02 | 1799 | Der Sandmann - Clara an Nathanael eliminated |
| Hoffmann 03 | 1799 | Der Sandmann - Nathanael an Lothar eliminated |
| Immermann 01 | 1818 | Der Karneval und die Somnambule |
| Kafka 01 | 1904 | In der Strafkolonie |
| Kafka 02 | 1904 | Ein Bericht für eine Akademie |
| Kafka 03 | 1904 | Betrachtung - Kinder auf der Landstraße |
| Kafka 04 | 1904 | Betrachtung - Entlarvung eines Bauernfängers |
| Kafka 05 | 1904 | Betrachtung - Der plötzliche Spaziergang |
| Kafka 06 | 1904 | Betrachtung - Entschlüsse |
| Kafka 07 | 1904 | Betrachtung - Der Ausflug ins Gebirge |
| Kafka 08 | 1904 | Betrachtung - Das Unglück des Junggesellen |
| Kafka 09 | 1904 | Betrachtung - Der Kaufmann |
| Kafka 10 | 1904 | Betrachtung - Zerstreutes Hinausschaun |
| Kafka 11 | 1904 | Betrachtung - Der Nachhauseweg |
| Kafka 12 | 1904 | Betrachtung - Die Vorüberlaufenden |
| Kafka 13 | 1904 | Betrachtung - Der Fahrgast |
| Kafka 14 | 1904 | Betrachtung - Kleider |
| Kafka 15 | 1904 | Betrachtung - Die Abweisung |
| Kafka 16 | 1904 | Betrachtung - Zum Nachdenken für Herrenreiter |
| Kafka 17 | 1904 | Betrachtung - Das Gassenfenster |
| Kafka 18 | 1904 | Betrachtung - Wunsch, Indianer zu werden |
| Kafka 19 | 1904 | Betrachtung - Die Bäume |
| Kafka 20 | 1904 | Betrachtung - Unglücklichsein |
| Kafka 21 | 1904 | Ein Brudermord |
| Kafka 22 | 1904 | Ein Landarzt |
| Kafka 23 | 1904 | Der Geier |
| Kafka 24 | 1904 | Vor dem Gesetz |
| Kafka 25 | 1904 | Ein Hungerkünstler |

| | | |
|---|---|---|
| Kafka 26 | 1904 | Nachts |
| Kafka 27 | 1904 | Das Schweigen der Sirenen |
| Kafka 28 | 1904 | Die Sorge des Hausvaters |
| Keller 01 | 1855 | Romeo und Julia auf dem Dorfe |
| Keller 02 | 1855 | Vom Fichtenbaum |
| Keller 03 | 1855 | Spiegel, das Kätzchen |
| Keller 04 | 1855 | Das Tanzlegendchen |
| Lessing 01 | 1755 | Der Besitzer des Bogens |
| Lessing 02 | 1755 | Die Erscheinung |
| Lessing 03 | 1755 | Der Esel mit dem Löwen |
| Lessing 04 | 1755 | Der Fuchs |
| Lessing 05 | 1755 | Die Furien |
| Lessing 06 | 1755 | Jupiter und das Schaf |
| Lessing 07 | 1755 | Der Knabe und die Schlange |
| Lessing 08 | 1755 | Minerva |
| Lessing 09 | 1755 | Der Rangstreit der Tiere |
| Lessing 10 | 1755 | Zeus und das Pferd |
| Löns 01 | 1890 | Der Werwolf - 1. Die Haidbauern |
| Löns 02 | 1890 | Der Werwolf - 2. Die Mansfelder |
| Löns 03 | 1890 | Der Werwolf - 3. Die Braunschweiger |
| Löns 04 | 1890 | Der Werwolf - 4. Die Weimaraner |
| Löns 05 | 1890 | Der Werwolf - 5. Die Marodebruede |
| Löns 06 | 1890 | Der Werwolf - 6. Die Bruchbauern |
| Löns 07 | 1890 | Der Werwolf - 7. Die Wehrwoelfe |
| Löns 08 | 1890 | Der Werwolf - 8. Die Schnitter |
| Löns 09 | 1890 | Der Werwolf - 9. Die Kirchenleute |
| Löns 10 | 1890 | Der Werwolf - 10. Die Hochzeiter |
| Löns 11 | 1890 | Der Werwolf - 11. Die Kaiserlichen |
| Löns 12 | 1890 | Der Werwolf - 12. Die Schweden |
| Löns 13 | 1890 | Der Werwolf - 13. Die Haidbauern |
| Meyer 01 | 1862 | Der Schuss von der Kanzel 1 |
| Meyer 02 | 1862 | Der Schuss von der Kanzel 2 |
| Meyer 03 | 1862 | Der Schuss von der Kanzel 3 |
| Meyer 04 | 1862 | Der Schuss von der Kanzel 4 |
| Meyer 05 | 1862 | Der Schuss von der Kanzel 5 |
| Meyer 06 | 1862 | Der Schuss von der Kanzel 6 |
| Meyer 07 | 1862 | Der Schuss von der Kanzel 7 |
| Meyer 08 | 1862 | Der Schuss von der Kanzel 8 |
| Meyer 09 | 1862 | Der Schuss von der Kanzel 9 |
| Meyer 10 | 1862 | Der Schuss von der Kanzel 10 |
| Meyer 11 | 1862 | Der Schuss von der Kanzel 11 |

| | | |
|---|---|---|
| Novalis 01 | 1787 | Heinrich von Ofterdingen - Die Erwartung 1 |
| Novalis 02 | 1787 | Heinrich von Ofterdingen - Die Erwartung 2 |
| Novalis 03 | 1787 | Heinrich von Ofterdingen - Die Erwartung 3 |
| Novalis 04 | 1787 | Heinrich von Ofterdingen - Die Erwartung 4 |
| Novalis 05 | 1787 | Heinrich von Ofterdingen - Die Erwartung 5 |
| Novalis 06 | 1787 | Heinrich von Ofterdingen - Die Erwartung 6 |
| Novalis 07 | 1787 | Heinrich von Ofterdingen - Die Erwartung 7 |
| Novalis 08 | 1787 | Heinrich von Ofterdingen - Die Erwartung 8 |
| Novalis 09 | 1787 | Heinrich von Ofterdingen - Die Erwartung 9 |
| Novalis 10 | 1787 | Heinrich von Ofterdingen - Die Erfuellung |
| Novalis 11 | 1787 | Hyazinth und Rosenblütchen |
| Novalis 12 | 1787 | Neue Fragmente - Sophie |
| Novalis 13 | 1787 | Neue Fragmente - Traktat vom Licht |
| Paul 01 | 1794 | Dr. Katzenbergers Badereise 1. |
| Paul 02 | 1794 | Dr. Katzenbergers Badereise 2. Reisezwecke |
| Paul 03 | 1794 | Dr. Katzenbergers Badereise 3. Ein Reisegefaehrte |
| Paul 04 | 1794 | Dr. Katzenbergers Badereise 4. Bona |
| Paul 05 | 1794 | Dr. Katzenbergers Badereise 5. Herr von Niess |
| Paul 06 | 1794 | Dr. Katzenbergers Badereise 6. Fortsetzung der Abreise |
| | | Dr. Katzenbergers Badereise 7. Fortgesetzte |
| Paul 07 | 1794 | Fortsetzung der Abreise |
| Paul 08 | 1794 | Dr. Katzenbergers Badereise 8. Beschluss der Abreise |
| | | Dr. Katzenbergers Badereise 9. Halbtagfahrt nach |
| Paul 09 | 1794 | St. Wolfgang |
| Paul 10 | 1794 | Dr. Katzenbergers Badereise 10. Mittags-Abenteuer |
| Paul 11 | 1794 | Dr. Katzenbergers Badereise 11. Wagen-Sieste |
| Paul 12 | 1794 | Dr. Katzenbergers Badereise 12. die Avantuere |
| Paul 13 | 1794 | Dr. Katzenbergers Badereise 13. Theodas ersten Tages Buch |
| Paul 14 | 1794 | Dr. Katzenbergers Badereise 14. Missgeburten-Adel |
| Paul 15 | 1794 | Dr. Katzenbergers Badereise 15. Hasenkrieg |
| Paul 16 | 1794 | Dr. Katzenbergers Badereise 16. Ankunft-Sitzung |
| Paul 17 | 1794 | Dr. Katzenbergers Badereise I. Huldigungpredigt |
| | | Dr. Katzenbergers Badereise II. Ueber Hebels |
| Paul 18 | 1794 | alemannische Gedichte |
| | | Dr. Katzenbergers Badereise III. Rat zu urdeutschen |
| Paul 19 | 1794 | Taufnamen |
| Paul 20 | 1794 | Dr. Katzenbergers Badereise IIII. Dr. Fenks Leichenrede |
| | | Dr. Katzenbergers Badereise V. Ueber den Tod |
| Paul 21 | 1794 | nach dem Tode |
| Paul 22 | 1794 | Dr. Katzenbergers Badereise 17. Blosse Station |
| Paul 23 | 1794 | Dr. Katzenbergers Badereise 18. Maennikes Seegefecht |
| Paul 24 | 1794 | Dr. Katzenbergers Badereise 19. Mondbelustigungen |
| Paul 25 | 1794 | Dr. Katzenbergers Badereise 20. Zweiten Tages Buch |

| | | |
|---|---|---|
| | | Dr. Katzenbergers Badereise 21. Hemmrad der |
| Paul 26 | 1794 | Ankunft im Badeorte |
| Paul 27 | 1794 | Dr. Katzenbergers Badereise 22. Niessiana |
| Paul 28 | 1794 | Dr. Katzenbergers Badereise 23. Ein Brief |
| Paul 29 | 1794 | Dr. Katzenbergers Badereise 24. Mittagtischreden |
| | | Dr. Katzenbergers Badereise 25. Musikalisches |
| Paul 30 | 1794 | Deklamatorium |
| Paul 31 | 1794 | Dr. Katzenbergers Badereise 26. Neuer Gastrollenspieler |
| Paul 32 | 1794 | Dr. Katzenbergers Badereise 27. Nachtrag |
| Paul 33 | 1794 | Dr. Katzenbergers Badereise 28. Darum |
| Paul 35 | 1794 | Dr. Katzenbergers Badereise 30. Tischgebet und Suppe |
| | | Dr. Katzenbergers Badereise 31. Aufdeckung und |
| Paul 36 | 1794 | Sternbedeckung |
| Paul 37 | 1794 | Dr. Katzenbergers Badereise 32. Erkennszene |
| | | Dr. Katzenbergers Badereise 33. Abendtisch-Reden |
| Paul 38 | 1794 | über Schauspiele |
| Paul 39 | 1794 | Dr. Katzenbergers Badereise 34. Brunnen-Beaengstigungen |
| Paul 40 | 1794 | Dr. Katzenbergers Badereise 35. Theodas Brief an Bona |
| Paul 41 | 1794 | Dr. Katzenbergers Badereise 36. Herzens-Interim |
| | | Dr. Katzenbergers Badereise 37. Neue Mitarbeiter |
| Paul 42 | 1794 | an allem |
| Paul 43 | 1794 | Dr. Katzenbergers Badereise I. Die Kunst, einzuschlafen |
| Paul 44 | 1794 | Dr. Katzenbergers Badereise II. Das Glueck |
| Paul 45 | 1794 | Dr. Katzenbergers Badereise III. Die Vernichtung |
| Paul 46 | 1794 | Dr. Katzenbergers Badereise 38. Wie Katzenberger … |
| Paul 47 | 1794 | Dr. Katzenbergers Badereise 39. Doktors Hoehlen-Besuch |
| Paul 48 | 1794 | Dr. Katzenbergers Badereise 40. Theodas Hoehlen-Besuch |
| Paul 49 | 1794 | Dr. Katzenbergers Badereise 41. Drei Abreisen |
| | | Dr. Katzenbergers Badereise 42. Theodas kuerzeste |
| Paul 50 | 1794 | Nacht der Reise |
| Paul 51 | 1794 | Dr. Katzenbergers Badereise 43. Praeliminar-Frieden ... |
| Paul 52 | 1794 | Dr. Katzenbergers Badereise 44. Die Stuben-Treffen |
| | | Dr. Katzenbergers Badereise 45. Ende der Reisen |
| Paul 53 | 1794 | und Noeten |
| | | Dr. Katzenbergers Badereise I. Wuensche fuer |
| Paul 54 | 1794 | Luthers Denkmal |
| Paul 55 | 1794 | Dr. Katzenbergers Badereise II. Ueber Charlotte Corday |
| Paul 56 | 1794 | Dr. Katzenbergers Badereise III. Polymeter |
| Pseudonym 01 | 2001 | Eine kleine Geschichte mit der Zeit |
| Pseudonym 02 | 2001 | Taumelnde Realitaet |
| Raabe 01 | 1871 | Im Siegeskranze |
| Raabe 02 | 1871 | Eine Silvester-Stimmung |
| Raabe 03 | 1871 | Ein Besuch |

| | | |
|---|---|---|
| Raabe 04 | 1871 | Deutscher Mondschein |
| Raabe 05 | 1871 | Theklas Erbschaft |
| Rieder 01 | 2001 | Liebe Mutter |
| Rieder 02 | 2001 | Brief an einen Toten |
| Rückert 01 | 1827 | Barbarossa |
| Rückert 02 | 1827 | Amor ein Besenbinder |
| Rückert 03 | 1827 | Der Frost |
| Rückert 04 | 1827 | Die goldne Hochzeit |
| Rückert 05 | 1827 | Erscheinung der Schnitterengel |
| Schnitzler 01 | 1897 | Der Sohn |
| Schnitzler 02 | 1897 | Albine |
| Schnitzler 03 | 1897 | Amerika |
| Schnitzler 04 | 1897 | Der Andere |
| Schnitzler 05 | 1897 | Die Braut |
| Schnitzler 06 | 1897 | Erbschaft |
| Schnitzler 07 | 1897 | Die Frau des Weisen |
| Schnitzler 08 | 1897 | Der Fürst ist im Hause |
| Schnitzler 09 | 1897 | Das Schicksal |
| Schnitzler 10 | 1897 | Welch eine Melodie |
| Schnitzler 11 | 1897 | Frühlingsnacht im Seziersaal |
| Schnitzler 12 | 1897 | Die Toten schweigen |
| Schnitzler 13 | 1897 | Er wartet auf den vazierenden Gott |
| Schnitzler 14 | 1897 | Mein Freund Ypsilon |
| Sealsfield 01 | 1829 | Das Cajuetenbuch - Die Praerie am Jacinto |
| Sealsfield 02 | 1829 | Das Cajuetenbuch 1 |
| Sealsfield 03 | 1829 | Das Cajuetenbuch 2 |
| Sealsfield 04 | 1829 | Das Cajuetenbuch 3 |
| Sealsfield 05 | 1829 | Das Cajuetenbuch 4 |
| Sealsfield 06 | 1829 | Das Cajuetenbuch 5 |
| Sealsfield 07 | 1829 | Das Cajuetenbuch 6 |
| Sealsfield 08 | 1829 | Das Cajuetenbuch 7 |
| Sealsfield 09 | 1829 | Das Cajuetenbuch 8 |
| Sealsfield 10 | 1829 | Das Cajuetenbuch 9 |
| Sealsfield 11 | 1829 | Das Cajuetenbuch 10 |
| Sealsfield 12 | 1829 | Das Cajuetenbuch 11 |
| Sealsfield 13 | 1829 | Das Cajuetenbuch 12 |
| Sealsfield 14 | 1829 | Das Cajuetenbuch 13 |
| Sealsfield 15 | 1829 | Das Cajuetenbuch 14 |
| Sealsfield 16 | 1829 | Das Cajuetenbuch 15 |
| Sealsfield 17 | 1829 | Das Cajuetenbuch 16 |
| Sealsfield 18 | 1829 | Das Cajuetenbuch - Der Fluch Kishogues |
| Sealsfield 19 | 1829 | Das Cajuetenbuch - Der Kapitaen |

| | | |
|---|---|---|
| Sealsfield 20 | 1829 | Das Cajuetenbuch - Callao 1825 |
| Sealsfield 21 | 1829 | Das Cajuetenbuch - Havanna 1816 |
| Sealsfield 22 | 1829 | Das Cajuetenbuch - Sehr Seltsam! |
| Sealsfield 23 | 1829 | Das Cajuetenbuch - Ein Morgen im Paradiese |
| Sealsfield 24 | 1829 | Das Cajuetenbuch - Selige Stunden |
| Sealsfield 25 | 1829 | Das Cajuetenbuch - Das Diner |
| Sealsfield 26 | 1829 | Das Cajuetenbuch - Der Abend |
| Sealsfield 27 | 1829 | Das Cajuetenbuch - Die Fahrt und die Kajuete |
| Sealsfield 28 | 1829 | Das Cajuetenbuch - Das Paradies der Liebe |
| Storm 01 | 1853 | Der Schimmelreiter |
| Sudermann 01 | 1893 | Die Reise nach Tilsit |
| Tucholsky 01 | 1913 | Schloss Gripsholm 1 |
| Tucholsky 02 | 1913 | Schloss Gripsholm 2 |
| Tucholsky 03 | 1913 | Schloss Gripsholm 3 |
| Tucholsky 04 | 1913 | Schloss Gripsholm 4 |
| Tucholsky 05 | 1913 | Schloss Gripsholm 5 |
| Wedekind 01 | 1891 | Mine-Haha I |
| Wedekind 02 | 1891 | Mine-Haha II |
| Wedekind 03 | 1891 | Mine-Haha III |
| Wedekind 04 | 1891 | Mine-Haha IV |
| Wedekind 05 | 1891 | Rabbi Esra |
| Wedekind 06 | 1891 | Frühlingsstürme |
| Wedekind 07 | 1891 | Silvester |
| Wedekind 08 | 1891 | Der Verführer |

# Фоносемантический анализ текста (на материале русских и английских звуко-цветовых ассоциаций)

*Лариса Прокофьева (Саратов, Россия)*

Фоносемантика как наука вступила в такую стадию своего развития, когда созданная теория с полноправным аппаратом и очерченной проблематикой проверяется расширением материала и аспектов исследования, получают экспериментальное подтверждение высказанные предположения, формируются системные классификации, создаются специальные словари (С. В. Воронин, Ю. А. Казарин, М. Магнус, И. Ю. Павловская, С. С. Шляхова, наши исследования и др.). При этом в особом положении оказалась прикладная часть теории – фоносемантический анализ текста, который стал наиболее разработанным разделом задолго до оформления самостоятельного направления в лингвистической науке. Прежде всего, это связано с тем, что первым объектом анализа стал текст художественный, причем поэтический, где взаимосвязь означаемого и означающего проявляется наиболее ярко и отчетливо. Хотя исследования звуковой структуры текста долгое время ограничивались описанием различных видов звуковых повторов, лишь изредка выходя за рамки собственно фоники, тем не менее, фоностилистические методы, выработанные за многие годы, активно используются при современном интерпретационном анализе текста, как поэтического, так и прозаического, являясь непременной его составляющей.[1]

Исследование взаимосвязи звука и цвета в художественном тексте предусматривает, прежде всего, изучение психического феномена, индивидуального опыта ассоциирования, как авторского, так и читательского, поэтому субъективный подход в таких наблюдениях является основным и научно обоснованным: «Изучение художественного текста с позиций психолингвистики делает возможным не только осветить темную силу человеческого сознания, но

[1] Ю.А. Казарин так определяет цель ФС анализа: «достижение за счет интерпретации фонетических смыслов углубленного и расширенного восприятия, понимания и усвоения смысловой системы и структуры стихотворения в целом» (Казарин, 2000: 134).

и высветить лучом знания всю радугу первичного единства одним из проявлений кого является комплекс звука и цвета» (Шуришина 1999: 52). Анализ текста на основе его звуко-цветовой ассоциативности, сопоставление, совмещение и, в конечном итоге, объединение лингвистического, психологического, физиологического, искусствоведческого и др. подходов способны продемонстрировать постепенное углубление в скрытую семантику художественного текста, что может открыть новые горизонты как в общей теории текста, так и в частных методиках его анализа.

Элементарный и исключительно массовый пример представляют разработанные российскими исследователями формализованные методы изучения текста, воплощенные в прикладных компьютерных программах анализа текста на уровне фоносемантики ВААЛ, DIATON, PSYLINE CD, и мн. др. Сразу отметим, что почти все они имеют коммерческую (полную) и демонстрационную (сокращенную) версию, и в силу определенных причин мы можем пользоваться только последней. Тем не менее, основные принципы, лежащие в основе каждой из программ, – одни и те же – принцип семантического дифференциала, разработанный Ч. Осгудом и примененный на практике А. П. Журавлевым. Соответственно, основной метод при наличии разнообразных дополнительных методик тоже один и тот же: подсчет и выявление значимых отклонений от средней частотности в речи (по данным Ленинградской фонетической школы), соотнесение их с матрицей оценок звукобукв русского языка по 25 бинарным шкалам (количество их может варьироваться), вычисление на этой основе градуированного набора максимально значимых для данного текста признаков.

Безусловно, идея «витает в воздухе», поэтому данная методика в упрощенном виде (без сведений по цветовой окраске согласных звукобукв русского языка) присутствует и в составе ВААЛа и ДИАТОНа. Нами обнаружена и ссылка и безымянную программу украинских специалистов по рекламе, направленную на анализ звукоцветовых соотношений, этимологическую и социолингвистическую экспертизу названия торговой марки. Автор Лученко заявляет, что провел опрос 5 000 информантов, составил матрицы фонетического значения звукобукв украинского языка, но демонстрирует только данные по цвето-звуковой ассоциативности гласных, причем абсолютно совпадающие с данными А. П. Жу-равлева.

Анализ существующих программ, способных анализировать звуко-цветовую ассоциативность (ЗЦА) в тексте, показал, что все

они, обладая достоинствами, имеют и некоторые недостатки, не позволяющие использовать при работе с текстом:

- большинство из них основывается лишь на данных по цветовой символике гласных русского языка, не учитывая согласные;
- нигде не подчеркивается национальная составляющая явления – даже украинская программа не показывает отличия ЗЦА русского языка от украинского;
- отсутствует возможность исследования текстов разных видов, типов и жанров;
- максимально нивелирована визуальная составляющая текста – даются лишь изображения самых частотных цветов по всему тексту, динамика в расчет не принимается и не показывается;
- в большинстве случаев отсутствует сопутствующая информация о наличии приемов семантизации звуковой стороны текста.

Таким образом, существование большого количества разнообразных программ, автоматизирующих процесс анализа звукоцветовой составляющей текста, не отменяет насущную необходимость универсального инструмента, способного учитывать разнообразные факторы, возникающие в процессе работы. Именно поэтому нами была создана универсальная русско-английская программа ЗВУКОЦВЕТ (программист Т. В. Миронова, руководитель проекта канд. физ.-мат. наук И. Л. Пластун), в которой учтены высказанные выше замечания.

Программа «Звукоцвет» написана на языке программирования C++ в интегрированной среде разработки C++ Builder 6. Код программы разбит на 4 модуля: модуль интерфейса, модуль для работы с английскими текстами, модуль для работы с русскими текстами и общий модуль для анализа текста, предполагающий работу с любым письменным текстом на русском или английском языках.

Программный продукт разработан для выполнения следующих задач:

1. Расчет частотности звукобукв русского или английского текста для прозаического, поэтического и драматургического текста с учетом их структуры и вывод результатов в таблицу.
2. Определение цветности текста на основании рассчитанной частотности звукобукв и представление результатов в виде графиков и диаграмм.
3. Определение наличия приемов аллитерации и ассонанса в тексте с помощью сравнения рассчитанной и среднестатистической частотностей звукобукв.

Для выполнения поставленных задач данная программа реализует следующие функции:

- загрузка текста из файлов типа *.rtf или *.txt;
- расстановка ударений с помощью «горячих клавиш» и кнопок на панели окна и вручную;
- автоматическое деление текста типа «поэзия» на строфы заданного размера;
- подсчет количества звукобукв в тексте;
- расчет частотности звукобукв текста с учетом типа текста;
- определение наличия приемов семантизации;
- вывод результатов в таблицу;
- определение цветности текста;
- вывод результатов в виде графиков и диаграмм.

Программа предназначена для универсального использования, поэтому предусмотрены три подмодуля для анализа прозаического (нехудожественного и художественного) текста, стихотворного и драматического произведения любого объема. Алгоритм работы с каждым родом и видом предусматривает некоторую структурную специфику, которая состоит, например, в делении на строки и строфы (в поэзии), если автором это деление было произведено. В прозе анализу целого текста предшествует анализ каждого абзаца, чем достигается эффект динамики ЗЦА от завязки через кульминацию к развязке. В драматургии предусмотрен анализ не только текста в целом, но и определение звуко-цветового сопровождения каждого персонажа, номинаций действующих лиц и авторских ремарок. Таким образом, обеспечивается «послойное» исследование цветовой фоносемантики текста с возможностью выбора для интерпретатора наиболее значимой информации из набора возможных.

Предполагается также возможность работы в сети Интернет для получения информации о словах из сетевых орфоэпических словарей и автоматической расстановки ударения в текстах большого объема. В поэтическом тексте или ритмизованной прозе размер и ритм может особым образом регулировать ударения, поэтому предусмотрена ручная установка этого параметра.

Очевидно, что задача исследования смыслового восприятия художественного текста в коммуникативном аспекте «перекрывает» задачу такого же рода в отношении текста нехудожественного. В понятие *художественно-эстетический компонент* входит и фоно-

семантическая составляющая, которая требует специального углубленного анализа. На наш взгляд, принципиально возможен анализ *любого* текста, но ценность такого исследования будет различной в зависимости от цели и задач данного речевого произведения. По нашей гипотезе национально обусловленные особенности ЗЦА будут значимо проявляться именно в нехудожественных текстах, тогда как в художественных они будут подвергаться существенной коррекции со стороны идиостиля, программирующего, в том числе и индивидуальную звуко-цветовую картину мира, отражаемую в тексте.

Для проверки гипотезы проанализировано по 100 нехудожественных текстов разных стилей и жанров на русском и английском языках (научные статьи, фрагменты учебников гуманитарного и естественнонаучного профиля, публицистические газетные и журнальные статьи, документе общего и специального характера, фрагменты текста законов, словарные дефиниции, статьи из энциклопедий) для выявления возможных различий. Безусловно, выбранные для анализа группы текстов не исчерпывают богатство типов нехудожественного текста, более того, внутри каждого из них можно выделить подгруппы, в которых могут обнаружиться значимые отличия. Но в данном случае важно было получить однозначный ответ на принципиальный вопрос, может ли ЗЦА в нехудожественном тексте соотноситься с общим настроем текста (эмоциональным или рациональным), является ли подобный анализ значимым для такого типа текста. Гипотетически в текстах по умолчанию «рационалистического» характера, реализующих только информативную функцию, не предусматривающих эмоциональный компонент, сложно ожидать яркости на фоносемантическом уровне. Тем не менее, в некоторых разновидностях нехудожественного текста, например, в публицистической статье (сближающейся стилистически с художественным), нельзя исключить вероятность проявления личностного характера отбора фонетических средств для достижения поставленной цели воздействия на адресата, но в общей массе анализируемых текстов они нивелируются, уступая место общеязыковой частотности и средней информативности фоно-семантического уровня.

В процессе отбора материала для исследования было проанализировано около 200 различных статей, из них только около 10 интуитивно нами отмечены как имеющие приемы семантизации звуковой сферы. Почти все они принадлежали профессиональным писателям или литературным работникам (В. Астафьев, В. Солоухин,

В. Песков, С. Кургинян и др.), т.е. только по теме являлись публицистическими, по форме и способу выражения мысли – художественными. В любом случае, общий отрицательный результат также значим, т.к. во-первых, позволяет подтвердить наличие обобщенного национально мотивированного фона текстов на русском (РЯ) и английском языках (АЯ), во-вторых, максимально сосредоточить внимание на эстетически значимых речевых произведениях. Результаты в системном виде представлены в Таблице 1.

Анализ результатов исследования позволяет утвердиться во мнении, что в нехудожественном тексте программируется общая национальная система ЗЦА, находящаяся в латентном состоянии в процессе его восприятия, тем самым находит еще одно подтверждение гипотеза о различиях в звуко-цветовой картине мира наций. В большинстве случаев и в английских, и в русских произведениях разных стилей и жанров рационалистического характера находит отражение желто-зеленая (англ.) и черно-белая (рус.) составляющие, тогда как в текстах, предусматривающих эмоциональный компонент (газетная и журнальная публицистика), реализуются красная (англ.) и сине-красная (рус.) составляющие. Отметим, что в некоторых текстах были выделены статистически значимые отклонения от средней частотности в данном языке, но для того чтобы интерпретировать результаты, необходимо оценить, можно ли на основании превышения частотности говорить о наличии некоего приема. Наши эксперименты по восприятию информантами квазилексических звукобуквенных комплексов с заранее рассчитанной цветовой ассоциативностью показали, что максимального соответствия прогнозируемые и реальные данные достигают при соотнесении со списками рангов встречаемости (информативности) звукобукв (Прокофьева 1998). Сводные данные по рангам встречаемости в русском (Белоногов, Фролов 1963) и английском (Singh 1997) языках приведены в Таблице 2.

На основании результатов с квазилексическими комплексами приходим к выводу, что информативность – еще один существенный факт, который надо учитывать при интерпретации фактов изменения частотности звукобукв в тексте: выделенные в текстах научного стиля превышения нормы *и, о, п., т* в русском и *e, t, w* в английском не оказывают существенного воздействия на его цветовой фон и, соответственно, на реципиента, т.к. являются высокочастотными и их ранг встречаемости в сознании человека не выходит за пределы латентного.

Таблица 1

Сводные результаты исследования нехудожественных текстов на русском и английском языках

| **Тип текста** | **Наличие/отсутствие отклонений от средней частотности звукобукв в тексте +/-** | | **Ведучий цвет текста** | | **Наличие/отсутствие отдельных значимих фрагментов в тексте +/-** | |
|---|---|---|---|---|---|---|
| | **РЯ** | **АЯ** | **РЯ** | **АЯ** | **РЯ** | **АЯ** |
| Научная статья | - чуть повышена частотность *п, и, т* | - | черно-белый | желто-зеленый | - | - |
| Учебник гуманитарного профиля | - чуть повышена частотность *о*) | - | черно-желтый | зелено-желтый | - | - |
| Учебник естественно-научного профиля | + частотность *и, п* превышена в 2 раза | - чуть повышена частотность w | черно-синий | зелено-синий | - | - |
| Газетная статья | + частотность *р, м* превышена в 1,4 раза | + в 1,3 раза превышена частотность *t, е* | черно-бело-красный | красно-желтый | + *красный* | + *красный* |
| Журнальная статья | + частотность *и, н* превышена в 2 раза | + в 1,5 раза превышена частотность *t* | черно-бело-синий | синий | + *синий* | + *красный* |
| Общий документ | - | - | черно-белый | желто-зеленый | - | - |
| Специальный документ | - | - | черно-белый | желто-зеленый | - | - |
| Текст закона | - | - | черно-белый | желто-зелено-красный | - | - |
| Словарная дефиниция | - чуть повышена частотность *п* | - чуть повышена частотность *s* | черно-белый | зеленый | - | - |
| Статья из энциклопедии | - | - | черно-белый | желто-зеленый | - | - |

Таблица 2
Сводные данные по рангам встречаемости звукобукв
в русском и английском языках

| **Звуко-буква РЯ**[2] | **Ранг** | **Звуко-буква РЯ** | **Ранг** | **Звуко-буква АЯ** | **Ранг** | **Звуко-буква АЯ** | **Ранг** |
|---|---|---|---|---|---|---|---|
| О | 1 | У | 16 | E | 1 | M | 14-15 |
| Е | 2 | З | 17 | T | 2 | W | 14-15 |
| А | 3 | Ы | 18 | A | 3 | F | 16 |
| И | 4 | Б | 19 | O | 4 | Y | 17-18 |
| Н | 5 | Ь | 20 | I | 5 | G | 17-18 |
| Т | 6 | Г | 21 | N | 6 | P | 19 |
| Р | 7 | Й | 22 | S | 7 | B | 20 |
| В | 8 | Х | 23 | H | 8 | V | 21 |
| С | 9 | Ч | 24 | R | 9 | K | 22 |
| Л | 10 | Ж | 25 | D | 10 | X | 23-24 |
| Д | 11 | Ш | 26 | L | 11 | J | 23-24 |
| М | 12 | Ю | 27 | C | 12-13 | Q | 25-26 |
| П | 13 | Щ | 28 | U | 12-13 | Z | 25-26 |
| К | 14 | Ц | 29 | | | | |
| Я | 15 | Э | 30 | | | | |

Качественно иная картина наблюдается в текстах публицистического стиля (газетные и журнальные статьи), в которых программа отметила существенное повышение информативности звукобукв в отдельных абзацах. Это отразилось и в общей оценке текста – к нейтральным ахроматическим цветам приблизились красный и синий. Надо отметить, что повышение частотности звукобукв, несущих в себе данную ассоциативную цветовую информацию, наблюдается в кульминационных в смысловом отношении абзацах, где автор старается максимально привлечь внимание читателя к обсуждаемой проблеме, поэтому дополнительно использует приемы семантизации фоники текста. Отклонения от обычного общеязыкового распределения звукобукв в тексте могут возникнуть и случайно, вследствие ограниченности набора букв и конечности их допускаемых языком сочетаний, причем в этом случае они, по большей мере, лишены смысла и функций. Опрос нескольких практикующих

[2] В таблице, составленной Н. Н. Соколовым, ранг Ё отдельно не выделяется, а учитывается вместе с Е. Это, вероятно, связано с преобладающей в последние десятилетия тенденцией к игнорированию буквы Ё в письменной речи.

журналистов о том, насколько сознательно они используют аллитерации и ассонансы в своей работе, показал, что обычно они не задумываются об этом, а получается «само собой».

Интересный материал, подтверждающий национальную обусловленность ЗЦА находим в работах Л.Н. Санжарова, который провел небольшой эксперимент с носителями украинского языка как родного (г. Измаил, Украина) по выявлению цветовых ассоциаций гласных звуков, а затем целостной оценке стихотворения Л.Украинки. Результаты показали, что только оценка [и] различается с русской (*желто-бело-голубой*, а не *синий*), тогда как автоматический и аудиторский анализ текста практически совпал.

Таким образом, наши эксперименты в общем и целом подтвердили правильность гипотезы, согласно которой в нехудожественном тексте кодируется национально обусловленная система звуко-цветовой ассоциативности, остающаяся при восприятии в подсознании в скрытом состоянии, активизируясь только в случаях ненамеренного (или намеренного) повышения информативности звукобукв, например, в публицистическом стиле. Соответственно, исследователь или работник PR, СМИ, политик и публичный деятель, специалист рекламного дела, обладающий методикой анализа ЗЦА, может прогнозировать появление той или иной эмоциональной реакции и даже направлять ее, организуя звуковые повторы различного рода в текстах любого типа.

Информация о звукобукве по умолчанию кодируется в любом тексте безотносительно к его содержанию. При этом отмечено, что даже в нехудожественном тексте национальная матрица ЗЦА проявлялась с большими или меньшими отклонениями от полученных «средних» данных. Выделенные закономерности позволили продолжить исследование с текстами художественными, подключив сведения из психологии восприятия. Понятно, что абсолютно однозначные результаты в ходе эксперимента по цветовому восприятию художественного текста в принципе невозможны, но все же обобщение множества полученных в ходе аудиторского и компьютерного экспериментов результатов способно выявить статистические закономерности звуко-цветового ассоциирования с достаточно высокой степенью вероятности. Совпадение или несовпадение реального восприятия и рассчитанного результата позволит говорить о различных типах проявления ЗЦА в художественном тексте.

Априори можно предположить, что любой текст может быть подвергнут анализу, но лишь по апробированной устойчивой методике, которая должна содержать универсальную основу и, тем не

менее, иметь возможности трансформации в зависимости от рода и вида текста, подвергающегося анализу. Методика исследования ЗЦА поэтического текста прошла длительную проверку (А. П. Журавлев, Е. Г. Сомова, Л. Н. Санжаров, Т. И. Шуришина, С. С. Шляхова, М. А. Балаш, Е. Н. Клеменова, С. В. Бондарь, и мн.др.[3]), тогда как прозаический текст в данном аспекте исследовался эпизодически и фрагментарно (Е. Н. Клеменова, Л. Н. Санжаров), контуры работы с драматургией только намечены. Отметим, что во всех вышеуказанных экспериментах практической основой послужили данные по цветовому ассоциированию гласных звукобукв русского языка. Скорее всего, в текстах с ассонансами без приемов семантизации согласных (что само по себе довольно редко) этого было бы достаточно, но в стихотворениях с намеренными аллитерациями, подчеркивающими определенную авторскую идею, данная методика не позволяет получить такие же однозначные результаты.

Безусловно, форма в поэзии предельно обнажена, она непосредственным образом связана с содержанием – этот факт не вызывает сомнений у исследователей стиха. Существует даже зафиксированная в Princeton Encyclopedia of Poetry and Poetics (1965) словарная статья, в которой утверждается, что «общий смысл поэтического произведения частично является функцией его звуковой организации». Осмысление звуков в поэтической речи (в отличие от функции «посредников» в обыденной речи) происходит закономерно и непроизвольно, поэзия «значения языковых знаков превращает в смыслы или окутывает их смыслами; осмысляет звучание их» (Вейдле 2002: 136), поэтому даже не выходя за пределы языка, она остается особым образованием, где разрозненные элементы картины мира сливаются в единое ценностное полотно. Подтверждение нашим мыслям находим у У. Вейнрейха: «звуковая сторона знака приобретает независимую символическую значимость («импрессионистическую» – звукоподражательную или «экспрессионистическую», т.е. синестетическую); особое семантическое отношение приписывается знакам, имеющим сходное означающее; короче говоря, зачаточные соотношения между содержанием и

[3] Методика достаточно проста: в стихотворениях соотносились присутствующие там цветовые номинации и результаты подсчета цветовой наполненности гласных ЗБ. Выявлены значительные совпадения, это дало возможность говорить о «неосознанной» инструментовке поэта, которая полностью совпадала с лексической и семантической наполненностью стихотворений.

выражением активно эксплуатируются, тогда как в «семантически нормальных» использованиях языка (т.е. в обыденной речи) эти соотношения произвольны» (Вейнрейх 1970: 169). Звукосмысловая и шире – звукосимволическая связь в поэзии образует «особое качество, позволяющее звуковым образам слов заменять зрительные представления» (Невзглядова 1968: 23), повторяющийся звук, выделенный в звучании, а тем самым и в сознании человека, вступает в ассоциативную связь с семантикой тех слов, которые его содержат. Кроме того, благодаря общему звуку в словах совершается взаимопроникновение смысла. Данный механизм находит свое непосредственное отражение в звуко-цветовой ассоциативности, к которой в процессе восприятия присоединятся эмоциональная реакция, обусловленная среди прочего и фоновым значением цвета.

По мнению К. Ф. Тарановского, известного специалиста по русской поэтике, в художественном тексте связь звуковых повторов со смыслом синестетического характера в большей степени осуществляется посредством ударных гласных: «контраст «темный – светлый» достигается противопоставлением гласных с самыми высокими первыми формантами (компактные) и с самыми низкими первыми формантами (диффузные)[4]» (Тарановский 2000: 349), тогда как согласные передают разнообразие оттенков. Наши наблюдения позволяют уточнить это положение – если в поэзии решающую роль попеременно играют то аллитерации, то ассонансы, то в прозе на первый план выступают аллитерации, в драматургии их значение менее выражено и во многом зависит от формы текста (поэтической или прозаической).

Таким образом, апробированную выше методику необходимо трансформировать в соответствии с задачами анализа художественного текста: изменению должна подвергнуться автоматизированная часть исследования ЗЦА, на остальных этапах наблюдается лишь необходимость приспособление способа выявления реакций реципиентов художественного текста к его объему и структуре. «Впечатляющее воздействие звуковой фактуры проявляется в поэтическом языке в двух направлениях: в выборе и в группировке фонем и их составляющих; эти два выразительных фактора, навевающих образы, хотя и скрыты, но присутствуют и в нашем обычном речевом поведении» (Якобсон 1983: 107) – эти слова Р. Якобсона

[4] Противопоставление темноты и светлости обычно приписывается оппозициям *у:и* и *о:е* (Тарановский 2000: 353).

из знаковой работы «В поисках сущности языка» укрепляют нас в мысли, что прозаические тексты в звуко-цветовом отношении могут в своей основе соотноситься с обычным речевым поведением, т.е. их комплексная оценка не будет выходить за пределы национально мотивированной. Ю. М. Лотман отмечал, что особой концентрации связи на фонетическом уровне достигают там, «где обычные языковые связи неполны или не мотивированы. И наоборот – на участках текста, где морфо-синтаксическая упорядоченность ясна, фонологическая ослаблена» (Лотман 1972:64). Художественный текст – сложная, упорядоченная авторской идеей структура, в которой фонетическая составляющая может «перейти» с уровня эксплицитной выраженности на уровень имплицитной, что не снижает значимости, но требует специальных усилий для ее выявления.

Разработанная нами целостная методика выявления глубинных, мотивированных языковым сознанием звуко-цветовых ассоциаций, а также закономерностей, обусловленных индивидуальным стилем творца речи включает наряду с фоностилистическими методами, активно использующимися при современном интерпретационном анализе текста, приемы простейших психолингвистических экспериментов, оценивающих реальность восприятия приемов реципиентами. Эксперименты показывают, что существуют индивидуальные реакции, которые в силу принадлежности творчески активным личностям, способным оказывать воздействие на окружающих (непосредственно или опосредованно), могут быть восприняты или творчески переработаны на уровне подсознания (реже – сознания) реципиентом. К таковым относятся синестетические модели музыкантов, художников, писателей и поэтов, выразивших свои индивидуальные ЗЦА. Авторские установки должны учитываться при изучении принципов, которыми руководствовался художник слова в непосредственном творчестве, так как могут прояснить особенности его работы над текстом, своеобразие его миропонимания и мироощущения.

О синестезии часто говорят как о компоненте восприятия произведений искусства. Например, предполагается, что живописное полотно может сопровождаться какими-то незрительными ощущениями, музыка предназначена пробуждать зрительные реакции, язык в стихотворениях способен вызывать живые чувственные впечатления и т.д. На наш взгляд, явление, наблюдаемое в художественной среде, менее перцептивно по своему качеству и опосредовано в большей степени метафорическим пониманием или

же накопленными ассоциациями, поэтому в чистом виде Сз не является, но попадает под определение Ст. Индивидуальные модели ЗЦА, не выходящие на уровень феномена Сз, способны активизироваться и частично «всплывать в светлое поле сознания» лишь под влиянием особых факторов-стимулов, во многом индивидуальных, причем, далеко не всегда эстетических. Тем не менее, возникающие в моменты восприятия произведений искусства индивидуальные реакции Ст характера оказывают дополнительное воздействие на личность, вызывая эмоциональное переживание, сопровождающее переживание эстетическое. Они важны для самого субъекта восприятия, но не оказывают такого же влияния на окружающих и в высокой степени автономны. Но существуют индивидуальные реакции, которые в силу принадлежности творчески активным личностям, способным оказывать воздействие на окружающих (непосредственно или опосредованно), могут быть восприняты или творчески переработаны на уровне подсознания (реже – сознания) реципиентом. К таким, по нашему мнению, относятся синестетические модели музыкантов, художников, писателей и поэтов, выразивших свои индивидуальные ЗЦА.

Представление об авторском понимании значимости синестетизма для искусства вообще и символистской поэзии в частности дают теоретические работы А. Блока. Например, в статье «Краски и слова» он находит глубинную изначальную родственность между живописью краской и живописью словом: «Говорят, слов больше, чем красок; но, может быть, достаточно для изящного писателя, для поэта – только таких слов, которые соответствуют краскам» (Блок 1982: 8), и далее: «Душа писателя поневоле заждалась среди абстракций, загрустила в лаборатории слов. Тем временем перед слепым взором ее бесконечно преломлялась цветовая радуга. И разве не выход для писателя – понимание зрительных впечатлений, уменье смотреть? Действие света и цвета освободительно. Оно улегчает душу, рождает прекрасную мысль» (Указ. соч.: 10). В этой небольшой статье А. Блок утверждает важное положение, которое становится исходной точкой для понимания идиостилевого своеобразия ЗЦА: обладая уникальным даром, поэт, тем не менее, должен учиться у природы ее внутренним закономерностям, ее краскам и звукам, пытаться отразить их, воплотить в слове: «Только часто прикасаясь взором к природе, отдаваясь свободно зримому и яркому простору, можно стряхивать с себя гнет боязни слов, расплывчатой и неуверенной мысли» (Указ. соч.: 11). Развивая мысль, предположим,

что индивидуальность может вступить в противоречие с природой, преобразуя ее в соответствии со своими эстетическими задачами, но результат такого преобразования в итоге может быть понятен самому творцу, но остаться недоступным для читателя. Проверить это предположение можно экспериментальным путем, если иметь в виду наличие/отсутствие авторских установок на ЗЦА.

В теоретических работах А.Блока находим замечания по поводу принципиальной необходимости соединения разных способов восприятия мира, их слиянии для создания произведения искусства. Так, в докладе «О современном состоянии русского символизма» он говорит о внутренней системе поэтического мира в ее развитии: «миры, предстающие взору в свете лучезарного меча, становятся все более зовущими; уже из глубины их несутся щемящие музыкальные звуки, призывы, шепоты, почти слова. Вместе с тем, они начинают окрашиваться (здесь возникает первое глубокое знание о цветах)» (Указ. соч.: 114). Налицо яркая синестетическая ассоциативность, возводимая в ранг обязательного приема. Принцип соответствия звука и цвета выражается Блоком эксплицитно, при этом он справедливо полагает, что способы его выражения – проявление воли и индивидуальности поэта.

Тенденция к слиянию звука и цвета ярко проявилась в творчестве А. Блока. Многочисленные исследователи его стиля часто ссылаются на факт, что тропеические средства, с помощью которых реализуется принцип звуковых и незвуковых соответствий, не только многочисленны, разнообразны, но и эволюционируют от цикла к циклу. Р. З. Миллер-Будницкая отмечала, что высший подъем кривой синестетизма приходится на цикл «Снежная маска», затем количество подобных тропов уменьшается (Миллер-Будницкая, 1930):

О, *запах пламенный духов!*
О, шелестящий миг!
О, речи магов и волхвов!
Пергамент желтых книг!

«Лазурью бледной месяц плыл…» (1906)

При этом отмечается, что среди разнообразных видов синестезий в поэзии А. Блока самое большое место занимают звукоцветовые аналогии:

Рукавом в окно мне машет,
Рукавом в окно мне машет,

*Красным криком* зажжена,
Так и манит, так и пляшет,
И ласкает скакуна.

«Прискакала дикой степью...» (1905)

Значительное количество синестетических тропов достаточно ярко, но все же опосредованным образом свидетельствует только о яркости ЗЦА, а не о синестезических способностях поэта (само их наличие является одним из признаков аксиологического аспекта отражения действительности). Непосредственное же указание на то, что поэт не отвергает наличие феномена у других, можно найти в его дневниковых записях. 20 января 1913 года Блок пишет о беседе с композитором Меттнером. Разговор коснулся явлений синестетизма в музыке, в частности, светового рояля А. Н. Скрябина. Сам факт «цветового слуха» не вызывает у него сомнений, но поэт отмечает индивидуальный характер проявления феномена у разных композиторов: «Красное do... для Меттнера – белое. Зато mi у всех (и у Скрябина, и у Римского-Корсакова, и у Меттнера) – голубое».

Никаких письменно зафиксированных авторских систем звуко-цветовых соответствий найти в работах Блока не удалось. Скорее всего, этой системы для него не существовало, а в непосредственном поэтическом творчестве его вела обостренная интуиция. Работы А. П. Журавлева, в которых анализируются стихотворения поэта, подтверждают этот вывод – именно творчество А. Блока стало прекрасной иллюстрацией проявлений творческого сверхсознания[5], которое опережает работу рационального мышления. Декодирование сигналов сверхсознания может служить источником авторских откровений для читателя.

Серия наших экспериментов по звуко-цветовому ассоциированию поэтических произведений А. Блока, проведенная в период с 1990 по 2003 гг. с использованием компьютерной программы KNEW (программисты канд. техн. наук Н. М. Брянцев и А. В. Демидов) и аудиторских опросов, продемонстрировала полные/регулярные совпадения цвето-эмоциональной и фоносемантической оценок текста (Прокофьева 1992, 2004). Вопрос о том, насколько сознательно

[5] «Вдохновение поэта – это активизация деятельности его подсознания и интуиции, которые в свою очередь активизируют работу сознания, образуя интеллектуальный сплав, который можно назвать сверхсознанием. Оно глубоко проникает в сущность явлений, обладает мощной воздействующей и прогностической силой» (Журавлев 2004: 14).

использует автор звуко-ассоциативные связи и насколько способен интерпретатор текста раскрыть дополнительную информацию сенсорно-эмотивного характера, касается психологии творчества и восприятия эстетической информации. Но в том случае, когда поэтическая мотивация как ассоциативное ощущение образных представлений входит в творческое задание автора (его поэтический идиостиль) и вызывает адекватные интеллектуальные и эмоциональные реакции читателя, можно говорить о системности ЗЦА средств. В этом плане не совсем корректно было бы говорить о программируемости А. Блоком подобных соответствий – такое «задание» исходит не из творческой воли автора, а из объективно существующего в природе человека явления Ст. Результаты наших экспериментов по восприятию текстов А. Блока демонстрируют большую степень вероятности фиксации данного явления информантами, при этом использование приемов семантизации звука помогает читателю глубже проникнуться «музыкой слова», включив тем самым в процесс понимания подсознательные импульсы, продуцирующие (среди прочих) ЗЦА.

Продемонстрируем комплексную методику анализа, состоящую из компьютерных расчетов, аудиторского эксперимента и психолингвистической интерпретации значений цвета, на примере стихотворения А. Блока «Та жизнь прошла…» 1914 года.

Автоматизированный анализ текста с помощью компьютерной программы представил следующий вариант его прогнозируемой цвето-звуковой наполненности:

*Та жизнь прошла*, КРАСНЫЙ
*И сердце спит*, ЗЕЛЕНЫЙ
*Утомленно.* БЕЛО-СИНИЙ
*И ночь опять пришла*, БЕЛО-ЖЕЛТЫЙ
*Бесстрашная - глядит* КРАСНЫЙ
*В мое окно*. БЕЛО-ЖЕЛТЫЙ
ЦВЕТ СТРОФЫ - БЕЛО-ЖЕЛТЫЙ ИНФ.: КРАСНЫЙ
*И выпал* снег, КРАСНЫЙ
*И не прогнать* БЕЛО-СИНИЙ
*Мне зимних чар...* СИНИЙ
*И не вернуть тех нег*, ЗЕЛЕНЫЙ
*И странно вспоминать*, БЕЛО-СИНИЙ
*Что был пожар*. БЕЛО-ЖЕЛТЫЙ
ЦВЕТ СТРОФЫ - БЕЛО-ЖЕЛТЫЙ ИНФ.: ЧЕРНО-БЕЛЫЙ
ЦВЕТ ТЕКСТА - БЕЛО-СИНИЙ ИНФ.: БЕЛО-СИНИЙ

Поясним выводимые результаты: цвета демонстрируют последовательную статистику частотности звукобукв в строке и строфе, помета ИНФ. показывает изменение цветовой картины на основе данных об информативности того или иной заукобуквы[6]. Ниже представлены статистические результаты компьютерного анализа данного текста (Таблица 3).

Сводные данные демонстрируют, что прием ассонанса в «чистом виде» не встречается, тогда как налицо значимые отклонения от средней частотности согласных *ш, ц, п, н, ж, г.* Из них максимальные индексы встречаемости у *п* и *н* – эти звукобуквы встретились в тексте с частотностью в 2 раза превышающей среднюю. Таким образом, прогнозируемая цветовая оценка текста на основании значимых звукоповторов – *черно-синяя*.

В аудиторском приняли участие ученики 10-11 классов гимназии № 1 г. Саратова (45 человек)[7]. В тексте присутствуют только имплицитные цветовые номинации *ночь, снег, пожар,* т.е. прогнозируемые оценки информантов под воздействием лексической семантики – *черный, белый, красный.* Результаты сведены в Таблицу 4.

Полученные результаты подтверждают мысль, что лексическая семантика, безусловно, оказывает решающее воздействие на восприятие реципиентов: 4 строка с цветовой номинацией *ночь* оценена как *черная и синяя*, 7 строка (*сн*ег) – как *белая*, 12 строка (*пожар*) – как *красная*. Заметим, что цветовые номинации корректируют и оценку строф, при этом можно было бы ожидать, что ассоциативно яркое слов *пожар* окажет влияние и на оценку всего текста, однако это не так – текст оценен как *синий*, что вполне соотносится с фоносемантическим расчетом. Вспомним, что наличие яркого приема аллитерации было отмечено у звукобукв *н* и *п*, причем выделить их определенную локализацию довольно трудно: звуки «рассыпаны» по всему стихотворению, несколько «сгущаясь» во 2 строфе. *Черно-синяя* оценка встречается в ответах информантов довольно часто, т.е. звукосимволическая система поэтического текста «работает». Информанты, как нам кажется,

[6] Например, информативность У, безусловно, выше информативности О, а информативность Ж выше Л, так как частотность О и Л выше, чем Ж и У.

[7] Эксперимент проводился по традиционной методике: диктор читал отдельно каждую строку, затем целиком строфу, затем все стихотворение; аудиторы записывали спонтанную цветовую реакцию на каждую РИЕ. Общение информантов друг с другом исключалось

вполне адекватно воспринимают фоносемантическую информацию в тех случаях, когда она не «перекрывается» лексической, оценивая строки без цветовых «подсказок» в соответствии с звукосимволическими возможностями звукобукв. При этом, конечно, речь не идет о непременном 100 % совпадении: всегда есть и статистически не вполне достоверное количество информантов, и индивидуальные ассоциации, и скрытые имплицитные цветовые номинации, способные актуализироваться в определенном контексте (например, в 9 строке слово зимних могло бы проявить «белую» оценку при усилении семантического поля слова «зима») и под. Важно отметить, что творческая интуиция А.Блока гармонично отбирает «нужные» для цветовой (и, соответственно, психологической) реакции звукобукв, организуя единство поэтического текста и его читателя.

Таблица 3

Результаты компьютерного анализа стихотворения А. Блока

| Звукобуква | Частотность в тексте | Статистика в речи | Прием аллитерация/ассонанс | Звукобуква | Частотность в тексте | Статистика в речи | Прием аллитерация / ассонанс |
|---|---|---|---|---|---|---|---|
| А | 7.86 | 9.5 | | П | 5.71 | 2.6 | * |
| Б | 0.71 | 1.8 | | Р | 6.43 | 3.8 | |
| В | 2.14 | 3.9 | | С | 5.00 | 4.9 | |
| Г | 2.86 | 1.5 | * | Т | 7.86 | 7.5 | |
| Д | 1.43 | 3.7 | | У | 0.71 | 2.9 | |
| Е | 8.57 | 8.9 | | Ф | 0.00 | 0.3 | |
| Ж | 1.43 | 0.8 | * | Х | 1.43 | 0.9 | |
| З | 1.43 | 1.5 | | Ц | 0,71 | 0,4 | * |
| И | 5.00 | 5.6 | | Ч | 1,43 | 2,0 | |
| Й | 0.00 | 1.3 | | Ш | 2,14 | 1,2 | * |
| К | 0.71 | 3.3 | | Щ | 0,00 | 0,3 | |
| Л | 4.29 | 3.7 | | Ы | 1,43 | 1,6 | |
| М | 2.86 | 3.2 | | Э | 0,00 | 0,5 | |
| Н | 12.14 | 6.4 | * | Ю | 0,00 | 0,6 | |
| О | 9.29 | 10.4 | | Я | 2,14 | 2,4 | |

Если «перевести» фоносемантическую цветовую информацию на язык психологии восприятия (Серов 1990), то можно представить следующий психологический рисунок текста:

1, 5, 7 строки – Сила, гнев

2, 10 строки – покой, уравновешенность
3, 8, 11 строки – слабость, холод, пустая бесконечность
4, 6, 12 строки + 1 и 2 строфа – интуиция, вера
9 строка – сдержанность, ум
весь текст – слабость, холод, пустая бесконечность.

Таблица 4
Результаты аудиторского эксперимента со стихотворением А. Блока

| **Структурная единица текста** | **Аудиторская оценка** | **Автоматизированная оценка** |
|---|---|---|
| 1 строка | черно-белый | красный |
| 2 строка | синий и красный | зеленый |
| *3 строка* | *синий* | *бело-синий* |
| 4 строка | черный и синий | бело-желтый |
| 5 строка | белый | красный |
| *6 строка* | *бело-синий* | *бело-желтый* |
| 1 строфа | синий | бело-желтый |
| 7 строка | белый | красный |
| *8 строка* | *синий* | *бело-синий* |
| **9 строка** | **синий** | **синий** |
| 10 строка | красный | зеленый |
| 11 строка | черный | бело-синий |
| 12 строка | красный | бело-желтый |
| 2 строфа | красный | бело-желтый |
| *Весь текст* | *синий* | *бело-синий* |

- полное совпадение с национальной матрицей русского языка (жирный шрифт),
- частичное совпадение (курсив), - отсутствие совпадений´

Анализ данного стихотворения с помощью демонстрационной (ограниченной) версии программы ВААЛ (Нейролингвистическая экспертная система) подтвердил наши «прогнозы», основанные только на цветовой информации, заложенной в звукобуквах, и дал следующие оценки: плохой; маленький; страшный; грубый, затем нежный; слабый; тусклый. Заметим, что приведенные выше психологические характеристики цветовой информации могут служить в определенном смысле «ключевыми словами» для интерпретации анализируемого стихотворения: гармония текста определяет восприятие читателем определенной фоносемантической информации,

которая создает ощущение бессилия и бесстрастности в душе лирического героя, оттого что ничего нельзя изменить, от воспоминаний о былой силе. Но интуиция, вера, ум оставляют герою возможность преодоления этой пустой бесконечности и внутри себя, и в окружающей его Бесконечности. Возможно, эта та основная мысль, которую хотел донести до читателя Автор.

Полученные результаты позволяют сделать вывод, что синестетически обусловленная звуко-цветовая картина мира состоит из нескольких компонентов, ведущими из которых является универсальная (бессознательная) способность человека ассоциировать звуки и цвета и национальное (подсознательное) свойство отражать специфику взгляда на мир через конкретный язык в образно-логическом и эстетическом восприятии. Преломление же этой картины в индивидуальном языковом сознании, ее функционирование и своеобразие исследователь может зафиксировать при помощи строгих методов выявления данной специфики в тексте, а затем интерпретировать результаты и продолжить изучение своеобразия ЗЦА в зависимости от индивидуального наполнения каждого речевого произведения. Являясь аналогом мироощущения, запечатленного в модели текста, звуко-цветовая картина мира не только продуцируется, но и может быть воспринята.

## *ЛИТЕРАТУРА*

Абдуллин И.Р. (1996): Цвето-музыкальные синестезии в поэзии Бальмонта // *Электроника, музыка, свет. Тезисы докладов*. Казань: Изд-во КАИ, 121-124.

Бальмонт К. (1917): *Светозвук в природе и световая симфония Скрябина*. Москва.: Нотный магазин Рос. муз. изд-ва.

Блок А.А. (1982): *Собр. соч. в 6 тт*. Москва: Прогресс.

Васильев И.Е. (1999): *Русский поэтический авангард XX в.* (Группа «41є»). Екатеринбург: УрГУ.

Вейдле В. (2002): *Эмбриология поэзии. Статьи по поэтике и теории искусства*. Москва: Языки славянской культуры.

Вейнрейх У. (1970): О семантической структуре языка // *Новое в лингвистике*. Москва: Прогресс. Вып. V. Языковые универсалии, 163-209.

Воронин С.В. (1982): *Основы фоносемантики*. Ленинград: Издво ЛГУ.

Воскресенская М.А. (2005): *Символизм как мировидение Серебряного века: социокультурные аспекты формирования общественного сознания российской культурной элиты рубежа XIX-XX веков*. Москва: Логос.

Галеев Б.М. (2004): Синестезия в эстетике и поэтике символизма // Синтез *в русской и мировой художественной культуре: Материалы Четвертой науч.-практ. конф., посвящ. памяти А.Ф. Лосева*. Москва: МПГУ, 50-55.

Казарин Ю.В. (2000): *Проблемы фоносемантики поэтического текста*: Учебное пособие. Екатеринбург: УрГУ.

Лотман Ю.М. (1972): *Анализ поэтического текста*. Ленинград: Просвещение.

Миллер-Будницкая Р.З. (1930): Символика цвета и синэстетизм в поэзии на основе лирики Блока // *Известия Крымск. пед ин-та*. Симферополь. Т. 3, 78-144.

Невзглядова Е. (1968): О звукосмысловых связях в поэзии // *Филологические науки,* 4, 23-34.

Павловская И.Ю. (2004): *Фоносемантический анализ речи*. СПб.: Изд-во СПб. ун-та.

Прокофьева Л.П. (2004): Лингвоцветовая картина мира Александра Блока // *Русская и сопоставительная филология: состояние и перспективы: Международная научная конференция, посвященная 200-летию Казанского университета* (Казань, 4-6 октября 2004 г.): Труды и материалы. Казань: Казан. гос. ун-т им. В.И. Ульянова-Ленина, 237-238.

Прокофьева Л.П. (1998): Понятие позиции звукобуквы в аспекте звукоцветовой ассоциативности // *Семантика языковых единиц*. Москва: МГОПУ, 331-333.

Прокофьева Л.П. (1992): Цветовая символика звука как категория идиостиля (на материале поэзии А. Блока и В. Набокова // Принципы *изучения художественного текста.* Ч. 2. Саратов: Изд-во СГПИ, 143-145.

Серов Н.В. (1990): *Хроматизм мира*. Л.: Васильевский остров.

Тарановский К. (2000): *О поэзии и поэтике* / Сост. М.Л. Гаспаров. Москва: Языки русской культуры. (Studia poetica).

Шуришина Т.И. (1999): *Актуальные проблемы стилистики текста (цветофоно-семантический аспект)*. Черновцы: Рута.

Эткинд Е.Г. (1998): *Материя стиха*. СПб.: Изд-во «Гуманитарный союз» (репринт Paris: Institut D'etudes Slaves, 1978).

Якобсон Р. (1983): В поисках сущности языка // *Семиотика*. Москва: Радуга, 102-117.

# Фоносемантический анализ текста

*Виктор Левицкий (Черновцы, Украина)*

Текст привлекает к себе внимание исследователей, в том числе и лингвистов, прежде всего своей содержательной стороной. Так, например, изучение частоты встречаемости в тексте таких слов, как *всегда, постоянно, никогда, все, никто, исключительно, только* и т.п. позволяют сделать вывод о степени догматичности автора текста (см. Ertel 1972: 257-258). Квантитативный анализ писем, оставленных самоубийцами и лже-самоубийцами, показал, что в письмах «истинных» самоубийц существенно реже встречаются прилагательные и наречия (по сравнению с глаголами); чаще встречаются глаголы в форме императива; число слов в предложении больше, чем в письмах «ложных» самоубийц, и т.д. (см. Osgood 1959; Merten 1983: 235).

Однако и формальная сторона текста (его фонетический уровень) издавна была объектом лингвистического анализа. Здесь лингвистов, как известно, интересовали такие явления, как аллитерация, ассонанс, ритмические повторы, рифма и т.п. Лишь постепенно исследователей стала интересовать «значимость» звуков, встречающихся в тексте, – прежде всего в поэтическом тексте. Наиболее часто цитируемой из работ такого рода является статья венгерского исследователя И. Фонадя, опубликованная в 1961 году в журнале «Word» (см. Fonagy 1961). Разделив стихотворные произведения Шандора Петефи на 6 «ласковых» и 6 «агрессивных», И. Фонадь установил, что большинство звуков встречается в обеих группах текстов с приблизительно одинаковой относительной частотой. Однако частота некоторых звуков существенно отличалась в каждой из двух групп: /l/, /m/, /n/ чаще встречаются в «нежных» произведениях, а /t/, /k/ и /r/ – в «агрессивных» (см. Fonagy 1961: 194-195). В целом такое распределение повторилось, по подсчетам И. Фонадя, в текстах Верлена, Гюго и немецкого поэта Рюккерта.[1]

[1] Наиболее важную литературу, посвященную фонетическому анализу текста (на начальном этапе развития такого направления) можно найти в книгах (Tsur 1992) и (Павловская 2001).

Менее известной в отечественной лингвистике была и, к сожалению, остается книга немецкого психолога З. Эртеля «Psychophonetik» (1969); содержание одной из статей З. Эртеля изложено выше.

З. Эртель провел серию экспериментов, целью которых было изучение звукосимволических свойств текстов. Два из этих экспериментов представляются нам наиболее интересными не только по полученным в них результатам, но и с точки зрения методики «фоносемантического» анализа текста (понятно, что З. Эртель не ведал, что занимался «фоносемантическим анализом»).

В одном из экспериментов было изучено распределение «слабых» и «сильных» согласных в опубликованных письмах известных личностей XIX столетия. Первый этап эксперимента состоял в том, что по просьбе экспериментатора одна из его сотрудниц (но, подчеркнем, не сам экспериментатор!), основываясь на биографических данных исследуемых личностей, составила 2 списка таких личностей: 15 «динамичных» и 14 «нединамичных» личностей. В первый список вошли, в частности, К. Маркс, Р. Вагнер, А. Гумбольдт, Л. ван Бетховен, Ф. Ницше, О. Бисмарк, Г. Гейне и др. На втором этапе специально обученные и натренированные сотрудники подсчитали частоту встречаемости звуков в первых 5 строках на каждой десятой странице книги, содержащей исследуемые письма. Наконец, на заключительном этапе эксперимента З. Эртель сравнил распределение частот 10 согласных, символические свойства которых были предварительно изучены экспериментатором в психолингвистических экспериментах. К этим согласным относились 5 «сильных» (p, t, k, f, s) и 5 «слабых» (b, d, g, v, z). Предполагалось, что в письмах «динамичных» личностей будут чаще встречаться «сильные» согласные, а в письмах «нединамичных» личностей – «слабые» согласные. Эта гипотеза подтвердилась относительно большинства из 15 «динамичных» личностей. Обратный результат зафиксирован в письмах А. Гумбольдта, Л. ван Бетховена; близки к нейтральному распределению частот письма Г. Гейне и О. Бисмарка.

Во втором из привлекших наше внимание экспериментов исследовались письма 7 творческих личностей (в том числе, Гете, Гумбольдта, Штифтера и др.). В письмах изучалось распределение перечисленных выше 10 «сильных» и «слабых» согласных и 10 кратких и долгих гласных (i, e, a, o, u). На этот раз категоризации были подвергнуты не «личности», а тексты, которые были

разделены по хронологическому принципу (каждые 10 лет жизни исследуемых авторов). Статистический анализ показал, что в целом частоты распределения исследуемых звуков в текстах первой и второй половины жизни творческих личностей коррелируют друг с другом, т.е. существенно не отклоняются от некоторой средней величины. Самым интересным было сравнение показателей «творческой продуктивности» личности и частоты употребления в текстах «сильных» и «слабых» звуков. Повышенное употребление сильных звуков совпадало с повышенной творческой активностью.

Что ж общего (с методической точки зрения) в экспериментах И. Фонадя и З. Эртеля? Общим, безусловно, является то, что во всех изложенных экспериментах (в том числе, и по «содержательному» анализу текста) все тексты подразделяются и противопоставляются по двум категориям: «динамичные-нединамичные», «агрессивные-нежные», принадлежащие истинным и ложным самоубийцам, догматическим и недогматическим личностям и т.п. Такой методический прием широко используется в таких науках, как биология и психология, где наряду с основной исследуемой группой объектов обязательно выделяется «контрольная» группа. Различия между методиками И. Фонадя и З. Эртеля состоят в том, что И. Фонадь в обеих противопоставленных группах текстов искал все звуки, употребление которых (по своей частоте) характеризует ту или иную группу, а З. Эртель заранее определял состав звуков, частоты которых подвергались изучению в каждой из двух групп. Назовем условно изложенную методику фоносемантического анализа текста *методикой Фонадя/Эртеля*.

Совершенно иным путем пошел А.П. Журавлев, который исходил из того, что «количественное выражение результатов измерения фонетического значения позволяет построить формальную процедуру анализа этого значения в поэтических текстах, в основе которой лежит сопоставление средних оценок символики звуков с отклонениями частотностей звуков в тексте от нормы» (Журавлев 1974: 99). Говоря более проще, А.П. Журавлев построил методику изучения «звуковой формы» стихотворного текста на тех же принципах, что и изучение фонетического значения слова. Как и при измерении фонетического значения слова, под «формой» художественного текста понимается употребление «звукобукв».

Методика анализа фонетического значения текста, предложенная А.П. Журавлевым, имеет те же принципиальные недостатки,

что и методика изучения фонетического значения слова: а) отбор текстов; б) интерпретация полученных результатов. По поводу первого – отбора текстов – А.П. Журавлев пишет: «анализу могут быть подвергнуты только такие произведения, в которых четко выражены или явно доминируют одна какая-либо тема, одно настроение, чувство» (там же, с. 108).

По поводу второго – интерпретации полученные результатов – А.П. Журавлев выдвигает такое условие: «необходимо, чтобы коннотативное содержание отобранных стихотворений допускало возможно более определенное интуитивное описание путем перечисления признаков лексикона» (там же, с. 108; под лексиконом имеется в виду перечень признаков, выданных ЭВМ после анализа текста).

Хотя в цитируемой монографии А.П. Журавлев довольно трезво оценивает «возможности и ограничения» своей методики (с. 112-116), в книге «Звук и смысл», как и следовало ожидать, тональность таких оценок меняется. Как и фонетическое значение слова, фонетическое значение стихотворного текста во всех приведенных А.П. Журавлевым примерах в той или иной степени гармонирует с его содержанием. Очевидно, чтобы объективно оценить соответствие вычисленного фонетического значения содержанию текста, необходимо сопоставить не интуитивные оценки содержания текста с найденным перечнем фонетических признаков, а объективные оценки содержания текста с соответствующими фонетическими признаками.

Такая процедура осуществлена в книге Л.П. Прокофьевой (см. Прокофьева 2007).

Подвергнув автоматизированному анализу огромное число художественных текстов (русских и английских писателей XIX-XX вв.) на основе частоты встречаемости «графонов» (учитывались также различия между ударными и безударными гласными), Л.П. Прокофьева на заключительном этапе сравнила результаты автоматического анализа текста (исследовалось его «цветовые» значение) с результатами «аудиторского» анализа (информантами, т.е. аудиторами, выступали школьники старших классов гимназии). Так, например, анализ стихотворения А. Блока «Та жизнь прошла…» выглядит следующим образом (см. табл. 1).

«Полученные результаты, – пишет Л.П. Прокофьева, – подтверждают мысль, что лексическая семантика, безусловно, оказывает решающее воздействие на восприятие реципиентов: 4 строка с цветовой номинацией *ночь* оценена как *черная* и *синяя*, 7 строка

(*снег*) – как *белая*, 12 строка (*пожар*) – как *красная*». И далее: «Информанты … вполне адекватно воспринимают фоносемантическую информацию в тех случаях, когда она не «перекрывается» лексической …» (Прокофьева 2007: 193-194)[2]. Сходные выводы были получены И.Ю. Павловский несколькими годами ранее: «При восприятии текстов аудиторы прежде всего опираются в своих оценках на смысловое содержание. Поэтому, чтобы обратить на себя внимание, звуковая инструментовка должна действовать «в унисон» с коннотативным значением текста» (Павловская 2001: 137). Этот вывод И.Ю. Павловской созвучен с теми выводами, к которым мы пришли, оценивая проявление звукового символизма в слове: «Экспериментальные данные показывают, что говорящий далек от того, чтобы искать определенные соответствия в каждом слове родного языка; звукосимволическое чутье «как бы дремлет» (по выражению А. Зиберера) в сознании человека и проявляется лишь тогда, когда слово, развиваясь в полном соответствии с фонетическими и морфологическими законами данного языка, случайно приобретает форму, соответствующую с точки зрения говорящего смыслу этого слова (при экспериментальном изучении звукосимволизма и моделируемся как раз «случайное» сближение ограниченного числа имен и ограниченного числа смыслов)» – см. Левицкий 1973: 90.

Таблица 1 (извлечение[3])

| Единица текста | Аудиторская оценка | Автоматизированная оценка |
|---|---|---|
| 1 строка | черно-белый | красный |
| 2 строка | синий и красный | зеленый |
| 3 строка | синий | бело-синий |
| 4 строка | черный и синий | бело-желтый |
| 5 строка | белый | красный |
| 6 строка | бело-синий | бело-жёлтый |
| 1 строфа | синий | бело-желтый |

Таким образом, исследования последних десятилетий показали, что соответствие фонетической формы текста и его содержания выглядит далеко не так, как это представлялось А.П. Журавлеву.

[2] См. выше статью Л. Прокофьевой, с. 247.
[3] См. выше в статье Л. Прокофьевой табл. 4.

Результаты, полученные И.Ю. Павловской и Л.П. Прокофьевой, являются убедительным подтверждением того, что носитель языка воспринимает коннотативное значение текста через его смысловое содержание, а не благодаря его фонетическому значению.

Попытаемся теперь проанализировать методику А.П. Журавлева с иной точки зрения – насколько она пригодна для изучения собственно фонетического значения текста. Это удобнее всего сделать на материале, содержащемся в монографии Л.П. Прокофьевой (разработанная ею программа для автоматизированного фоносемантического анализа текста основана на тех же принципах, которые заложены в работах А.П. Журавлева, – см. Прокофьева 2007: 167).

Л.П. Прокофьева, как показано выше, изучала цветовую окрашенность формы текста – графонов, входящих в его состав и образующих текст. При этом исследовались не только поэтические и прозаические (художественные) тексты, но и «нехудожественные» тексты на русском и английском языках (научные статьи, учебники, публицистические статьи и т.п.) – см. Прокофьева 2007: 172.

Понятно, что частота употребления тех или иных звуков (звукобукв) в тексте зависит от того, какие в нем употребляются слова. Допустим, например, что в учебнике по сердечно-сосудистой хирургии очень часто употребляются слова *сердце, сердечный, кровь, кровяной, кровеносный, аорта, коронарный, кардиология, хирургия, шунтирование, сосуд, артерия;* благодаря этому в тексте будет наблюдаться повышенная частота употребления звуков /р/, /к/ и /с/ – тех самых звуков, которые, по данным Л.П. Прокофьевой, ассоциируются в русском языке с красным цветом. В результате – в соответствии с программой вычисления фонетического значения – текст учебника по хирургии «окрасится» в красный цвет.

Значит ли это, что авторы учебника (сознательно или подсознательно) подбирали такие звучания, чтобы привести в гармонию содержание и форму учебника? Или они преследовали иные цели? Ответ очевиден. Поскольку все буквы русского или английского алфавита, как установила Л.П. Прокофьева, в большей или меньшей степени ассоциируются с тем или иным цветом, а сам алфавит является своего рода небольшим заумным текстом с постоянным составом графонов, и русский, и английский алфавит сами по себе должны иметь определенную окраску.

Попытаемся определить «цвета» русского и английского алфавитов, воспользовавшись данными Л.П. Прокофьеовой. Простейшая программа нахождения суммарной величины фонетического

значения некоторой совокупности букв, как и букв в слове, должна строиться на том, что частота буквы умножается на его оценку по данной шкале. Частоты букв русского и английского алфавитов приведены в монографии Л.П. Прокофьевой (с. 272). Как найти цветовые значения букв? Воспользуемся для этого нахождением веса каждой буквы в обозначении цвета, разделив частоту этой буквы в ответах испытуемых (в экспериментах Л.П. Прокофьевой – см. с. 264), по данному цвету на суммарную частоту той же буквы в ответах испытуемых по всем цветам. Например, 514 испытуемых приписали букве А значение «красный», всего в оценке буквы А принимало участие 859 испытуемых; следовательно, вес буквы А в обозначении красного цвета равен

$$\frac{514}{859} = 0.598 \approx 0.6.$$

Вес той же буквы в обозначении белого цвета равен

$$\frac{127}{859} = 0.148.$$

Будем учитывать для каждой буквы только такие веса, которые превышают 15 %, т.е. равны или больше 0,16 (исключения сделаны только для англ. orange и brown, где учтен графон H с весом 0,13).

Теперь несложно найти долю красного цвета в русском алфавите – см. табл. 2

Таблица 2

Доля красного цвета в русском алфавите

| буквы | частота | вес | цветовое значение |
|---|---|---|---|
| а | 0,95 | 0,6 | 0,570 |
| я | 0,24 | 0,52 | 0,125 |
| к | 0,33 | 0,44 | 0,145 |
| р | 0,38 | 0,37 | 0,141 |
| м | 0,05 | 0,36 | 0,018 |
| ю | 0,06 | 0,31 | 0,019 |
| л | 0,37 | 0,18 | 0,067 |
| ф | 0,03 | 0,16 | 0,005 |
| всего | | | 1,09 |

Сводные данные о цветовой окраске русского и английского алфавитов представлены в таблицах 3 и 4.

Как видно из этих таблиц, в ранжированном порядке цвета русского алфавита располагаются так: красный, зеленый, синий, черный, белый, желтый; цвета английского алфавита располагаются следующим образом: зеленый, желтый, красный, белый, коричневый, оранжевый; далее следуют синий, фиолетовый и черный. Эти данные можно теперь сравнить с данными самой Л.П. Прокофьевой, которая с помощью более сложной программы (нам осталось неясным только, каким образом в этих программах графонам приписывались цветовые значения) проанализировала «нехудожественные тексты». В большинстве случаев, – делает вывод Л.П. Прокофьева, – и в русских, и в английских произведениях рационалистического характера находит отражение желто-зеленая (англ.) и черно-белая (рус.) составляющие…» (Прокофьева 2007: 174). Как видим, данные по английскому языку в нашем эксперименте и в эксперименте Л.П. Прокофьевой полностью совпадают, а по русскому языку расходятся.

Если бы такие расхождения были обусловлены различиями использованных процедур, то результаты были бы различными по обоим языкам. Но поскольку для одного языка они идентичны, следует предложить, что тексты учебников и научных статей на русском языке не представляют, как предполагает Л.П. Прокофьева «национальной» системы звуко-цветовой ассоциативности». Можно предполагать, что черно-белая окрашенность русских текстов является некоторым отклонением от нормы. Такие отклонения, по данным Л.П. Прокофьевой, наблюдаются, например, в учебнике естественнонаучного профиля (частоты *и, п* превышены в 2 раза – цвет черно-синий) в газетной статье (частотность *р, м* превышена в 1,4 раза).

Однако независимо от того, какие из полученных данных наиболее близки к «национальной системе», следует сделать один важный вывод: какой бы текст на русском, английском или любом другом языке мы не подвергли бы анализу по методике А.П. Журавлева (еще раз подчеркнем, что в экспериментах Л.П. Прокофьевой использовалась видоизмененная процедура А.П. Журавлева), в результате этот текст, будь то поэзия, проза, учебник по грамматике или математике, обязательно обнаружит фонетическое значение – либо по шкале цвета, либо любой из 25 (19) шкал, использованных А.П. Журавлевым. И при желании всегда можно найти гармонию между смысловым содержанием текста и его фонетическим значением. Из этого следует только одно: методика, которую мы условно называем методикой А.П. Журавлева менее пригодна для фоносемантического анализа текста, чем методика

Фонадя/Эртеля. Она оказывается наиболее эффективной для изучения фонетического значения стихотворного текста при условии, что результаты автоматического анализа сопровождаются и проверяются с помощью аудиторского анализа.

Таблица 3
Цветовая окраска русского алфавита

| графоны | частота | красный | синий | зеленый | желтый | черный | белый | фиолетовый | коричневый | всего |
|---|---|---|---|---|---|---|---|---|---|---|
| а | 0,95 | 0,57 | | | | | | | | 0,570 |
| б | 0,18 | | | | | | 0,077 | | | 0,077 |
| в | 0,39 | | 0,129 | | | | | | | 0,129 |
| г | 0,15 | | 0,035 | | | | | | 0,027 | 0,062 |
| д | 0,37 | | | | | 0,067 | | | | 0,067 |
| е+ё | 0,89 | | | 0,801 | | | | | | 0,801 |
| ж | 0,08 | | | | 0,038 | | | | | 0,038 |
| з | 0,15 | | | 0,083 | | | | | | 0,083 |
| и | 0,56 | | 0,235 | | | | | | | 0,235 |
| й | 0,13 | | 0,046 | | | | 0,036 | | | 0,082 |
| к | 0,33 | 0,145 | | | | | | | | 0,145 |
| л | 0,37 | 0,067 | 0,067 | | 0,074 | | | | | 0,208 |
| м | 0,32 | 0,018 | | | | | | | | 0,018 |
| н | 0,64 | | 0,198 | | | | | | | 0,198 |
| о | 1,04 | | | | 0,291 | | 0,364 | | | 0,655 |
| п | 0,26 | | | | | 0,083 | | | | 0,083 |
| р | 0,38 | 0,141 | | | | | | | | 0,141 |
| с | 0,49 | | 0,168 | | | | | | | 0,168 |
| т | 0,75 | | | | | 0,225 | | | | 0,225 |
| у | 0,29 | | 0,075 | 0,078 | | | | | | 0,153 |
| ф | 0,03 | 0,005 | 0,006 | | | | | 0,006 | | 0,018 |
| х | 0,09 | | | | | 0,023 | 0,017 | | | 0,040 |
| ц | 0,04 | | | | 0,016 | | | | | 0,016 |
| ч | 0,20 | | | | | 0,098 | | | | 0,098 |
| ш | 0,12 | | | | | 0,039 | | | | 0,039 |
| щ | 0,03 | | | | | 0,007 | 0,005 | | | 0,012 |
| ы | 0,16 | | | | | 0,035 | | | 0,035 | 0,070 |
| э | 0,05 | | | 0,009 | 0,010 | | | | | 0,019 |
| ю | 0,06 | 0,019 | | | | | | | | 0,019 |
| я | 0,24 | 0,125 | | | | | | | | 0,125 |
| всего | | 1,09 | 0,959 | 0,971 | 0,429 | 0,577 | 0,499 | 0,006 | 0,062 | |
| ранг | | 1 | 3 | 2 | 6 | 4 | 5 | 8 | 7 | |

Таблица 4
Цветовая окраска английского алфавита

| графоны | частота | red | yellow | green | white | brown | blue | orange |
|---|---|---|---|---|---|---|---|---|
| A | 0,82 | 0,138 | | | | | | |
| B | 0,15 | | | | | | 0,049 | |
| C | 0,28 | | 0,050 | 0,056 | | | | 0,056 |
| D | 0,43 | | | 0,082 | | 0,116 | | |
| E | 1,27 | | 0,216 | 0,216 | 0,191 | | | |
| F | 0,22 | | | 0,051 | | | | |
| G | 0,20 | | | 0,064 | | | | |
| H | 0,61 | | 0,092 | | 0,092 | 0,079 | | 0,079 |
| I | 0,02 | | 0,003 | | 0,007 | | | |
| J | 0,02 | | | 0,003 | | | | |
| K | 0,08 | 0,014 | 0,088 | | | | 0,014 | |
| L | 0,40 | | | | | | 0,072 | |
| M | 0,24 | 0,053 | | | | | 0,012 | |
| N | 0,67 | | | 0,114 | | 0,107 | | |
| O | 0,75 | | | | 0,188 | | | 0,113 |
| P | 0,19 | 0,039 | | | | | | |
| Q | 0,01 | 0,002 | | | | | | |
| R | 0,60 | 0,282 | | | | | | |
| S | 0,63 | | 0,139 | | | | | |
| T | 0,91 | 0,016 | | 0,182 | | | | |
| U | 0,28 | | | | | | 0,045 | |
| V | 0,10 | | | | | | | |
| W | 0,24 | | | | 0,048 | | 0,053 | |
| X | 0,02 | | | | | | | |
| Y | 0,20 | | 0,074 | | | | | |
| Z | 0,01 | | | | | | | |
| Всего | | 0,544 | 0,662 | 0,768 | 0,526 | 0,302 | 0,245 | 0,248 |
| Ранг | | 3 | 2 | 1 | 4 | 5 | 7 | 6 |

Если бы такие расхождения были обусловлены различиями использованных процедур, то результаты были бы различными по обоим языкам. Но поскольку для одного языка они идентичны, следует предложить, что тексты учебников и научных статей на русском языке не представляют, как предполагает Л.П. Прокофьева «национальной» системы звуко-цветовой ассоциативности». Можно предполагать, что черно-белая окрашенность русских текстов является некоторым отклонением от нормы. Такие отклонения, по данным Л.П. Прокофьевой, наблюдаются, например, в учебнике естественно-

научного профиля (частоты *и, п* превышены в 2 раза – цвет черно-синий) в газетной статье (частотность *р, м* превышена в 1,4 раза).

Однако независимо от того, какие из полученных данных наиболее близки к «национальной системе», следует сделать один важный вывод: какой бы текст на русском, английском или любом другом языке мы не подвергли бы анализу по методике А.П. Журавлева (еще раз подчеркнем, что в экспериментах Л.П. Прокофьевой использовалась видоизмененная процедура А.П. Журавлева), в результате этот текст, будь то поэзия, проза, учебник по грамматике или математике, обязательно обнаружит фонетическое значение – либо по шкале цвета, либо любой из 25 (19) шкал, использованных А.П. Журавлевым. И при желании всегда можно найти гармонию между смысловым содержанием текста и его фонетическим значением. Из этого следует только одно: методика, которую мы условно называем методикой А.П. Журавлева менее пригодна для фоносемантического анализа текста, чем методика Фонадя/Эртеля. Она оказывается наиболее эффективной для изучения фонетического значения стихотворного текста при условии, что результаты автоматического анализа сопровождаются и проверяются с помощью аудиторского анализа.

Методика Фонадя/Эртеля приспособлена, как показано выше, для решения более широкого круга задач.

Вернемся еще раз к этой методике. Она, как показано выше, допускает два основных процедурных варианта: 1) экспериментатор подвергает анализу частоту встречаемости всех фонетических единиц в каждой из двух противопоставленных групп текстов и находит те единицы, частоты которых превышены в первой и во второй группе; 2) экспериментатор исследует в каждой группе текстов частоты только тех единиц, которые по данным психометрических измерений (например, по шкале силы или оценки) **соответствуют** шкале, по которой противопоставлены группы текстов («динамичные-нединамичные тексты»). Только второй процедурный вариант обеспечивает экспериментатору выполнение подлинно фоносемантического анализа текста, т.е. такого анализа, целью которого является поиск звукосмысловых ассоциаций, основанных на **символических** значениях звуков.

Это не значит, что первый вариант должен быть полностью исключен из практики фоносемантического анализа текста. Его использование требует от экспериментатора дополнительного анализа попавших «в финал» звуков. Если, например, «ласковыми»

звуками, как у Фонадя, оказались l, m, n, то необходимо сопоставить эти данные объективного звукосимволизма с данными субъективного звукосимволизма. Как установлено в фоносемантике, звуки l, m, n, действительно, сосредоточены в различных языках на полюсе шкалы «приятный» (см. Левицкий 1973: …). Но если в числе фаворитов в той или иной группе текстов оказываются звуки, не соответствующие заданной шкале, они не могут приниматься во внимание как звуки с определенным символическим значением. Их следует отнести к звукам с функциональным значением в данном языке (о терминах «иконическое» и «функциональное» значение см. в (Левицкий 2008: …)). При группировке текстов по шкалам вовсе не обязательно использовать нефундаментальные шкалы («ласковый-грубый»; «веселый-грустный» и т.д.), поскольку в психолингвистике давно установлено, что все множество шкал сводится к трем основным: шкала оценки («приятный-неприятный»), шкала силы («слабый-сильный») и шкала активности («медленый-быстрый», «активный-пассивный»). В некоторых экспериментах З. Эртель объединил даже 2 последних шкалы в одну «динамичный-нединамичный». Практика показывает, что объединение текстов по такой шкале, например, как «минорный-мажорный» (см. Найдеш 1998) в конечном итоге переформулируется в шкалу «неприятный-приятный», т.к. в «минорных» текстах превалирует /r/, а в «мажорных» – /l/. Такое распределение соответствует шкале «неприятный-приятный», а не, например, шкале «веселый/грустный», т.к. в русском языке (по данным А.П. Журавлева /r/ имеет значение «веселый» (2,6). Итак, можно рекомендовать исследователям фоносемантических свойств текстов группировать тексты по трем или даже двум основным шкалам – шкале оценки и шкале динамики. Если тексты сгруппированы правильно, в динамичных текстах следует ожидать превышения частот «сильных» и «активных» звуков, а в нединамичных – «слабых» и «медленных».

В практике фоносемантического анализа могут встретиться однако и такие процедурные варианты, когда тексты не группируются по осгудовским шкалам. В тексте могут выявиться несколько частотных или даже один сверхчастотный звук. Так, например, Ю.Ю. Павловская, основываясь на проведенных ею экспериментах, полагает, что в тексте целесообразно выделять один или несколько ключевых звуков. Таким ключевым звуком в некоторых текстах, исследованных И.Ю. Павловский, оказался звук w, входящий в состав англ. слова wild. Такой вывод созвучен с

нашей гипотезой о том, что наличие в звучании слова хотя бы одного звука (особенно ударных гласных), соответствующего семантике этого слова, дает испытуемому основание считать, что слово обладает фонетической мотивированностью (см. Левицкий 1994: 38). Очевидно, тот же механизм действует и при оценке испытуемыми текста: «если в тексте невозможно выделить ключевой звук или звуки, то аудиторы не в силах правильно уловить фоносемантическую символику текста» (Павловская 2001: 137). Можно согласиться, таким образом, с тем, что наличие в слове или тексте какого-либо звука, соответствующего содержательной стороне той или иной единицы языка или речи (слово, предложение, стихотворная строфа, небольшой текст), может оказаться решающим фактором в восприятии этой единицы языка или речи как «фонетически мотивированной». И все же экспериментатор должен избегать слишком поспешных и прямолинейных выводов. Н.Л. Львова обнаружила сверхчастое употребление в английской публицистике начального сочетания pr. При использовании методики Л.П. Прокофьевой следовало бы сделать вывод, что текст окрашен в красный цвет (именно так по ее данным окрашена английская публицистика – см. Прокофьева 2007: 174). При использовании методики И.Ю. Павловской следовало бы сделать вывод, что исследованный Н.Л. Львовой текст «неприятный», т.к. «ключевой» звукокомплекс pr оценивается аудиторами И.Ю. Павловской как «неприятный» (см. Павловская 2001: 110). Все дело, однако, в том, что в тексте, о котором идет речь, очень часто встречается слово president… Так при чем же здесь звуковой символизм?

## *ЛИТЕРАТУРА*

Журавлёв А.П. (1974): *Фонетическое значение*. – Ленинград: ЛГУ.

Левицький В.В. (1973): *Семантика и фонетика*. Пособие подготовленное на материале экспериментальных исследований. – Черновцы: ЧГУ.

Левицкий В.В. (1994): Фонетическая мотивированность слова // *Вопросы языкознания,* №1, 26-37.

Левицкий В.В. (2008): *Семантические и фонетические связи в лексике индоевропейского праязыка. Опыт квантитативного анализа этимологического словаря*. – Черновцы, Рута.

Львова Н.Л. (2005): Фонетико-семантичне дослідження трьох видів текстів // *Науковий вісник Чернівецького університету: Зб. наук. пр*. – Вип. 232. Германська філологія. – Чернівці: Рута, 180-191.

Найдеш О.В. (1998): Явище фоносемантизму в німецькомовній поезії / Найдеш О.В. // *Науковий вісник Чернівецького університету: збірник наукових праць*. – Вип. 41: Германська філологія. – Чернівці: ЧДУ, 14-25.

Павловская И.Ю. (2001): *Фоносемантический анализ речи*. – СПб.: Изд-во С.-Петербургского ун-та.

Прокофьева Л.П. (2007): *Звуко-цветовая ассоциативность: универсальное, национальное, индивидуальное*. – Саратов: Изд-во Саратовского медицинского ун-та.

Ertel S. (1969): *Psychophonetik (Untersuchungen über Lautsymbolik und Motivation)*. – Göttingen.

Ertel S. (1972): Erkenntnis und Dogmatismus // *Psychologische Rundschau*, № 23, 241-269.

Fonagy J. (1961): Communication in Poetry // *Word*, vol. 17, № 2.

Merten K. (1983): *Inhaltsanalyse. Einführung in Theorie, Methode und Praxis*. – Opladen: Westdeutscher Verlag.

Osgood Ch. (1959): The Representational Model and Relevant Research Methods // Pool I. (ed.) *Trends in Content Analysis*. – Urbana: University of Illinois Press, 33-88.

Tsur R. (1992): *What Makes Sound Patterns Expressive?* – Durham and London: Duke University Press.

# Теоретико-методологические предпосылки анализа текста англоязычной шутки

*Виктория Самохина (Дмитренко) (Харьков, Украина)*

Современное состояние лингвистической науки характеризуется полипарадигматичностью и синтетичностью концепций, подходов и **методологий**, что создало все необходимые предпосылки для изучения текста англоязычной шутки, который представляет собой сложный культурный феномен, аккумулирующий в себе свойства различных областей человеческого бытия, активно регулирующий юмористическую коммуникацию.

**Актуальность** данного исследования определяется наличием в современной лингвистике существенных теоретико-методологических предпосылок для изучения англоязычного юмора малых форм, а также многоаспектностью объекта исследования. Научное описание современного юмора малых форм может претендовать на полноту и самодостаточность только в том случае, если оно осуществлено с использованием концептуального аппарата различных научных направлений.

Терминологическая неопределенность понятий «юмористический интрадискурс», «шутка», отсутствие системного подхода к теории функционально-коммуникативной стилистики текста обуславливает необходимость дальнейшего всестороннего анализа их природы с использованием новейших методов, ориентированных на проникновение в глубинные механизмы языкового творчества.

**Методологической базой** исследования является антропологический по своей сущности текстово-дискурсивный анализ, который утверждает в лингвистике принцип функционализма и интегрирует когнитивно-прагматический (А.Д. Белова, Л.И. Белехова, О.П. Воробьева, С.А. Жаботинская, В.В. Козловский, И.М. Колегаева, В.А. Кухаренко, М.М. Полюжин, А.Н. Приходько и др.) и функционально-семантический (Ю.А. Зацный, Р.П. Зоривчак, А.Э. Левицкий, В.В. Левицкий, Л.Ф. Омельченко, Л.И. Сахарчук, С.А. Швачко) аспекты деятельностного подхода к изучению и функционированию языковых единиц в тексте и дискурсе.

Методологической основой исследования служит также концепция об экстралингвистической сущности фольклорного текста, подход к данному типу текста как средству хранения и передачи информации (в статике) и как к когнитивному процессу, выступающему в роли орудия познания (в динамике).

Синтезированный характер исследования обуславливает использование ряда взаимодействующих **методов и методик** традиционной функционально-семантической парадигмы компонентного анализа – для инвентаризации, систематизации и функционально-прагматической классификации текстов англоязычных шуток; фрагменты концептуального анализа применяются для когнитивно-семантической интерпретации юмора малых форм; контекстологического анализа – для выявления дискурсивых реализаций текстов англоязычных шуток как речевых жанров; интерпретационно-текстового анализа – для толкования пропозиционального содержания компонентов шутки с использованием методики прагмастилистического и когнитивно-семантического анализа и дальнейшей интерпретации антропонимов, тропов, интертекстуальных ссылок и т.д., т.е. экспликации когнитивного механизма категории комического – инконгруэнтности – в тексте англоязычной шутки; этнопрагматический метод дает возможность выявить этнокультурное своеобразие менталитета британской и американской народных смеховых культур, отображенного в юморе, и проследить это влияние на выбор способов маркирования комического в шутках.

Возможность масштабного, общекультурного рассмотрения юмора малых форм была создана фундаментальными исследованиями карнавальных и смеховых традиций Европы (Бахтин 1995; Рюмина 2003; Фрейденберг 1987) и Древней Руси (Лихачев, Панченко, Понырко 1984). В данном случае, особую актуальность приобретает выявление амбивалентности мировосприятия, которая определила существование двух параллельных моделей мира – официальной, канонической, и неофициальной, игровой, карнавальной.

Не меньшую значимость при рассмотрении комических текстов малых форм имеет обширный корпус концепции игры как феномена человеческой культуры. Среди разнообразных интерпретаций игры с позиции различных научных направлений следует особо отметить квалификацию игры как культуросозидающей социальной практики (Земская, Китайгородская, Розанова 1987; Левицкий 2007; Санников 2002; Сниховская 2005;

Хейзинга 1997). Закономерным развитием теорий этой группы стало введение в практику анализа термина "комическая ситуация" (Рюмина 2003).

Исследования в области лингвостилистики не могут оставаться в стороне от ведущих тенденций современной лингвистики. Обращение к коммуникативному аспекту языковых явлений неизбежно ставят вопросы, решением которых занимается новое направление стилистики – функционально-коммуникативная стилистика текста (ФКСТ), которая требует комплексного рассмотрения проблемы взаимодействия коммуникативного и стилистического аспектов функционирования языковых единиц. До настоящего момента данное направление занималось изучением, преимущественно, научного дискурса. Юмористический текст впервые становится объектом исследования в рамках ФКСТ.

Ориентация современной лингвистики на функциональный аспект языковых единиц в контексте дискурс-анализа определила новые подходы к изучению англоязычного фольклорного юмористического текста малой формы в рамках функционально-коммуникативной стилистики текста, что дало возможность выявить его текстово-дискурсивные параметры и представить как холистическую единицу в единстве формальных, текстово-композиционных и функционально-коммуникативных параметров.

Экстраполяция функционально-коммуникативного подхода на антропологичекую методологию исследования позволила осуществить анализ текста англоязычной шутки как интеллектуально-эмоционльной деятельности, выявить пути и способы, которые обуславливают и модифицируют его параметры.

Юмор – наиболее распространенная современная форма комического, которая когнитивно исходит из глубокого чувства трагизма и существует только в ситуациях общения, в юмористическом интрадискурсе. Ядро юмора сосредоточено в стимуле, и таким стимулом является комическая инконгруэнтность. Смешное видится как продукт инконгруэнтности – своеобразной игры противоречиями, противопоставлениями, несовместимостями.

Инконгруэнтность находит выражение в языке и речи малых фольклорных юмористических текстов – шуток, которые являются основной формой выражения юмористического интрадискурса (т.е. речевым жанром и текстотипом), в рамках которого происходит переключение с языка институциональной на язык прагмалингвистической сферы общения.

Текст шутки выполняет регулятивную функцию в дискурсе, есть способ речевого влияния на коммуникативное поведение участников общения, т.е. шутка – это логико-понятийная, ситуативно-вербальная единица коммуникации, ингерентными признаками которой являются игровой элемент комического, направленность на ситуативный конфликт, вызванный как ее языково-речевыми, так и содержательными свойствами, обманутое ожидание и принцип инконгруэнтности, имеющие сквозной характер и реализующиеся на всех уровнях. На их основе гиперболизованная, условная реальность создается в шутке на короткое время, имеет культурное пространство и по форме представляет собой текст малой формы с неожиданной, инконгруэнтной концовкой, характеризующейся разряжением напряжения от обманутого ожидания, проявляющегося в смеховой реакции.

Юмористическая коммуникация подчиняется стилистическому принципу нарушения стандартных норм и экспликации оценки. В этом проявляется ее нормативная инконгруэнтность, т.е. несоответствие в шутке является ее нормой и источником комической тональности.

Текстотипологическое изучение шутки поставило вопрос о ее текстовой норме, опирающейся на трехаспектную характеристику: по разновидностям (семантическим, стилистическим, структурно-композиционным) составляющих ее единиц; по комбинаторике этих единиц; по размещению их в тексте шутки. Таким образом, норма текста шутки приобретает концептуальный композитивный характер.

Двухчастность текста шутки – диктумно-модусное устройство (объективное и субъективное начала) – является ее ингерентным признаком, а наличие двух несовместимых фреймов создают в нем инконгруэнтность. В рамках одного фрейма ситуация преподносится как объективная. На этот фрейм накладывается фрейм субъективный, в котором гиперболизируются черты персонажа шутки или ситуации. Пока такая ситуация с наложением фреймов существует, она ассоциируется с единственным фреймом как реально-вымышленная. В плане семантики инконгруэнтность проявляется в карикатурном изображении образа, обыгрывании двусмысленности; в плане прагматики – в высмеивании персонажа, состоящем в понимании его статуса в виде шутливого переворачивания общепринятых норм; в плане синтактики – в соотношении текста шутки с его формальными типами.

Нарративная шутка (или анекдот) является наиболее распространенной разновидностью англоязычной шутки, в жанровой

структуре которой выделяются две группы – группа нарративных элементов и группа экспрессивных элементов. Нарративность анекдота связана с театральностью, исполнением, т.е., с игрой. Комический эффект в анекдоте достигается благодаря концептуальному раскрытию его композиционной модели, эксплицирующейся в блоках: заглавии, интродуктивном, компликативном, эксплозивном и эксплицированию ситуационной и лингвистической двусмысленности, т.е. инконгруэнтности.

На ситуативном уровне нарушения норм в тексте анекдота наблюдаются на следующих уровнях: субъект – объект; человек – пространство, время; человек – деятельность. Нарушения лингвистических норм: игра звуками, спунеризмы, каламбуры (омографы, омофоны, омонимы, омофоноиды, тропы). Средства логико-речевой инконгруэнтности: двойное значение (двусмысленность), иронические фигуры речи, малапропизм, оксюморон, прием «наивная правда», прием «реверс» и др. Разновидностью инконгруэнтности является бинарность, которая проявляется на всех уровнях: коммуникативном (в адресованности), коммуникативной структуры (тема-рема); содержательном (инконгруэнтность объективной и вымышленной реальности); персонажной подсистемы (бинарные оппозиции по рангу, по принадлежности к определенному полу, по положению в обществе); в переворачивании трагического в комическое и т.д.

Описание нормативно-текстового образца шутки предполагает сохранение специфики данного явления, что означает ее трехаспектную характеристику:

а) по разновидностям – семантическим, стилистическим, структурно-композиционным – составляющих ее единиц; б) по комбинаторике этих единиц; в) по размещению их в тексте шутки. Таким образом, *норма* шутки имеет концептуальный композитивный характер: это – композиция языковых средств, отвечающая типовой целеустановке и сумме типовых условий определенной коммуникативной ситуации. Изучение текста шутки представляется правомерным на следующих нормативных основаниях: с опорой на критерий авторитетности текста, ориентацией на законы жанра и стиля, учете типовых условий общения. Текстовая норма шутки состоит, по нашему мнению, в трояком подходе к ней. Такой подход представлен О.П. Воробьевой, выделившей статус текста согласно трем системным ценностям: текст-макрознак; текст-коммуникат; текст-дискурс (Воробьева 1991: 15-17).

Каждый уровень системной целостности текста шутки предполагает характерный для него тип восприятия: для текста-макрознака – узнавание, идентификацию; для текста-коммуниката – понимание, интерпретацию; для текста-дискурса – декодирование, расшифровку.

Несмотря на значимость, древнее происхождение шутки, представления о ней весьма неопределенны.

Так, Советский энциклопедический словарь (Сов. энцикл. словарь 1982) вообще не дает определения шутке. В нем 1599 страниц, и тематически он даже перенасыщен. В чем же дело? Может быть, дать определение шутке не представляется возможным? Почему тогда, когда мы говорим: «Знаешь эту шутку?» люди начинают улыбаться, значит, они понимают суть данного явления. Может быть, это имели ввиду составители данного словаря, не включив статью о шутке? Зачем объяснять то, что и так хорошо известно.

С другой стороны, «The American Heritage Dictionary», насчитывающий 1568 страниц, предлагает следующее определение шутки:

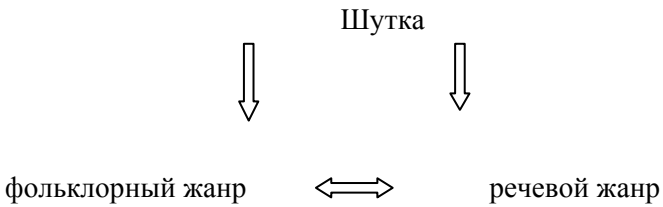
Шутка (сущ.) 1. короткий, забавный рассказ, особенно такой, который содержит пуант. 2. Забавная или шуточная ремарка; каламбур. 3. озорной трюк; проделка, шутка. 4. Забавный, шуточный случай или ситуация. 5. Что-то, что не воспринимается всерьез; тривиальность. 6. Предмет развлечения или смеха, посмешище. – (глагол). шутить, подшучивать. 1. Рассказывать или показывать шутки. 2. Говорить в шутливом тоне; быть веселым. – Смеяться над кем-либо, дразнить.

Эти существительные относятся к формам юмористических высказываний или действий. Joke и jest, которые обозначают что-то сказанное или сделанное, практически взаимозаменяемы, хотя jest встречается сейчас реже в этом смысле. Witticism относится к вербальному юмору, обычно с интеллектуальным контекстом и оформленный в фразу. Quip подразумевает легкую, колкую, добродушную ремарку и sally – внезапное, умное или остроумное высказывание. Crack и wisecrack относятся к легкомысленным или высмеюващим остроумным репликам или к экспромптным ремаркам в ответ на особую ситуацию. Gag в основном применим к неприличной комической ремарке или, реже, к комической побочной театральной сцене (American Heritage 1983: 690).

Как видно из американского источника, определение шутки весьма широкое. Тем не менее, эти определения, видимо, отражают суть одного и того же явления, которое представлено вариативными формами выражения, имеющими один общий «стержень» – наличие

так называемого «punch line», или пуанта, без которого ни одна из этих разновидностей существовать не может.

Текст шутки как фольклорный жанр не был предметом монографического исследования. Вместе с тем, англоязычная шутка впервые становится объектом анализа и как речевого жанра. Общее, что объединяет шутку в этих подходах – это понятие «жанр»:




Мы рассматриваем шутку как фольклорный и речевой жанр, т.е. как динамичное и статичное текстово-дискурсивное образование. Следовательно, шутку можно считать разновидностью фольклорных жанров, типовой структурной моделью, относящейся к классу аппелятивных текстов, и обладающей способностью реализовать определенную жизненную установку.

С точки зрения лингвистики текста, шутка рассматривается как тип текста. Доказательством тому – основные стандартные критерии определения текста, выделенные в работах как отечественных – И.Р. Гальперина, Ю.М. Лотмана, О.И. Москальской, О.П. Воробьевой и др., так и зарубежных лингвистов – К. Гаузенблас, Х. Изенберг, В. Ингве, Б. Палек, М. Пфютце, П. Сгалл, J. Culler, W. Dressler, N.E. Enkvist, P. Hartmann.

Текст шутки является самой распространенной формой юмора, поддающийся изучению в силу своей относительно стабильной структуры.

Как тип текста шутка характеризуется следующими чертами: завершенность, информативность, членимость, когезия и когерентность, воспроизводимость, модальность, экспрессивность, действенность (максимально достигать поставленной говорящим цели), уместность, интертекстуальность, функционализм и другие. Исследование типологических особенностей текста шутки подтверждает положение о том, что вся система средств художественного выражения данного типа текста, его стиль, находятся в сложной, но ощутимой связи с его жанровой принадлежностью и отвечает следующим основным параметрам: малый объем, который обычно измеряется количеством

строк (в среднем 1-7), архитектоническое оформление (в пределах 1-6 предложений-высказываний); стереотипность концептуальной композиционной модели; неосложненность синтаксиса, краткость лингвистических единиц, конституирующих текст, целевая установка.

Таким образом, шутку можно рассматривать как юмористическую смысловую миниатюру, которой свойственны: 1) коммуникативная четкость (новое и данное эксплицитно выделены и противопоставлены друг другу); 2) информативная плотность (минимальное количество информационных единиц в шутке ниже за счет различных включений коннотативного характера).

Для текста шутки характерна такая содержательно-фактуальная информация, которая тяготеет к типологизации в описании событий. В подобных текстах «постепенно вырабатываются модели, облегчающие быстрое и эффективное декодирование. Вот почему, вчитываясь в более или менее однотипные тексты, мы начинаем улавливать структурные особенности, способствующие ускорению процесса понимания в связи с определённой степенью предсказуемости». (Гальперин 1977: 526-527).

Наиболее типичными стилевыми чертами текста шутки являются следующие:

1. современность. Так, исторические анекдоты скорее необходимо считать историческими рассказами, воспринимающимися адресатом как реальные комические факты действительности. Более того, вряд ли они уже являются частью фольклора: ведь изустная форма передачи ими практически утеряна: их не передают из уст в уста, а читают в специальных сборниках с соответствующими комментариями. Некоторые разновидности шуток (напр., шутки об исторических личностях), хотя и создаются по канонам традиций фольклора, не могут оставаться традиционным, долговечным жанром. Вероятно, необходимо принять во внимание текучесть такого жанра, переход одного в другой, так же, как и существование произведений переходного типа, которые иногда трудно или вряд ли возможно отнести к тому или иному жанру;
2. жанровая модель – краткая когнитивная форма-модель шутки, управляющая юмористическим интрадискурсом;
3. интертекстовый римейк – возрождение старых, известных шуток и создание на их основе новых, «по мотивам»;
4. иррадиация экспрессивности – перенос коннотативного фона слова на определённый фрагмент текста шутки;

5. мимезис – правдоподобие шутки, её сходство, но не тождество с действительностью;
6. обманутое ожидание – нарушение предсказуемости дальнейшего развития событий в шутке;
7. серийность – явление текстовой синтагматики, наличие циклов, сборников;
8. ситуативные знания – фоновые знания о ситуации общения, о теме, в т.ч. совокупность всех имплицитных знаний, используемых в шутке, представленных текстовыми сигналами, но не вытекающими из содержания шутки;
9. эвокативность – способность шутки вызывать в сознании адресата большое количество ассоциаций;
10. диалогичность юмористической коммуникации, направленная либо на адекватную вербальную юмористическую (неюмористическую), либо на невербальную (смеховую) реакцию адресата. Диалогичность (в терминах М.М. Бахтина) связана в шутке «перекличкой» с другими прецедентными текстами, что вносит в текст шутки оттенок экспрессивности, которая направлена на создание комического эффекта;
11. иносказательность, намёк – шутки с их тенденцией к иносказательности и намёку, когда адресат сам извлекает скрытый смысл, получая от этого удовольствие;
12. карнавальное сознание, характеризующееся стремлением к фамилиаризации, снижению, приземлению всего, что воспринимается как значимое и серьёзное во внекарнавальной жизни, всех официальных ценностей и фетишей;
13. мгновенное восприятие юмористического текста (всё, что в нём пародируется или просто упоминается, должно входить в фон знания адресата и быть для него актуальным). Бессмысленно высмеивать никому не известный текст и также бессмысленно использовать его в качестве средства осмеяния.;
14. лазеечный адресат – по М.М. Бахтину, нададресат, понимание текста, которое возможно в будущем (ср., «дошло, как до слона на третьи сутки»);
15. воспроизводимость, вариативность, интертекстуальность;
16. краткость, пуантированность;
17. коагуляция – высвечивание на этапе внутреннего программирования протовербальной схемы будущего юмористического высказывания в сознании адресата;

18. эмпатия – способ понимания шутки, основанный на сопереживании, в частности, проявляющийся в смеховой реализации.

Эти стилевые черты, или, как их называют в новейших разработках, стилевые доминанты, юмористического стиля, отражают типичные, регулярно повторяющиеся и поэтому стандартизированные характеристики, общие для каждой шутки. Они являются результатом стилеобразующих экстралингвистических факторов, которые определяют особую системность в употреблении тех, а не иных, языковых структур, текстовых единиц внутри данного текстового целого.

*ЛИТЕРАТУРА*

Бахтин М.М. (1965): *Творчество Франсуа Рабле и народная культура Средневековья и Ренессанса*. – Москва: Худ. лит.

Воробьёва О.П. (1993): *Текстовые категории и фактор адресата*. – Київ: Вища школа.

Гальперин И.Р. (1977): Грамматические категории // *Известия АН СССР*. – Сер. лит. и яз., № 6. – Т. 36, 526-527.

Земская Е.А.; Китайгородская М.В.; Розанова Н.Н. (1987): Языковая игра // *Русская разговорная речь. Фонетика. Морфология. Лексика. Жест*. – Москва: Наука, 172-214.

Левицкий А.Э. (2007): Комическое: играем языком // *Логический анализ языка. Языковые механизмы комизма* / Отв. ред. Н.Д. Арутюнова. – Москва: Изд-во «Индрик», 295-307.

Лихачёв Д.С.; Панченко А.М.; Понырко Н.В. (1984): *Смех в Древней Руси*. – Ленинград: Наука.

Рюмина Т.М. (2003): *Эстетика списка. Смех или виртуальная реальность*. – Москва: УРСС.

Санников В.З. (2002): *Русский язык в зеркале языковой игры*. – 2-е изд., испр. и доп. – Москва: Языки славянской культуры.

Сніховська І.Е. (2005): *Механізми, засоби та прийоми мовної гри в сучасній англійській мові*. – Автореф. дис. … канд. філ. наук: 10.02.04. – Харків.

*Советский энциклопедический словарь*. – Москва: Советская энциклопедия, 1982.

Фрейденберг О.М. (1997): *Поэтика сюжета и жанра*. – Москва: Лабиринт.

Хейзинга Й. (1992): *Homo Ludens. В тени завтрашнего дня* / Й. Хейзинга. – Москва: Прогресс; Прогресс – Академия.

*The American Heritage Dictionary*. 2nd College edition, Boston: Houghton Mifflin Company, 1985.

# Length and Style in Slovak

*Zuzana Venitová and Emília Nemcová*[1] *(Trnava, Slovakia)*

## *1. INTRODUCTION*

Studies of sentence length are more than 100 years old (cf. Sherman 1888) and the literature concerning it is rather immense (cf. Köhler 1995; Best 2001; http://www.gwdg.de/~best/litlist.htm). A part of the studies are about probabilistic modelling, others about length in relation to sentence structure, style, and language "type". Furthermore pedagogical problems, text complexity, language history, etc. are studied based on sentence length. The aim of our short contribution is (1) to show that Slovak texts follow Sherman's law (cf. Altmann 1988) too, and (2) that there is – based on sentence length – a striking difference between fiction and technical texts.

Sentence length is measured in the number of clauses which are the immediate constituents of sentences. Measuring the sentence length in number of clauses seems to be a more appropriate way, since another measurement (e.g. in the number of words, morphemes etc.) implies the skipping of a language level i.e. an unnecessary assumption of the ceteris paribus condition and the additional consideration of further parameters. However, the definition and identification of clauses must be linguistically decided. The identification of the clauses in our Slovak material follows the suggestions, made by Niehaus (1997):

(a) a clause is a part of sentence containing a finite verb or a participle, e.g. *Hodiny tupo šťukali semtam a Mara Turjanka, zamýšľajúc sa* (zamýšľala sa) *nad svojím smutným stavom, zaháňala rukou myšlienky, prihládzajúc si* (prihládzala si) *vlasy, podtískajúc* (podtískala) *ich pod šatku, ale slzy jednak nemohla udržať.* (1 sentence, 6 clauses);
(b) each sentence contains at least one clause, e.g. *Dobre!*
(c) ellipses are restored and the given parts are considered clauses – a consequence of rule (b), e.g. *Ale akože neísť, keď ma vždy volajú,*

[1] Address correspondence to: zuzkave@zoznam.sk or milka@stonline.sk.

*a zarobím, aj sa vyspím dosť, aj dve pláce cez deň.* (1 sentence, 5 clauses; the last part has an elliptic verb);

(d) indirect and direct speech are not separated; e.g. *Zuzka len plecom mrdla: „Nuž nepôjdem!"* (1 sentence, 2 clauses);

(e) epentheses, separated by commas or hyphens, are counted as full parts of the sentences, e.g. *Stein zostavil až 23 rtg príznakov akútnej pankreatitídy, ktoré môžu byť priame – vyvolané zväčšením pankreasu – alebo nepriame – prevažne funkčné zmeny a odchýlky.* (1 sentence, 3 clauses);

Definitions of linguistic entities are matter of convention. They are neither true nor false, but they are either prolific or not prolific. They are prolific if the data generated by them abide by the hypothesis set up a priori. The data are not given, but they are the result of operational definitions. Even if the operational definitions of the clauses for two languages are different, the data are comparable because they arose from the interpretation of the same hypothesis. The definitions may differ but we need only the numbers of clauses.

The fact that sentence length is a factor of style is generally known. It can be characteristic for a text sort or even the individual style of an author. For quantitative linguistics three aspects are relevant: (a) frequency distribution, (b) text characterization, (c) the relation of frequency to other properties of the text and sentence and (d) its relation to the morphological character of language. Since the last two points will not be scrutinized here in detail, they nevertheless should be mentioned briefly. Concerning (c), the sentence has like any other speech entity a potentially infinite number of properties, i.e. syntactic, semantic, discourse, psycholinguistic, speech act, poetic and other properties whose relationship to its length can be analysed. Concerning (d), there are many differences: some languages use relative sentences, other ones prefer attributes, phrases and clauses may be constructed in different ways, there could be a difference in the synthetic and the analytic way of expression, etc. Considering all possible syntactic mechanisms, in two languages the same sentence can have different length. Omitting these synergetic (c) and typological problems (d) we restrict ourselves to point (a) and (b).

## *2. TEXT SAMPLES*

In order to compare different styles we analysed 10 fictional texts written by Slovak authors, mostly in reedited versions, and 10 technical texts from medicine, biology, legislation, mineralogy, geography, art and instructions. The data were obtained with pencil-and-paper method because there are no tagged Slovak texts available yet.

Following texts were analysed:

**Fictional texts:**

L1 URBAN, M.: *Živý bič*. Slovenský spisovateľ, Bratislava 1990.
L2 BEDNÁR, A.: *Sklený vrch*. Smena, Bratislava 1974.
L3 TAJOVSKÝ, J. G.: *Horký chlieb*. Horký chlieb a iné poviedky, Mladé letá, Bratislava 1960.
L4 JESENSKÝ, J.: *Vydaj*. Malomestské rozprávky, Tranoscius, Liptovský Mikuláš 1996.
L5 TIMRAVA: *Na Ondreja*. Tatran, Martin 1975.
L6 HURBAN, J. M.: *Olejkár*. Tatran, Bratislava 1977.
L7 VAJANSKÝ, S. HURBAN: *Rubačova žienka*. http://zlatyfond.sme.sk/dielo/64/Vajansky_Rubacova-zienka/1.
L8 NÁDAŠÍ – JÉGÉ, L.: *Adam Šangala*. Tatran, Bratislava 1970.
L9 CHALUPKA, J.: *Bendeguz*. Slovenské vydavateľstvo krásnej literatúry, Bratislava 1959. L10 – JAROŠ, P.: *Nemé ucho, hluché oko*. Slovenský spisovateľ, Bratislava 1984.

**Technical texts:**

T1 MORAVEC, R. a kol.: *Diabetická mikroangiopatia* Choroby pankreasu. Osveta, Mar-tin 1987.
T2 KRAMPLOVÁ, Z.: *Glykozidy*. Prírodná lekáreň, Príroda, Bratislava 1988.
T3 SEVERKA, V.: *Historický vývoj organizmov*. Všeobecná biológia (Vysokoškolské skriptá), Trnavská univerzita, Trnava 1999.
T4 SABO, M. a kol.: *Licenčná zmluva na predmety priemyselného vlastníctva*. Právne formy podnikateľských vzťahov (Vysokoškolské učebné texty), IURA EDITION, Bratislava 1995.
T5 DUBOVSKÝ, J., MARŠÁLEK, L.: *Zmeny rastlinných organizmov pri domestikácii*. Genetika rastlín, Slovenské vydavateľstvo pôdohospodárskej literatúry, Bratislava 1968.
T6 HERČKO, I.: *Kalcit*. Minerály Slovenska, Osveta, Martin 1984.

T7 BURDA a kol.: *Význam biogénnych a stopových prvkov*. Poľnohospodárska výroba. Slovenské vydavateľstvo pôdohospodárskej literatúry, Bratislava 1960.
T8 BELLA, P. – *Domica*. SLOVENSKO Sprístupnené jaskyne, DTP štúdio Grafon.
T9 KATUŠČÁK, D.: *Úprava seminárnej práce*. Ako písať záverečné a kvalifikačné práce, Enigma, Nitra 2004.
T10 Kolektív autorov: *Rozmanitosť románskeho umenia*. Dejiny umenia, Mladé letá, Bratislava 1998.

## *3. TESTING THE FIRST HYPOTHESIS*

According to Sherman-Altmann's law sentence length, measured in clause numbers in the majority of cases abides by the zero-truncated or positive negative binomial distribution given as

$$P_x = \binom{k+x-1}{x} \frac{p^k q^x}{1-p^k}, \quad x = 1,2,3,... \qquad (1)$$

whose parameters are interpreted linguistically (cf. Altmann 1988, 2005). In case that $k = 1$, one obtains the (1-displaced) geometric distribution and if $k \to \infty$, $q \to 0$ and $kq \to a$, it converges to the positive Poisson distribution with parameter $a$ which is in many cases sufficient for fitting and enables us to take into account also "short" data having few classes for fitting the negative binomial.

In some cases, e.g. in text T6, the data are too short to be tested using the chi-square criterion even using the Poisson distribution because there are no degrees of freedom. In such cases one can solve the problem with a simple "trick" namely adding to the data an $x_{max}$ +1 with frequency 0 yielding the necessary DF = 1.

The fitting of the negative binomial and the Poisson distributions to the frequencies of clauses has been performed by using appropriate software (Altmann-Fitter). The raw data and the results of fitting are presented in Table 1.

Table 1
Fitting the positive negative binomial distribution to Slovak texts
(L = prosaic text, T = technical text)

| | L1 | | L2 | | L3 | | L4 | |
|---|---|---|---|---|---|---|---|---|
| x | $f_x$ | $NP_x$ | $f_x$ | $NP_x$ | $f_x$ | $NP_x$ | $f_x$ | $NP_x$ |
| 1 | 57 | 55.34 | 33 | 34.91 | 25 | 24.84 | 83 | 78.89 |
| 2 | 50 | 45.96 | 27 | 23.35 | 29 | 26.56 | 54 | 59.27 |
| 3 | 20 | 30.43 | 15 | 14.61 | 24 | 21.80 | 33 | 35.16 |
| 4 | 26 | 17.58 | 8 | 8.83 | 7 | 15.19 | 24 | 18.08 |
| 5 | 6 | 9.27 | 4 | 5.22 | 12 | 9.45 | 6 | 8.44 |
| 6 | 5 | 4.58 | 3 | 3.04 | 6 | 5.41 | 4 | 3.68 |
| 7 | 1 | 2.15 | 1 | 1.75 | 4 | 2.91 | 2 | 2.48 |
| 8 | 1 | 0.97 | 1 | 0.99 | 2 | 2.84 | | |
| 9 | 0 | 0.42 | 0 | 0.57 | | | | |
| 10 | 0 | 0.18 | 2 | 0.73 | | | | |
| 11 | 1 | 0.13 | | | | | | |
| | k=4.1075; DF=5; p=0.6748; P=0.08 $X^2$=9.87; | | k=1.4825; DF=5; p=0.4610; P=0.90 $X^2$=1.63; | | k=5.5929; DF=5; p=0.6757; P=0.28 $X^2$ = 6.27; | | k=4.4181; DF=4; p=0.7227; P=0.47 $X^2$ = 3.58; | |

| x | L5 | | L6 | | L7 | | L8 | |
|---|---|---|---|---|---|---|---|---|
| 1 | 16 | 15.76 | 21 | 20.43 | 69 | 67.18 | 52 | 54.76 |
| 2 | 21 | 21.41 | 19 | 20.32 | 58 | 60.97 | 58 | 52.23 |
| 3 | 22 | 20.03 | 16 | 15.14 | 42 | 37.29 | 37 | 36.78 |
| 4 | 11 | 14.50 | 9 | 9.40 | 12 | 17.29 | 19 | 21.30 |
| 5 | 10 | 8.65 | 6 | 5.13 | 9 | 6.49 | 8 | 10.74 |
| 6 | 5 | 4.42 | 2 | 2.55 | 1 | 2.05 | 6 | 4.88 |
| 7 | 2 | 1.99 | 1 | 1.17 | 1 | 0.73 | 3 | 2.04 |
| 8 | 1 | 1.26 | 1 | 0.85 | | | 0 | 0.80 |
| | | | | | | | 0 | 0.30 |
| | | | | | | | 0 | 0.11 |
| | | | | | | | 1 | 0.05 |
| | k=30.2243; DF=5; p=0.9129; P=0.93 $X^2$=1.39 | | k=7.0408; DF=4; p=0.7527;P=0.98 $X^2$=0.43 | | k=89.7520; DF=3; p=0.98; P=0.31 $X^2$=3.60 | | k=8.3268; DF=5; p=0.7954; P=0.78 $X^2$=2.48 | |

| x | L9 | | L10 | | T1 | | T2 | |
|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 14.70 | 38 | 39.09 | 26 | 28.36 | 51 | 52.35 |
| 2 | 22 | 20.02 | 37 | 34.07 | 23 | 17.20 | 36 | 32.13 |
| 3 | 24 | 18.20 | 21 | 22.83 | 3 | 6.45 | 11 | 13.56 |
| 4 | 15 | 12.42 | 14 | 13.00 | | Poisson | 4 | 4.42 |
| 5 | 2 | 6.78 | 5 | 6.61 | | | 1 | 1.19 |
| 6 | 3 | 3.09 | 6 | 5.40 | | | 1 | 0.34 |
| 7 | 0 | 1.21 | | | | | | |
| 8 | 1 | 0.59 | | | | | | |
| | k=1040.4659; DF=4; p=0.9974; P=0.10 $X^2$=7.82; | | k=5.5291; DF=3; p=0.7330; P=0.81 $X^2$=0.97; | | a=0.6064; DF=1; $X^2$=4; P=0.05 | | k=30.9833; DF=2; p=0.9616; P=0.56 $X^2$=1.17; | |

| | T3 | | T4 | | T5 | | T6 | |
|---|---|---|---|---|---|---|---|---|
| 1 | 91 | 94.04 | 16 | 16.33 | 38 | 38.15 | 28 | 28.44 |
| 2 | 45 | 36.97 | 11 | 11.12 | 27 | 26.79 | 14 | 11.09 |
| 3 | 5 | 9.71 | 7 | 5.43 | 12 | 12.06 | 0 | 2.48 |
| 4 | 0 | 1.92 | 2 | 3.12 | | Poisson | | Poisson |
| 5 | 2 | 0.35 | | | | | | |
| | k=399.2687; DF=1; p=0.9980; P=0.04 $X^2$=4.16; | | k=12.2692; DF=1; p=0.8973; P=0.35; $X^2$=0.87; | | a=0.7024; DF=1; $X^2$=0.0025; P=0.96 | | a=0.3899, DF=1 $X^2$=3.24; P=0.07 | |

| | T7 | | T8 | | T9 | | T10 | |
|---|---|---|---|---|---|---|---|---|
| | 35 | 33.71 | 51 | 61.53 | 56 | 56.33 | 70 | 73.40 |
| | 11 | 13.60 | 16 | 14.60 | 26 | 24.91 | 43 | 37.58 |
| | 7 | 5.02 | 1 | 1.88 | 7 | 8.04 | 13 | 12.88 |
| | 0 | 1.77 | | Poisson | 3 | 2.72 | 2 | 4.15 |
| | 2 | 0.90 | | | | | | |
| | k=1.6864; DF=1; p=0.6996; P=0.22 $X^2$=1.50; | | a=0.2372; DF=1; $X^2$=0.55; P=0.46; | | k=9.4755; DF=1; p=0.9156; P=0.64 $X^2$=0.21; | | k=276.9131;DF=1; p=0.9963; P=0.15; $X^2$=2.05; | |

As can be seen, the fitting is not sufficient only in one case (T3). Considering the results, we may state that most probably the generation of a technical text begins with a fully random Poisson process whose traces can still be seen in some texts (T1, T5, T6, T8, T9) in which the variability is small, but also in texts T2, T3, T4, T10 in which the convergence of negative binomial to Poisson is still evident ($k \to \infty$, $q \to 0$, $kq \to a$). Hence, at least in Slovak, the negative binomial distribution

arises hand in hand with the increase of length variability (dispersion). In fiction texts only two cases (L5, L7) display an evident convergence to Poisson.

A survey of parameters and tests is presented in Table 2.

Table 2
Fitting the negative binomial and related distributions to Slovak data

| Text | Negative binomial | | Poisson | | | |
|---|---|---|---|---|---|---|
| | k | p | a | $X^2$ | DF | P |
| L1 | 4.1075 | 0.6748 | | 9.87 | 6 | 0.08 |
| L2 | 1.4825 | 0.4610 | | 1.63 | 5 | 0.90 |
| L3 | 5.5929 | 0.6757 | | 6.27 | 6 | 0.28 |
| L4 | 4.4181 | 0.7227 | | 3.58 | 4 | 0.47 |
| L5 | 30.2243 | 0.9129 | | 1.39 | 6 | 0.93 |
| L6 | 7.0408 | 0.7527 | | 0.43 | 4 | 0.98 |
| L7 | 89.7520 | 0.9800 | | 3.60 | 3 | 0.31 |
| L8 | 8.3268 | 0.7954 | | 2.48 | 6 | 0.78 |
| L9 | 1040.4659 | 0.9974 | | 7.82 | 4 | 0.10 |
| L10 | 5.5291 | 0.7330 | | 0.97 | 3 | 0.81 |
| T1 | | | 1.0740 | 6.69 | 2 | 0.03 |
| T2 | 30.9833 | 0.9616 | | 1.17 | 2 | 0.56 |
| T3 | 399.2687 | 0.9980 | | 4.16 | 1 | 0.04 |
| T4 | 12.2692 | 0.8973 | | 0.87 | 1 | 0.35 |
| T5 | | | 1.2213 | 4.81 | 2 | 0.09 |
| T6 | | | 0.3899 | 3.24 | 1 | 0.07 |
| T7 | 1.6864 | 0.6996 | | 1.60 | 1 | 0.22 |
| T8 | | | 0.5216 | 1.53 | 1 | 0.22 |
| T9 | 9.4755 | 0.9156 | | 0.21 | 1 | 0.64 |
| T10 | 276.9131 | 0.9963 | | 206 | 1 | 0.16 |

## *3. TESTING THE SECOND HYPOTHESIS*

If there is a significant difference in sentence length between two genres, it must be possible to show it using some simple statistical tests. Usually one performs a discriminant analysis (cf. Kelih et al. 2006) or sets up a taxonomy but in our case a much simpler method is sufficient. Later on, when texts from more genres will be at our disposal, a more complex analysis will be necessary. One common possibility to perform discrimination is the computing of Ord's indicators *I* and *S* (cf. Ord 1972) defined as

$$\mathrm{I} = \frac{m_2}{\bar{x}} \qquad (2)$$

and

$$\mathrm{S} = \frac{m_3}{m_2}\ , \qquad (3)$$

where the moments are defined as

$$\bar{x} = \frac{1}{N}\sum_x x f_x$$

$$m_r = \frac{1}{N}\sum_x (x-\bar{x})^r f_x \ , \mathrm{r} = 2,3$$

and obtain the results in Table 3.

In this table all means ($m_1$) and variances of fictional texts are greater than those of technical texts. A simple test for the average of means would show that there is a significant difference. Plotting the <I,S> point in a Cartesian coordinate system clearly shows that fictional and technical texts are situated in different areas of the plot (cf. Figure 1).

Though all points lie almost on the same line, all fictional texts have I > 0.61 while all technical texts have I < 0.59. One can test the difference between fictional and technical tests using the normal criterion

$$U = \frac{\bar{I}_F - \bar{I}_T}{\sqrt{V(\bar{I}_F) + V(\bar{I}_T)}}\ . \qquad (4)$$

Table 3
Ord's indicators for 20 Slovak texts

| Text | $m_1$ | $m_2$ | $m_3$ | I | S |
|---|---|---|---|---|---|
| F1 | 2.4371 | 2.5454 | 6.8958 | 1.0444 | 2.7091 |
| F2 | 2.5213 | 3.4836 | 12.6127 | 1.3817 | 3.6206 |
| F3 | 2.9633 | 3.2280 | 5.3279 | 1.0893 | 1.6505 |
| F4 | 2.2039 | 1.8031 | 2.8207 | 0.8181 | 1.5644 |
| F5 | 3.0682 | 2.6999 | 3.1974 | 0.8800 | 1.1843 |
| F6 | 2.6667 | 2.4889 | 4.0326 | 0.9333 | 1.6202 |
| F7 | 2.1719 | 1.4340 | 1.8649 | 0.6603 | 1.3005 |
| F8 | 2.5163 | 2.3910 | 6.1643 | 0.9502 | 2.5781 |
| F9 | 2.8831 | 1.7656 | 2.3739 | 0.6124 | 1.3445 |
| F10 | 2.4132 | 1.9449 | 2.5927 | 0.8059 | 1.3331 |
| T1 | 1.5577 | 0.3621 | 0.1246 | 0.2325 | 0.3441 |
| T2 | 1.7596 | 0.9326 | 1.4845 | 0.5300 | 1.5918 |
| T3 | 1.4406 | 0.4842 | 0.7640 | 0.3361 | 1.5779 |
| T4 | 1.8611 | 0.8418 | 0.5479 | 0.3361 | 0.6509 |
| T5 | 1.6623 | 0.5353 | 0.2431 | 0.3220 | 0.1937 |
| T6 | 1.3333 | 0.2222 | 0.0741 | 0.1667 | 0.3333 |
| T7 | 1.6000 | 0.9309 | 1.6538 | 0.5818 | 1.7766 |
| T8 | 1.2647 | 0.2240 | 0.1565 | 0.1771 | 0.6987 |
| T9 | 1.5326 | 0.5968 | 0.6671 | 0.3894 | 1.1178 |
| T10 | 1.5859 | 0.5395 | 0.4208 | 0.3402 | 0.7780 |

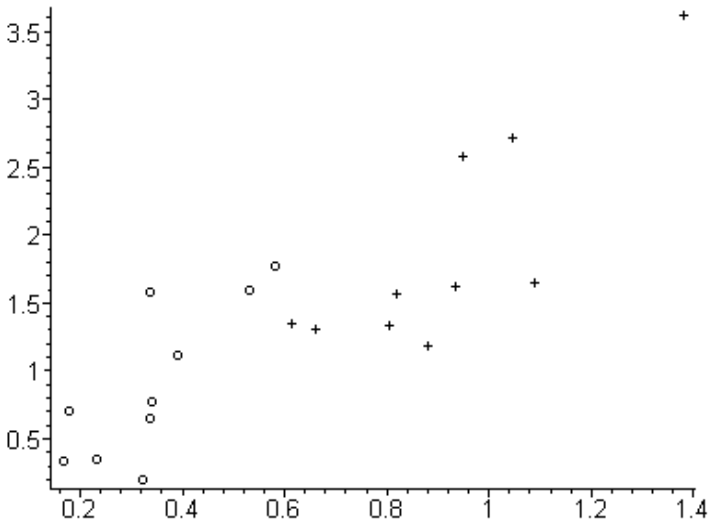



Figure 1. Ord's indicators of sentence length
(circles = technical texts; crosses = fictional texts)

Taking the values from the fifth column of Table 3 we obtain

$$\overline{I}_F = 0.91756; \quad \sigma^2_{\overline{I}_F} = 0.004453984$$

$$\overline{I}_T = 0.34199 \quad \sigma^2_{\overline{I}_T} = 0.00165473$$

Inserting these values in (4) we obtain

$$u = \frac{0.91756 - 0.34199}{\sqrt{0.004453984 + 0.00165473}} = \mathbf{7.36,}$$

which is a highly significant result. Hence, technical texts have a much smaller variation coefficient represented by Ord´s indicator I. We assume that this tendency is quite general, holding not only for Slovak. Nevertheless, the assumption must be tested in several languages.

*REFERENCES*

Altmann, G. (1988): Verteilungen der Satzlängen. *Glottometrika 9, 147-170*.

Altmann, G. (2005): Diversification processes. In: Köhler, R.; Altmann, G.; Piotrowski, R.G., *Quantitative linguistics. An international handbook: 646-658*. Berlin/New York: de Gruyter.

Best, K.-H. (2001): *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt Verlag.

Kelih, E.; Grzybek, P.; Antić, G.: Stadlober, E. (2006): Quantitative Text Typology. The Impact of Sentence Length. In: Spiliopoulou, Myra; Kruse, Rudolf; Nürnberger, Andreas; Borgelt, C.; Gaul, Wolfgang (eds.): *From Data and Information Analysis to Knowledge Engineering.* Heidelberg/ Berlin: Springer, 382-389.

Köhler, R. (1995): *Bibliography of quantitative linguistics*. Amsterdam/Philadelphia: Benjamins.

Niehaus, B. (1997): Untersuchung zur Satzlängenhäufigkeit im Deutschen. *Glottometrika 16, 213-275.*

Ord, J.K. (1972): *Families of frequency distributions*. London: Griffin.

Sherman, L.A. (1888): Some observations upon the sentence-length in English prose. *University of Nebraska Studies 1, 119-130.*

# The Well-Formedness of Two Psychoanalytic Word Categories in Portuguese Texts

*Andrew Wilson (Lancaster, United Kingdom )*

## *1 INTRODUCTION*

Freud (1900) proposed the existence of two different modes of thought, which he called respectively primary process and secondary process. Primary process thought is, by nature, sensation-oriented, concrete, and ignorant of time, space, and social institutions. It is most closely associated with early childhood, but is also hypothesized to predominate in dreams and other altered states of consciousness such as hypnosis, mystical religious experiences, and under the influence of certain psychotropic drugs. In contrast, secondary process thought – which is the normal conscious mode of cognition for adults – is logical and oriented to time, space, and society. Very similar distinctions in thought processes have also been suggested by such theoretically diverse scholars such as Lévy-Bruhl (1910/1966), Goldstein (1939), Werner (1948), Piaget (1954), Sorokin (1957), Aulagnier (1975/2001), Bucci (1997), and Kristeva (1996/1997).

As a measure of primary and secondary process thought in texts, Martindale (1975) developed the Regressive Imagery Dictionary (hereafter "RID"), a computer-readable lexicon for content analysis. Originally built for the analysis of English-language texts, the RID now exists in versions for several other languages, including Portuguese. It contains two main summary categories – primary and secondary process – as well as a few further categories, such as a set of emotion types. Each of the two main summary categories is made up of several subcategories, and, in the case of primary process, these subcategories also contain further, smaller subcategories. Table 1 shows, in English, the composition of the RID, together with some sample words for each subcategory.

The RID has demonstrated good construct validity in many studies over the past thirty years or so. That is to say that frequency patterns of primary and secondary process words, which were predicted on the basis of the theoretical literature, have been found to hold empirically when operationalized using the RID. For example, in relation to the hypothesis linking primary process thought with early childhood, it

has been found that stories composed by older children contain fewer primary process words than those composed by younger children (West, Martindale & Sutton-Smith, 1985). An increase in the use of primary process words has been strongly linked to drug-induced altered states (Martindale & Fischer, 1977; West, Martindale, Hines & Roth, 1983), and fluctuations in the frequency of primary process across the text of the bible have been shown to correlate with a five-stage theory of mystical development (West, 1991; Wilson, 2007). A study of language produced under hypnosis did not find an increase in primary process over the normal waking state, but it did find a significant decrease in secondary process (Elter-Nodvin, 2000). In accordance with psycho-analytic theory, primary process lexis has also been found to be elevated in the speech of schizophrenics (West & Martindale, 1988), in fetish fantasies (Wilson, 2002), and in the folk tales of more primitive societies (Martindale, 1976).

The present study aims to examine the validity of the RID from a different angle, i.e., that of the rank-frequency distributions of words within its two main component categories. Although research on the rank-frequency distributions of words in texts has a long history (cf. Prün, 1999; Baayen, 2001; Altmann, 2002), and is equally well established in contemporary quantitative linguistics, there has so far been relatively little research on the distribution of word frequencies within other linguistic categories that have been identified in a text. One exception to this observation is the work by Uhlířova (1995), who examined the rank-frequency distribution of word forms within word-length categories and found that they could be modelled by the well-known Zipf-Mandelbrot distribution. Another, more recent, exception is the work by Popescu, Best & Altmann (2007), who have examined the distribution of word forms within part-of-speech categories. This study has perhaps an even wider import than Uhlířova's, since it deals with interpretive categories constructed by scientists rather than with categories which derive from purely formal properties of the text (such as word lengths).

Popescu, Best & Altmann (2007) base their approach on word-frequency spectra. A word-frequency spectrum for a text is obtained when the number of different words that fall into each frequency category is listed in the natural ascending order of the frequency categories – i.e., first, the number of words which occur only once, then the number which occur twice, then the number which occur three times, and so on. Proceeding from these frequency spectra, Popescu, Best & Altmann (2007) found that the right-truncated Zeta distribution, which is the simplest and most commonly used model for the frequency spectra

of words within a text, could also be fitted to the frequency spectra of words within discrete part-of-speech categories in texts – for example, the number of nouns that occur once, the number that occur twice, and so on.[1]

This observation leads straightforwardly to the following general hypothesis about the frequencies of words within categories: "if a linguistic class is constructed "naturally", then its elements abide by a proper rank-frequency distribution of the Zipf type" (Strauss, Fan & Altmann, 2008, 94). The present study aims to test whether this hypothesis holds for the primary and secondary process categories in the RID when they are applied in the content analysis of individual texts. If the right-truncated Zeta distribution can be fitted adequately to the words within both categories, then – in addition to possessing psychoanalytic construct validity – the categories would also appear to constitute well-formed natural properties of texts.

## *2 DATA AND METHOD*

The main set of data for this study consisted of six complete books selected from a Portuguese translation of the Bible (Bíblia Católica v2.0, retrieved April 20, 2006 from http://www.bibliacatolica.com.br). The books chosen were the four gospels of Matthew, Mark, Luke and John, together with the Apocalypse (also known as the Book of Revelation) and the Epistle to the Romans. The latter two books were particularly chosen for this preliminary experiment as they were known from previous studies to have very different overall frequencies of primary and secondary process words (West, 1991; Wilson, 2007).

Since translation texts can sometimes be problematic for quantitative modeling, as they are not spontaneously composed texts in the target language, a further sample was also included. This was chosen at random (subject to availability as an electronic text) from the reading list in Portuguese literature for the University of Cambridge's degree course in Portuguese. It was the 19th century romance *A Dama-Pé-De-Cabra* by Alexandre Herculano (retrieved March 19, 2007, from http://www.gutenberg.org/1/7/0/0/17005/, as part of his *Lendas e Narrativas, Tomo II*).

[1] This is the original distribution proposed by Zipf (1935), simply truncated at the maximum word frequency for a given text. A fair amount of recent work, such as that of Uhlířová (1995), has tended to neglect this distribution in favour of other related distributions such as the Zipf-Mandelbrot distribution; however, since it has only one parameter, Zipf's original zeta distribution has the advantage of being simpler than its relatives, and this makes it much easier to interpret.

Table 1
Categories and subcategories in the RID, based on Martindale (1975)

| Major subcategory | Minor subcategory | Example words |
|---|---|---|
| ***PRIMARY PROCESS*** | | |
| Drive | | |
| | Oral | *breast, drink, lip* |
| | Anal | *sweat, rot, dirty* |
| | Sex | *lover, kiss, naked* |
| Sensation | | |
| | General sensation | *fair, charm, beauty* |
| | Touch | *touch, thick, stroke* |
| | Taste | *sweet, taste, bitter* |
| | Odor | *breath, perfume, scent* |
| | Sound | *hear, voice, sound* |
| | Vision | *see, light, look* |
| | Cold | *cold, winter, snow* |
| | Hard | *rock, stone, hard* |
| | Soft | *soft, gentle, tender* |
| Perceptual Disinhibition | | |
| | Passivity | *die, lie, bed* |
| | Voyage | *wander, desert, beyond* |
| | Random movement | *wave, roll, spread* |
| | Diffusion | *shade, shadow, cloud* |
| | Chaos | *wild, crowd, ruin* |
| Regressive Cognition | | |
| | Unknown | *secret, strange, unknown* |
| | Timeless | *eternal, forever, immortal* |
| | Altered consciousness | *dream, sleep, wake* |
| | Brink passage | *road, wall, door* |
| | Narcissism | *eye, heart, hand* |
| | Concreteness | *at, where, over* |
| Icarian Imagery | | |
| | Ascend | *rise, fly, throw* |
| | Height | *up, sky, high* |
| | Descend | *fall, drop, sink* |
| | Depth | *down, deep, beneath* |
| | Fire | *sun, fire, flame* |
| | Water | *sea, water, stream* |
| ***SECONDARY PROCESS*** | | |
| Abstract Thought | | *know, may, thought* |
| Social Behavior | | *say, tell, call* |
| Instrumental Behavior | | *make, find, work* |
| Restraint | | *must, stop, bind* |
| Order | | *simple, measure, array* |
| Temporal Reference | | *when, now, then* |
| Moral Imperative | | *should, right, virtue* |

Table 2 shows, for the seven texts, their chi-square ratios of primary to secondary process, calculated according to the formula pattern proposed by Ziegler, Best & Altmann (2002, 76): $X^2$ = (primary process – secondary process)$^2$ / (primary process + secondary process). This equation is distributed as the chi-square with one degree of freedom. A book leans significantly towards primary or secondary process content if the chi-square value is greater than 3.84 ($p < 0.05$); the direction of the trend can be determined simply from the process type with the higher frequency.

Table 2
Chi-square ratios of primary to secondary process for the seven books

| **Book** | **F(PP)** | **F(SP)** | $X^2$ | **Trend** |
|---|---|---|---|---|
| Matthew | 1657 | 1546 | 3.85 | PP > SP |
| Mark | 1078 | 863 | 23.82 | PP > SP |
| Luke | 1592 | 1608 | 0.08 | not significant |
| John | 1338 | 1276 | 1.47 | not significant |
| Romans | 596 | 768 | 21.69 | SP > PP |
| Apocalypse | 1143 | 492 | 259.21 | PP > SP |
| Herculano | 779 | 349 | 163.92 | PP > SP |

To identify the primary and secondary process words in the texts and measure their frequencies, a computerized content analysis was performed using the Portuguese version of the RID by Cardoso e Cunha, Detry, Hogenraad & Martindale (version of 28 August 1996). The RID was applied to the texts using the PROTAN suite of programs for content analysis (Hogenraad, Daubies, Bestgen & Mahau, 2003). PROTAN first divides the input file into the segments premarked by the analyst (in this case, the individual books). The majority of words in these segments are then reduced by another procedure to their basic, uninflected forms. The reduced text is finally matched against the entries in the RID. For each text segment, PROTAN produces a frequency count of each requested RID category, which shows how many word occurrences fell into that category. PROTAN can also produce a listing of each individual word form that fell into a given category and of its frequency in each text segment: this listing was used as the basis for the frequency spectrum calculations in this paper. Note that the frequencies studied in this paper are those for the actual word forms in the texts, not the reduced forms used by PROTAN for its internal dictionary matching procedure.

For each text, two frequency spectra were produced: one for the frequencies of words in the category of primary process and one for

the frequencies of words in the category of secondary process. Using the Altmann Fitter software, the right-truncated Zeta distribution was then fitted to the empirical frequencies for each spectrum. The right-truncated Zeta distribution is given by:

$$P_x = x^{-a} / F(R),\ x = 1,2,...,R$$

where $a$ = a parameter and $F(R) = \sum_{i=1}^{R} i^{-a}$ .

The probability of the chi-squared test between the empirical and estimated frequencies was used as a measure of goodness of fit. The distribution was considered to be a good model if $P(X^2) \geq 0.05$.

## *3 RESULTS*

Tables 3 to 16 show the results of fitting the right-truncated Zeta distribution to the fourteen frequency spectra. In the tables, the following abbreviations are used:

x = the frequency class (e.g. 1 = words with a frequency of 1)
g(x) = the number of words falling into frequency class x in the observed data
$NP_x$ = the estimated value of g(x) generated by the right-truncated Zeta distribution
a = the parameter of the right-truncated Zeta distribution
R = the value of x at which the right-truncated Zeta distribution is truncated
DF = the number of degrees of freedom of the chi-squared test
$X^2$ = value of the chi-squared test
$P(X^2)$ = probability of the chi-squared test

It will be seen that the right-truncated Zeta distribution could be fitted successfully in all cases. In the majority of cases, the value of $P(X^2)$ was very high, demonstrating an excellent quality of fit. In three cases – Matthew (primary process), Mark (primary process), and Apocalypse (secondary process) – $P(X^2)$ was rather low, but even here the quality of fit for the theoretical model was still acceptable.

Table 3: Matthew. Primary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 275 | 265.87 | 26 | 1 | 0.47 | 51 | 0 | 0.13 | 76 | 0 | 0.06 | 101 | 0 | 0.03 | 126 | 0 | 0.02 | 151 | 0 | 0.02 |
| 2 | 61 | 69.09 | 27 | 1 | 0.44 | 52 | 0 | 0.12 | 77 | 0 | 0.06 | 102 | 0 | 0.03 | 127 | 0 | 0.02 | 152 | 0 | 0.02 |
| 3 | 23 | 31.41 | 28 | 0 | 0.41 | 53 | 0 | 0.12 | 78 | 0 | 0.06 | 103 | 0 | 0.03 | 128 | 0 | 0.02 | 153 | 0 | 0.02 |
| 4 | 21 | 17.95 | 29 | 1 | 0.38 | 54 | 1 | 0.11 | 79 | 0 | 0.05 | 104 | 0 | 0.03 | 129 | 0 | 0.02 | 154 | 0 | 0.01 |
| 5 | 10 | 11.63 | 30 | 0 | 0.36 | 55 | 0 | 0.11 | 80 | 0 | 0.05 | 105 | 0 | 0.03 | 130 | 0 | 0.02 | 155 | 0 | 0.01 |
| 6 | 7 | 8.16 | 31 | 0 | 0.34 | 56 | 0 | 0.11 | 81 | 1 | 0.05 | 106 | 0 | 0.03 | 131 | 0 | 0.02 | 156 | 0 | 0.01 |
| 7 | 7 | 6.05 | 32 | 0 | 0.32 | 57 | 0 | 0.1 | 82 | 0 | 0.05 | 107 | 0 | 0.03 | 132 | 0 | 0.02 | 157 | 0 | 0.01 |
| 8 | 6 | 4.67 | 33 | 0 | 0.3 | 58 | 0 | 0.1 | 83 | 0 | 0.05 | 108 | 0 | 0.03 | 133 | 0 | 0.02 | 158 | 0 | 0.01 |
| 9 | 2 | 3.71 | 34 | 0 | 0.28 | 59 | 0 | 0.1 | 84 | 0 | 0.05 | 109 | 0 | 0.03 | 134 | 0 | 0.02 | 159 | 0 | 0.01 |
| 10 | 6 | 3.02 | 35 | 1 | 0.26 | 60 | 0 | 0.09 | 85 | 0 | 0.05 | 110 | 0 | 0.03 | 135 | 0 | 0.02 | 160 | 0 | 0.01 |
| 11 | 8 | 2.51 | 36 | 0 | 0.25 | 61 | 0 | 0.09 | 86 | 0 | 0.05 | 111 | 0 | 0.03 | 136 | 0 | 0.02 | 161 | 0 | 0.01 |
| 12 | 4 | 2.12 | 37 | 0 | 0.24 | 62 | 0 | 0.09 | 87 | 0 | 0.05 | 112 | 0 | 0.03 | 137 | 0 | 0.02 | 162 | 0 | 0.01 |
| 13 | 4 | 1.82 | 38 | 0 | 0.23 | 63 | 0 | 0.08 | 88 | 0 | 0.04 | 113 | 0 | 0.03 | 138 | 0 | 0.02 | 163 | 0 | 0.01 |
| 14 | 0 | 1.57 | 39 | 0 | 0.21 | 64 | 0 | 0.08 | 89 | 0 | 0.04 | 114 | 0 | 0.03 | 139 | 0 | 0.02 | 164 | 0 | 0.01 |
| 15 | 1 | 1.37 | 40 | 0 | 0.2 | 65 | 0 | 0.08 | 90 | 0 | 0.04 | 115 | 0 | 0.03 | 140 | 0 | 0.02 | 165 | 0 | 0.01 |
| 16 | 2 | 1.21 | 41 | 0 | 0.19 | 66 | 0 | 0.08 | 91 | 0 | 0.04 | 116 | 0 | 0.03 | 141 | 0 | 0.02 | 166 | 0 | 0.01 |
| 17 | 2 | 1.08 | 42 | 0 | 0.19 | 67 | 0 | 0.07 | 92 | 0 | 0.04 | 117 | 0 | 0.03 | 142 | 0 | 0.02 | 167 | 0 | 0.01 |
| 18 | 0 | 0.96 | 43 | 0 | 0.18 | 68 | 0 | 0.07 | 93 | 0 | 0.04 | 118 | 0 | 0.02 | 143 | 0 | 0.02 | 168 | 0 | 0.01 |
| 19 | 0 | 0.87 | 44 | 0 | 0.17 | 69 | 0 | 0.07 | 94 | 0 | 0.04 | 119 | 0 | 0.02 | 144 | 0 | 0.02 | 169 | 0 | 0.01 |
| 20 | 1 | 0.79 | 45 | 0 | 0.16 | 70 | 0 | 0.07 | 95 | 0 | 0.04 | 120 | 0 | 0.02 | 145 | 0 | 0.02 | 170 | 0 | 0.01 |
| 21 | 0 | 0.71 | 46 | 0 | 0.16 | 71 | 0 | 0.07 | 96 | 0 | 0.04 | 121 | 0 | 0.02 | 146 | 0 | 0.02 | 171 | 0 | 0.01 |
| 22 | 0 | 0.65 | 47 | 0 | 0.15 | 72 | 0 | 0.07 | 97 | 0 | 0.04 | 122 | 0 | 0.02 | 147 | 0 | 0.02 | 172 | 0 | 0.01 |
| 23 | 0 | 0.6 | 48 | 0 | 0.14 | 73 | 1 | 0.06 | 98 | 0 | 0.04 | 123 | 0 | 0.02 | 148 | 0 | 0.02 | 173 | 0 | 0.01 |
| 24 | 0 | 0.55 | 49 | 0 | 0.14 | 74 | 0 | 0.06 | 99 | 0 | 0.04 | 124 | 0 | 0.02 | 149 | 0 | 0.02 | 174 | 0 | 0.01 |
| 25 | 2 | 0.51 | 50 | 0 | 0.13 | 75 | 0 | 0.06 | 100 | 0 | 0.03 | 125 | 0 | 0.02 | 150 | 0 | 0.02 | 175 | 0 | 0.01 |
| | **a = 1.94** | | | **R = 176** | | | $X^2$ **= 37.71** | | | | **DF = 28** | | | **P($X^2$) = 0.10** | | | | | 176 | 1 | 0.01 |

Table 4: Matthew. Secondary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 276 | 271.5 | 26 | 0 | 0.39 | 51 | 0 | 0.1 | 76 | 0 | 0.05 | 101 | 0 | 0 | 126 | 0 | 0.02 | 151 | 0 | 0 |
| 2 | 64 | 67.49 | 27 | 1 | 0.36 | 52 | 0 | 0.1 | 77 | 0 | 0.04 | 102 | 0 | 0 | 127 | 0 | 0.02 | 152 | 0 | 0 |
| 3 | 28 | 29.89 | 28 | 0 | 0.34 | 53 | 1 | 0.1 | 78 | 0 | 0.04 | 103 | 0 | 0 | 128 | 0 | 0.02 | 153 | 0 | 0 |
| 4 | 21 | 16.77 | 29 | 1 | 0.31 | 54 | 0 | 0.1 | 79 | 0 | 0.04 | 104 | 0 | 0 | 129 | 0 | 0.02 | 154 | 0 | 0 |
| 5 | 9 | 10.72 | 30 | 1 | 0.29 | 55 | 1 | 0.1 | 80 | 0 | 0.04 | 105 | 0 | 0 | 130 | 0 | 0.02 | 155 | 0 | 0 |
| 6 | 9 | 7.43 | 31 | 0 | 0.27 | 56 | 0 | 0.1 | 81 | 0 | 0.04 | 106 | 0 | 0 | 131 | 0 | 0.02 | 156 | 0 | 0 |
| 7 | 4 | 5.45 | 32 | 0 | 0.26 | 57 | 0 | 0.1 | 82 | 0 | 0.04 | 107 | 0 | 0 | 132 | 0 | 0.02 | 157 | 0 | 0 |
| 8 | 5 | 4.17 | 33 | 0 | 0.24 | 58 | 0 | 0.1 | 83 | 0 | 0.04 | 108 | 0 | 0 | 133 | 0 | 0.01 | 158 | 0 | 0 |
| 9 | 2 | 3.29 | 34 | 0 | 0.23 | 59 | 1 | 0.1 | 84 | 0 | 0.04 | 109 | 0 | 0 | 134 | 0 | 0.01 | 159 | 0 | 0 |
| 10 | 2 | 2.66 | 35 | 1 | 0.22 | 60 | 0 | 0.1 | 85 | 0 | 0.04 | 110 | 0 | 0 | 135 | 0 | 0.01 | 160 | 0 | 0 |
| 11 | 2 | 2.2 | 36 | 0 | 0.2 | 61 | 0 | 0.1 | 86 | 0 | 0.04 | 111 | 0 | 0 | 136 | 0 | 0.01 | 161 | 0 | 0 |
| 12 | 3 | 1.85 | 37 | 0 | 0.19 | 62 | 0 | 0.1 | 87 | 0 | 0.03 | 112 | 0 | 0 | 137 | 0 | 0.01 | 162 | 0 | 0 |
| 13 | 0 | 1.57 | 38 | 0 | 0.18 | 63 | 0 | 0.1 | 88 | 0 | 0.03 | 113 | 0 | 0 | 138 | 0 | 0.01 | 163 | 0 | 0 |
| 14 | 0 | 1.36 | 39 | 0 | 0.17 | 64 | 0 | 0.1 | 89 | 0 | 0.03 | 114 | 0 | 0 | 139 | 0 | 0.01 | 164 | 0 | 0 |
| 15 | 2 | 1.18 | 40 | 1 | 0.16 | 65 | 0 | 0.1 | 90 | 0 | 0.03 | 115 | 0 | 0 | 140 | 0 | 0.01 | 165 | 0 | 0 |
| 16 | 0 | 1.04 | 41 | 0 | 0.16 | 66 | 0 | 0.1 | 91 | 0 | 0.03 | 116 | 0 | 0 | 141 | 0 | 0.01 | 166 | 0 | 0 |
| 17 | 2 | 0.92 | 42 | 0 | 0.15 | 67 | 0 | 0.1 | 92 | 0 | 0.03 | 117 | 0 | 0 | 142 | 0 | 0.01 | 167 | 1 | 0 |
| 18 | 1 | 0.82 | 43 | 0 | 0.14 | 68 | 0 | 0.1 | 93 | 0 | 0.03 | 118 | 0 | 0 | 143 | 0 | 0.01 | | | |
| 19 | 1 | 0.73 | 44 | 0 | 0.14 | 69 | 0 | 0.1 | 94 | 0 | 0.03 | 119 | 0 | 0 | 144 | 0 | 0.01 | | | |
| 20 | 1 | 0.66 | 45 | 0 | 0.13 | 70 | 0 | 0.1 | 95 | 0 | 0.03 | 120 | 0 | 0 | 145 | 0 | 0.01 | | | |
| 21 | 0 | 0.6 | 46 | 0 | 0.12 | 71 | 0 | 0.1 | 96 | 0 | 0.03 | 121 | 0 | 0 | 146 | 0 | 0.01 | | | |
| 22 | 1 | 0.55 | 47 | 0 | 0.12 | 72 | 0 | 0.1 | 97 | 0 | 0.03 | 122 | 0 | 0 | 147 | 0 | 0.01 | | | |
| 23 | 0 | 0.5 | 48 | 0 | 0.11 | 73 | 1 | 0.1 | 98 | 0 | 0.03 | 123 | 0 | 0 | 148 | 0 | 0.01 | | | |
| 24 | 0 | 0.46 | 49 | 0 | 0.11 | 74 | 0 | 0.1 | 99 | 0 | 0.03 | 124 | 0 | 0 | 149 | 0 | 0.01 | | | |
| 25 | 0 | 0.42 | 50 | 0 | 0.11 | 75 | 0 | 0.1 | 100 | 0 | 0.03 | 125 | 0 | 0 | 150 | 0 | 0.01 | | | |
| | **a = 2.01** | | | **R = 167** | | | **$X^2$ = 10.62** | | | | **DF = 25** | | | **P($X^2$) = 0.99** | | | | | | |

Table 5: Mark. Primary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 202 | 193 | 21 | 1 | 0.6 | 41 | 0 | 0.2 | 61 | 0 | 0.1 | 81 | 0 | 0 | 101 | 0 | 0 |
| 2 | 46 | 51.1 | 22 | 2 | 0.5 | 42 | 0 | 0.2 | 62 | 0 | 0.1 | 82 | 0 | 0 | 102 | 0 | 0 |
| 3 | 19 | 23.5 | 23 | 0 | 0.5 | 43 | 0 | 0.1 | 63 | 0 | 0.1 | 83 | 0 | 0 | 103 | 0 | 0 |
| 4 | 13 | 13.6 | 24 | 0 | 0.4 | 44 | 0 | 0.1 | 64 | 0 | 0.1 | 84 | 0 | 0 | 104 | 0 | 0 |
| 5 | 8 | 8.84 | 25 | 0 | 0.4 | 45 | 1 | 0.1 | 65 | 0 | 0.1 | 85 | 0 | 0 | 105 | 0 | 0 |
| 6 | 7 | 6.24 | 26 | 0 | 0.4 | 46 | 0 | 0.1 | 66 | 0 | 0.1 | 86 | 0 | 0 | 106 | 0 | 0 |
| 7 | 5 | 4.64 | 27 | 0 | 0.4 | 47 | 0 | 0.1 | 67 | 0 | 0.1 | 87 | 0 | 0 | 107 | 0 | 0 |
| 8 | 4 | 3.6 | 28 | 0 | 0.3 | 48 | 0 | 0.1 | 68 | 0 | 0.1 | 88 | 0 | 0 | 108 | 0 | 0 |
| 9 | 8 | 2.87 | 29 | 1 | 0.3 | 49 | 0 | 0.1 | 69 | 0 | 0.1 | 89 | 0 | 0 | 109 | 0 | 0 |
| 10 | 3 | 2.35 | 30 | 0 | 0.3 | 50 | 0 | 0.1 | 70 | 0 | 0.1 | 90 | 0 | 0 | 110 | 0 | 0 |
| 11 | 2 | 1.96 | 31 | 0 | 0.3 | 51 | 0 | 0.1 | 71 | 0 | 0.1 | 91 | 0 | 0 | 111 | 0 | 0 |
| 12 | 0 | 1.66 | 32 | 0 | 0.3 | 52 | 0 | 0.1 | 72 | 0 | 0.1 | 92 | 0 | 0 | 112 | 0 | 0 |
| 13 | 0 | 1.42 | 33 | 0 | 0.2 | 53 | 0 | 0.1 | 73 | 0 | 0.1 | 93 | 0 | 0 | 113 | 0 | 0 |
| 14 | 0 | 1.23 | 34 | 1 | 0.2 | 54 | 0 | 0.1 | 74 | 0 | 0.1 | 94 | 0 | 0 | 114 | 0 | 0 |
| 15 | 4 | 1.08 | 35 | 0 | 0.2 | 55 | 0 | 0.1 | 75 | 0 | 0.1 | 95 | 0 | 0 | 115 | 0 | 0 |
| 16 | 1 | 0.95 | 36 | 0 | 0.2 | 56 | 0 | 0.1 | 76 | 0 | 0.1 | 96 | 0 | 0 | 116 | 1 | 0 |
| 17 | 1 | 0.85 | 37 | 0 | 0.2 | 57 | 0 | 0.1 | 77 | 0 | 0.1 | 97 | 0 | 0 | | | |
| 18 | 0 | 0.76 | 38 | 0 | 0.2 | 58 | 0 | 0.1 | 78 | 0 | 0.1 | 98 | 0 | 0 | | | |
| 19 | 0 | 0.69 | 39 | 0 | 0.2 | 59 | 0 | 0.1 | 79 | 0 | 0 | 99 | 0 | 0 | | | |
| 20 | 1 | 0.62 | 40 | 0 | 0.2 | 60 | 0 | 0.1 | 80 | 0 | 0 | 100 | 0 | 0 | | | |
| | **a = 1.91** | | | **R = 116** | | | **$X^2$ = 26.23** | | | | **DF = 23** | | | **P($X^2$) = 0.29** | | | |

Table 6: Mark. Secondary Process

| **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 205 | 201 | 13 | 1 | 0.9 | 25 | 0 | 0.2 | 37 | 0 | 0.1 | 49 | 0 | 0.1 | 61 | 0 | 0 | 73 | 0 | 0 |
| 2 | 48 | 46.4 | 14 | 0 | 0.8 | 26 | 0 | 0.2 | 38 | 0 | 0.1 | 50 | 0 | 0.1 | 62 | 0 | 0 | 74 | 0 | 0 |
| 3 | 16 | 19.7 | 15 | 1 | 0.7 | 27 | 0 | 0.2 | 39 | 0 | 0.1 | 51 | 0 | 0.1 | 63 | 0 | 0 | 75 | 0 | 0 |
| 4 | 9 | 10.7 | 16 | 1 | 0.6 | 28 | 0 | 0.2 | 40 | 1 | 0.1 | 52 | 0 | 0.1 | 64 | 0 | 0 | 76 | 0 | 0 |
| 5 | 5 | 6.66 | 17 | 2 | 0.5 | 29 | 0 | 0.2 | 41 | 0 | 0.1 | 53 | 0 | 0 | 65 | 0 | 0 | 77 | 0 | 0 |
| 6 | 4 | 4.53 | 18 | 0 | 0.4 | 30 | 0 | 0.2 | 42 | 0 | 0.1 | 54 | 0 | 0 | 66 | 0 | 0 | 78 | 0 | 0 |
| 7 | 4 | 3.27 | 19 | 3 | 0.4 | 31 | 0 | 0.1 | 43 | 0 | 0.1 | 55 | 0 | 0 | 67 | 0 | 0 | 79 | 0 | 0 |
| 8 | 2 | 2.46 | 20 | 0 | 0.4 | 32 | 1 | 0.1 | 44 | 0 | 0.1 | 56 | 0 | 0 | 68 | 0 | 0 | 80 | 0 | 0 |
| 9 | 0 | 1.92 | 21 | 0 | 0.3 | 33 | 1 | 0.1 | 45 | 0 | 0.1 | 57 | 0 | 0 | 69 | 0 | 0 | 81 | 0 | 0 |
| 10 | 3 | 1.54 | 22 | 1 | 0.3 | 34 | 0 | 0.1 | 46 | 0 | 0.1 | 58 | 0 | 0 | 70 | 0 | 0 | 82 | 1 | 0 |
| 11 | 1 | 1.25 | 23 | 0 | 0.3 | 35 | 0 | 0.1 | 47 | 0 | 0.1 | 59 | 0 | 0 | 71 | 0 | 0 | | | |
| 12 | 0 | 1.04 | 24 | 0 | 0.2 | 36 | 0 | 0.1 | 48 | 0 | 0.1 | 60 | 0 | 0 | 72 | 0 | 0 | | | |
| | **a = 2.12** | | | **R = 82** | | | ***X²* = 11.32** | | | | **DF = 16** | | | **P(*X²*) = 0.79** | | | | | | |

Table 7: Luke. Primary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 271 | 262.9 | 31 | 0 | 0.33 | 61 | 0 | 0.09 | 91 | 0 | 0.04 | 121 | 0 | 0.02 | 151 | 0 | 0.02 | 181 | 0 | 0 |
| 2 | 58 | 68.29 | 32 | 0 | 0.31 | 62 | 0 | 0.09 | 92 | 0 | 0.04 | 122 | 0 | 0.02 | 152 | 0 | 0.02 | 182 | 0 | 0 |
| 3 | 28 | 31.04 | 33 | 0 | 0.29 | 63 | 0 | 0.08 | 93 | 0 | 0.04 | 123 | 0 | 0.02 | 153 | 0 | 0.01 | 183 | 0 | 0 |
| 4 | 18 | 17.74 | 34 | 0 | 0.28 | 64 | 0 | 0.08 | 94 | 0 | 0.04 | 124 | 0 | 0.02 | 154 | 0 | 0.01 | 184 | 0 | 0 |
| 5 | 17 | 11.49 | 35 | 1 | 0.26 | 65 | 0 | 0.08 | 95 | 0 | 0.04 | 125 | 0 | 0.02 | 155 | 0 | 0.01 | 185 | 0 | 0 |
| 6 | 6 | 8.06 | 36 | 0 | 0.25 | 66 | 0 | 0.08 | 96 | 0 | 0.04 | 126 | 0 | 0.02 | 156 | 0 | 0.01 | 186 | 0 | 0 |
| 7 | 6 | 5.97 | 37 | 0 | 0.23 | 67 | 0 | 0.07 | 97 | 0 | 0.04 | 127 | 0 | 0.02 | 157 | 0 | 0.01 | 187 | 0 | 0 |
| 8 | 5 | 4.61 | 38 | 0 | 0.22 | 68 | 0 | 0.07 | 98 | 0 | 0.04 | 128 | 0 | 0.02 | 158 | 0 | 0.01 | 188 | 0 | 0 |
| 9 | 6 | 3.66 | 39 | 0 | 0.21 | 69 | 0 | 0.07 | 99 | 0 | 0.03 | 129 | 0 | 0.02 | 159 | 0 | 0.01 | 189 | 0 | 0 |
| 10 | 4 | 2.99 | 40 | 0 | 0.2 | 70 | 0 | 0.07 | 100 | 0 | 0.03 | 130 | 0 | 0.02 | 160 | 0 | 0.01 | 190 | 0 | 0 |
| 11 | 1 | 2.48 | 41 | 0 | 0.19 | 71 | 0 | 0.07 | 101 | 0 | 0.03 | 131 | 0 | 0.02 | 161 | 0 | 0.01 | 191 | 0 | 0 |
| 12 | 5 | 2.09 | 42 | 0 | 0.18 | 72 | 0 | 0.06 | 102 | 0 | 0.03 | 132 | 0 | 0.02 | 162 | 0 | 0.01 | 192 | 0 | 0 |
| 13 | 2 | 1.79 | 43 | 0 | 0.18 | 73 | 0 | 0.06 | 103 | 0 | 0.03 | 133 | 0 | 0.02 | 163 | 0 | 0.01 | 193 | 0 | 0 |
| 14 | 1 | 1.55 | 44 | 0 | 0.17 | 74 | 0 | 0.06 | 104 | 0 | 0.03 | 134 | 0 | 0.02 | 164 | 0 | 0.01 | 194 | 0 | 0 |
| 15 | 1 | 1.36 | 45 | 0 | 0.16 | 75 | 0 | 0.06 | 105 | 0 | 0.03 | 135 | 0 | 0.02 | 165 | 0 | 0.01 | 195 | 0 | 0 |
| 16 | 1 | 1.2 | 46 | 0 | 0.15 | 76 | 0 | 0.06 | 106 | 0 | 0.03 | 136 | 0 | 0.02 | 166 | 0 | 0.01 | 196 | 1 | 0 |
| 17 | 2 | 1.06 | 47 | 0 | 0.15 | 77 | 0 | 0.06 | 107 | 0 | 0.03 | 137 | 0 | 0.02 | 167 | 0 | 0.01 | | | |
| 18 | 1 | 0.95 | 48 | 0 | 0.14 | 78 | 1 | 0.06 | 108 | 0 | 0.03 | 138 | 0 | 0.02 | 168 | 0 | 0.01 | | | |
| 19 | 0 | 0.86 | 49 | 0 | 0.14 | 79 | 0 | 0.05 | 109 | 0 | 0.03 | 139 | 0 | 0.02 | 169 | 0 | 0.01 | | | |
| 20 | 1 | 0.78 | 50 | 0 | 0.13 | 80 | 0 | 0.05 | 110 | 0 | 0.03 | 140 | 0 | 0.02 | 170 | 0 | 0.01 | | | |
| 21 | 1 | 0.71 | 51 | 0 | 0.13 | 81 | 0 | 0.05 | 111 | 0 | 0.03 | 141 | 0 | 0.02 | 171 | 0 | 0.01 | | | |
| 22 | 0 | 0.64 | 52 | 0 | 0.12 | 82 | 0 | 0.05 | 112 | 0 | 0.03 | 142 | 0 | 0.02 | 172 | 0 | 0.01 | | | |
| 23 | 2 | 0.59 | 53 | 0 | 0.12 | 83 | 0 | 0.05 | 113 | 0 | 0.03 | 143 | 0 | 0.02 | 173 | 0 | 0.01 | | | |
| 24 | 0 | 0.54 | 54 | 0 | 0.11 | 84 | 0 | 0.05 | 114 | 0 | 0.03 | 144 | 0 | 0.02 | 174 | 0 | 0.01 | | | |
| 25 | 2 | 0.5 | 55 | 0 | 0.11 | 85 | 0 | 0.05 | 115 | 0 | 0.03 | 145 | 0 | 0.02 | 175 | 0 | 0.01 | | | |
| 26 | 1 | 0.47 | 56 | 0 | 0.1 | 86 | 0 | 0.05 | 116 | 0 | 0.03 | 146 | 0 | 0.02 | 176 | 0 | 0.01 | | | |
| 27 | 0 | 0.43 | 57 | 1 | 0.1 | 87 | 0 | 0.04 | 117 | 0 | 0.03 | 147 | 0 | 0.02 | 177 | 0 | 0.01 | | | |
| 28 | 0 | 0.4 | 58 | 0 | 0.1 | 88 | 0 | 0.04 | 118 | 0 | 0.02 | 148 | 0 | 0.02 | 178 | 0 | 0.01 | | | |
| 29 | 1 | 0.38 | 59 | 0 | 0.09 | 89 | 0 | 0.04 | 119 | 0 | 0.02 | 149 | 0 | 0.02 | 179 | 0 | 0.01 | | | |
| 30 | 0 | 0.35 | 60 | 0 | 0.09 | 90 | 0 | 0.04 | 120 | 0 | 0.02 | 150 | 0 | 0.02 | 180 | 0 | 0.01 | | | |
| | **a = 1.94** | | | **R = 196** | | | **$X^2$ = 19.56** | | | | **DF = 28** | | | **P($X^2$) = 0.88** | | | | | | |

Table 8: Luke. Secondary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 329 | 325 | 31 | 2 | 0.2 | 61 | 1 | 0.1 | 91 | 0 | 0 | 121 | 0 | 0 | 151 | 0 | 0 |
| 2 | 72 | 73.6 | 32 | 0 | 0.2 | 62 | 0 | 0.1 | 92 | 0 | 0 | 122 | 0 | 0 | 152 | 0 | 0 |
| 3 | 29 | 30.9 | 33 | 0 | 0.2 | 63 | 0 | 0.1 | 93 | 0 | 0 | 123 | 0 | 0 | 153 | 0 | 0 |
| 4 | 18 | 16.7 | 34 | 0 | 0.2 | 64 | 1 | 0 | 94 | 0 | 0 | 124 | 0 | 0 | 154 | 0 | 0 |
| 5 | 8 | 10.3 | 35 | 0 | 0.2 | 65 | 0 | 0 | 95 | 0 | 0 | 125 | 0 | 0 | 155 | 0 | 0 |
| 6 | 6 | 7 | 36 | 1 | 0.2 | 66 | 0 | 0 | 96 | 0 | 0 | 126 | 0 | 0 | 156 | 0 | 0 |
| 7 | 3 | 5.03 | 37 | 0 | 0.1 | 67 | 0 | 0 | 97 | 0 | 0 | 127 | 0 | 0 | 157 | 0 | 0 |
| 8 | 2 | 3.78 | 38 | 0 | 0.1 | 68 | 0 | 0 | 98 | 0 | 0 | 128 | 0 | 0 | 158 | 0 | 0 |
| 9 | 3 | 2.94 | 39 | 0 | 0.1 | 69 | 0 | 0 | 99 | 0 | 0 | 129 | 0 | 0 | 159 | 0 | 0 |
| 10 | 3 | 2.34 | 40 | 0 | 0.1 | 70 | 0 | 0 | 100 | 0 | 0 | 130 | 0 | 0 | 160 | 0 | 0 |
| 11 | 2 | 1.91 | 41 | 0 | 0.1 | 71 | 0 | 0 | 101 | 0 | 0 | 131 | 0 | 0 | 161 | 0 | 0 |
| 12 | 1 | 1.59 | 42 | 0 | 0.1 | 72 | 0 | 0 | 102 | 0 | 0 | 132 | 0 | 0 | 162 | 0 | 0 |
| 13 | 3 | 1.34 | 43 | 0 | 0.1 | 73 | 0 | 0 | 103 | 0 | 0 | 133 | 0 | 0 | 163 | 0 | 0 |
| 14 | 0 | 1.14 | 44 | 0 | 0.1 | 74 | 0 | 0 | 104 | 0 | 0 | 134 | 0 | 0 | 164 | 0 | 0 |
| 15 | 2 | 0.98 | 45 | 2 | 0.1 | 75 | 0 | 0 | 105 | 0 | 0 | 135 | 0 | 0 | 165 | 0 | 0 |
| 16 | 3 | 0.86 | 46 | 0 | 0.1 | 76 | 0 | 0 | 106 | 0 | 0 | 136 | 0 | 0 | 166 | 0 | 0 |
| 17 | 0 | 0.75 | 47 | 0 | 0.1 | 77 | 0 | 0 | 107 | 0 | 0 | 137 | 0 | 0 | 167 | 0 | 0 |
| 18 | 0 | 0.67 | 48 | 0 | 0.1 | 78 | 0 | 0 | 108 | 0 | 0 | 138 | 0 | 0 | 168 | 0 | 0 |
| 19 | 1 | 0.59 | 49 | 0 | 0.1 | 79 | 0 | 0 | 109 | 0 | 0 | 139 | 0 | 0 | 169 | 0 | 0 |
| 20 | 0 | 0.53 | 50 | 0 | 0.1 | 80 | 0 | 0 | 110 | 0 | 0 | 140 | 0 | 0 | 170 | 0 | 0 |
| 21 | 0 | 0.48 | 51 | 0 | 0.1 | 81 | 0 | 0 | 111 | 0 | 0 | 141 | 0 | 0 | 171 | 0 | 0 |
| 22 | 0 | 0.43 | 52 | 0 | 0.1 | 82 | 0 | 0 | 112 | 0 | 0 | 142 | 0 | 0 | 172 | 0 | 0 |
| 23 | 0 | 0.39 | 53 | 0 | 0.1 | 83 | 0 | 0 | 113 | 0 | 0 | 143 | 0 | 0 | 173 | 0 | 0 |
| 24 | 0 | 0.36 | 54 | 0 | 0.1 | 84 | 0 | 0 | 114 | 0 | 0 | 144 | 0 | 0 | 174 | 0 | 0 |
| 25 | 0 | 0.33 | 55 | 0 | 0.1 | 85 | 0 | 0 | 115 | 0 | 0 | 145 | 0 | 0 | 175 | 0 | 0 |
| 26 | 0 | 0.3 | 56 | 0 | 0.1 | 86 | 0 | 0 | 116 | 0 | 0 | 146 | 0 | 0 | 176 | 0 | 0 |
| 27 | 2 | 0.28 | 57 | 0 | 0.1 | 87 | 0 | 0 | 117 | 0 | 0 | 147 | 0 | 0 | 177 | 0 | 0 |
| 28 | 0 | 0.26 | 58 | 0 | 0.1 | 88 | 0 | 0 | 118 | 0 | 0 | 148 | 0 | 0 | 178 | 0 | 0 |
| 29 | 0 | 0.24 | 59 | 0 | 0.1 | 89 | 0 | 0 | 119 | 0 | 0 | 149 | 0 | 0 | 179 | 1 | 0 |
| 30 | 0 | 0.22 | 60 | 0 | 0.1 | 90 | 1 | 0 | 120 | 0 | 0 | 150 | 0 | 0 | | | |
| | **a = 2.14** | | | **R = 179** | | | **$X^2$ = 12.49** | | | | | **DF = 21** | | | **P($X^2$) = 0.92** | | |

Table 9: John. Primary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 146 | 151 | 32 | 0 | 0.3 | 63 | 0 | 0.1 | 94 | 0 | 0 | 125 | 0 | 0 | 156 | 0 | 0 | 187 | 0 | 0 |
| 2 | 47 | 42.8 | 33 | 0 | 0.3 | 64 | 0 | 0.1 | 95 | 0 | 0 | 126 | 0 | 0 | 157 | 0 | 0 | 188 | 0 | 0 |
| 3 | 19 | 20.5 | 34 | 2 | 0.3 | 65 | 0 | 0.1 | 96 | 0 | 0 | 127 | 0 | 0 | 158 | 0 | 0 | 189 | 0 | 0 |
| 4 | 12 | 12.2 | 35 | 0 | 0.2 | 66 | 0 | 0.1 | 97 | 0 | 0 | 128 | 0 | 0 | 159 | 0 | 0 | 190 | 0 | 0 |
| 5 | 7 | 8.11 | 36 | 0 | 0.2 | 67 | 0 | 0.1 | 98 | 0 | 0 | 129 | 0 | 0 | 160 | 0 | 0 | 191 | 0 | 0 |
| 6 | 6 | 5.82 | 37 | 0 | 0.2 | 68 | 0 | 0.1 | 99 | 0 | 0 | 130 | 0 | 0 | 161 | 0 | 0 | 192 | 0 | 0 |
| 7 | 8 | 4.4 | 38 | 0 | 0.2 | 69 | 0 | 0.1 | 100 | 0 | 0 | 131 | 0 | 0 | 162 | 0 | 0 | 193 | 0 | 0 |
| 8 | 2 | 3.45 | 39 | 0 | 0.2 | 70 | 0 | 0.1 | 101 | 0 | 0 | 132 | 0 | 0 | 163 | 0 | 0 | 194 | 0 | 0 |
| 9 | 3 | 2.79 | 40 | 0 | 0.2 | 71 | 1 | 0.1 | 102 | 0 | 0 | 133 | 0 | 0 | 164 | 0 | 0 | 195 | 0 | 0 |
| 10 | 2 | 2.3 | 41 | 0 | 0.2 | 72 | 1 | 0.1 | 103 | 0 | 0 | 134 | 0 | 0 | 165 | 0 | 0 | 196 | 0 | 0 |
| 11 | 4 | 1.94 | 42 | 0 | 0.2 | 73 | 0 | 0.1 | 104 | 0 | 0 | 135 | 0 | 0 | 166 | 0 | 0 | 197 | 0 | 0 |
| 12 | 1 | 1.65 | 43 | 0 | 0.2 | 74 | 0 | 0.1 | 105 | 0 | 0 | 136 | 0 | 0 | 167 | 0 | 0 | 198 | 0 | 0 |
| 13 | 2 | 1.43 | 44 | 0 | 0.2 | 75 | 0 | 0.1 | 106 | 0 | 0 | 137 | 0 | 0 | 168 | 0 | 0 | 199 | 0 | 0 |
| 14 | 2 | 1.25 | 45 | 0 | 0.2 | 76 | 0 | 0.1 | 107 | 0 | 0 | 138 | 0 | 0 | 169 | 0 | 0 | 200 | 0 | 0 |
| 15 | 2 | 1.1 | 46 | 0 | 0.1 | 77 | 0 | 0.1 | 108 | 0 | 0 | 139 | 0 | 0 | 170 | 0 | 0 | 201 | 0 | 0 |
| 16 | 2 | 0.98 | 47 | 0 | 0.1 | 78 | 0 | 0.1 | 109 | 0 | 0 | 140 | 0 | 0 | 171 | 0 | 0 | 202 | 0 | 0 |
| 17 | 1 | 0.88 | 48 | 0 | 0.1 | 79 | 0 | 0.1 | 110 | 0 | 0 | 141 | 0 | 0 | 172 | 0 | 0 | 203 | 0 | 0 |
| 18 | 1 | 0.79 | 49 | 0 | 0.1 | 80 | 0 | 0.1 | 111 | 0 | 0 | 142 | 0 | 0 | 173 | 0 | 0 | 204 | 0 | 0 |
| 19 | 1 | 0.72 | 50 | 0 | 0.1 | 81 | 0 | 0.1 | 112 | 0 | 0 | 143 | 0 | 0 | 174 | 0 | 0 | 205 | 0 | 0 |
| 20 | 0 | 0.65 | 51 | 0 | 0.1 | 82 | 0 | 0.1 | 113 | 0 | 0 | 144 | 0 | 0 | 175 | 0 | 0 | 206 | 0 | 0 |
| 21 | 0 | 0.6 | 52 | 0 | 0.1 | 83 | 0 | 0.1 | 114 | 0 | 0 | 145 | 0 | 0 | 176 | 0 | 0 | 207 | 0 | 0 |
| 22 | 0 | 0.55 | 53 | 0 | 0.1 | 84 | 0 | 0.1 | 115 | 0 | 0 | 146 | 0 | 0 | 177 | 0 | 0 | 208 | 0 | 0 |
| 23 | 1 | 0.51 | 54 | 0 | 0.1 | 85 | 0 | 0.1 | 116 | 0 | 0 | 147 | 0 | 0 | 178 | 0 | 0 | 209 | 0 | 0 |
| 24 | 1 | 0.47 | 55 | 1 | 0.1 | 86 | 0 | 0.1 | 117 | 0 | 0 | 148 | 0 | 0 | 179 | 0 | 0 | 210 | 0 | 0 |
| 25 | 1 | 0.44 | 56 | 0 | 0.1 | 87 | 0 | 0.1 | 118 | 0 | 0 | 149 | 0 | 0 | 180 | 0 | 0 | 211 | 0 | 0 |
| 26 | 1 | 0.41 | 57 | 0 | 0.1 | 88 | 0 | 0 | 119 | 0 | 0 | 150 | 0 | 0 | 181 | 0 | 0 | 212 | 0 | 0 |
| 27 | 0 | 0.38 | 58 | 0 | 0.1 | 89 | 0 | 0 | 120 | 0 | 0 | 151 | 0 | 0 | 182 | 0 | 0 | 213 | 1 | 0 |
| 28 | 0 | 0.35 | 59 | 0 | 0.1 | 90 | 0 | 0 | 121 | 0 | 0 | 152 | 0 | 0 | 183 | 0 | 0 | | | |
| 29 | 0 | 0.33 | 60 | 0 | 0.1 | 91 | 0 | 0 | 122 | 0 | 0 | 153 | 0 | 0 | 184 | 0 | 0 | | | |
| 30 | 0 | 0.31 | 61 | 0 | 0.1 | 92 | 0 | 0 | 123 | 0 | 0 | 154 | 0 | 0 | 185 | 0 | 0 | | | |
| 31 | 0 | 0.29 | 62 | 0 | 0.1 | 93 | 0 | 0 | 124 | 0 | 0 | 155 | 0 | 0 | 186 | 0 | 0 | | | |
| | **a = 1.82** | | | **R = 213** | | | **$X^2$ = 17.98** | | | | **DF = 26** | | | **P($X^2$) = 0.88** | | | | | | |

Table 10: John. Secondary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 166 | 156 | 26 | 1 | 0.4 | 51 | 0 | 0.1 | 76 | 0 | 0.1 | 101 | 0 | 0 | 126 | 0 | 0 | 151 | 0 | 0 |
| 2 | 41 | 43.1 | 27 | 0 | 0.4 | 52 | 0 | 0.1 | 77 | 0 | 0.1 | 102 | 0 | 0 | 127 | 0 | 0 | 152 | 0 | 0 |
| 3 | 15 | 20.3 | 28 | 0 | 0.3 | 53 | 0 | 0.1 | 78 | 0 | 0.1 | 103 | 0 | 0 | 128 | 0 | 0 | 153 | 0 | 0 |
| 4 | 16 | 11.9 | 29 | 0 | 0.3 | 54 | 0 | 0.1 | 79 | 0 | 0.1 | 104 | 0 | 0 | 129 | 0 | 0 | 154 | 0 | 0 |
| 5 | 5 | 7.88 | 30 | 0 | 0.3 | 55 | 0 | 0.1 | 80 | 0 | 0.1 | 105 | 0 | 0 | 130 | 0 | 0 | 155 | 0 | 0 |
| 6 | 4 | 5.62 | 31 | 1 | 0.3 | 56 | 0 | 0.1 | 81 | 0 | 0.1 | 106 | 0 | 0 | 131 | 0 | 0 | 156 | 0 | 0 |
| 7 | 3 | 4.23 | 32 | 0 | 0.3 | 57 | 0 | 0.1 | 82 | 1 | 0 | 107 | 0 | 0 | 132 | 0 | 0 | 157 | 0 | 0 |
| 8 | 5 | 3.3 | 33 | 0 | 0.2 | 58 | 0 | 0.1 | 83 | 0 | 0 | 108 | 0 | 0 | 133 | 0 | 0 | 158 | 0 | 0 |
| 9 | 2 | 2.65 | 34 | 0 | 0.2 | 59 | 0 | 0.1 | 84 | 0 | 0 | 109 | 0 | 0 | 134 | 0 | 0 | 159 | 0 | 0 |
| 10 | 2 | 2.18 | 35 | 1 | 0.2 | 60 | 0 | 0.1 | 85 | 0 | 0 | 110 | 0 | 0 | 135 | 0 | 0 | 160 | 0 | 0 |
| 11 | 1 | 1.83 | 36 | 1 | 0.2 | 61 | 1 | 0.1 | 86 | 0 | 0 | 111 | 0 | 0 | 136 | 0 | 0 | 161 | 0 | 0 |
| 12 | 1 | 1.56 | 37 | 1 | 0.2 | 62 | 0 | 0.1 | 87 | 0 | 0 | 112 | 0 | 0 | 137 | 0 | 0 | 162 | 0 | 0 |
| 13 | 0 | 1.34 | 38 | 0 | 0.2 | 63 | 0 | 0.1 | 88 | 0 | 0 | 113 | 0 | 0 | 138 | 0 | 0 | 163 | 0 | 0 |
| 14 | 0 | 1.17 | 39 | 1 | 0.2 | 64 | 0 | 0.1 | 89 | 0 | 0 | 114 | 0 | 0 | 139 | 0 | 0 | 164 | 1 | 0 |
| 15 | 0 | 1.03 | 40 | 0 | 0.2 | 65 | 0 | 0.1 | 90 | 0 | 0 | 115 | 0 | 0 | 140 | 0 | 0 | | | |
| 16 | 2 | 0.91 | 41 | 0 | 0.2 | 66 | 0 | 0.1 | 91 | 0 | 0 | 116 | 0 | 0 | 141 | 0 | 0 | | | |
| 17 | 3 | 0.82 | 42 | 1 | 0.2 | 67 | 0 | 0.1 | 92 | 0 | 0 | 117 | 0 | 0 | 142 | 0 | 0 | | | |
| 18 | 0 | 0.73 | 43 | 0 | 0.2 | 68 | 0 | 0.1 | 93 | 0 | 0 | 118 | 0 | 0 | 143 | 0 | 0 | | | |
| 19 | 0 | 0.66 | 44 | 0 | 0.1 | 69 | 1 | 0.1 | 94 | 0 | 0 | 119 | 0 | 0 | 144 | 0 | 0 | | | |
| 20 | 1 | 0.6 | 45 | 0 | 0.1 | 70 | 0 | 0.1 | 95 | 0 | 0 | 120 | 0 | 0 | 145 | 0 | 0 | | | |
| 21 | 0 | 0.55 | 46 | 0 | 0.1 | 71 | 0 | 0.1 | 96 | 0 | 0 | 121 | 0 | 0 | 146 | 0 | 0 | | | |
| 22 | 0 | 0.51 | 47 | 0 | 0.1 | 72 | 0 | 0.1 | 97 | 0 | 0 | 122 | 0 | 0 | 147 | 0 | 0 | | | |
| 23 | 1 | 0.47 | 48 | 0 | 0.1 | 73 | 0 | 0.1 | 98 | 0 | 0 | 123 | 0 | 0 | 148 | 0 | 0 | | | |
| 24 | 0 | 0.43 | 49 | 0 | 0.1 | 74 | 0 | 0.1 | 99 | 0 | 0 | 124 | 0 | 0 | 149 | 0 | 0 | | | |
| 25 | 0 | 0.4 | 50 | 0 | 0.1 | 75 | 0 | 0.1 | 100 | 0 | 0 | 125 | 0 | 0 | 150 | 0 | 0 | | | |
| | | **a = 1.85** | | | **R = 164** | | | ***X²* = 17.98** | | | | **DF = 26** | | | **P(*X²*) = 0.88** | | | | | |

Table 11: Apocalypse. Primary Process

| **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | **x** | **g(x)** | **NPx** | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 165 | 163 | 13 | 0 | 1.5 | 25 | 1 | 0.5 | 37 | 0 | 0.2 | 49 | 0 | 0.1 | 61 | 1 | 0.1 | 73 | 0 | 0.1 |
| 2 | 46 | 45.6 | 14 | 3 | 1.3 | 26 | 1 | 0.4 | 38 | 0 | 0.2 | 50 | 0 | 0.1 | 62 | 0 | 0.1 | 74 | 0 | 0.1 |
| 3 | 23 | 21.7 | 15 | 0 | 1.1 | 27 | 0 | 0.4 | 39 | 0 | 0.2 | 51 | 0 | 0.1 | 63 | 0 | 0.1 | 75 | 1 | 0.1 |
| 4 | 9 | 12.8 | 16 | 1 | 1 | 28 | 0 | 0.4 | 40 | 0 | 0.2 | 52 | 0 | 0.1 | 64 | 0 | 0.1 | | | |
| 5 | 8 | 8.51 | 17 | 1 | 0.9 | 29 | 0 | 0.3 | 41 | 0 | 0.2 | 53 | 1 | 0.1 | 65 | 0 | 0.1 | | | |
| 6 | 7 | 6.1 | 18 | 0 | 0.8 | 30 | 0 | 0.3 | 42 | 0 | 0.2 | 54 | 0 | 0.1 | 66 | 1 | 0.1 | | | |
| 7 | 3 | 4.6 | 19 | 1 | 0.7 | 31 | 0 | 0.3 | 43 | 0 | 0.2 | 55 | 0 | 0.1 | 67 | 0 | 0.1 | | | |
| 8 | 3 | 3.6 | 20 | 0 | 0.7 | 32 | 0 | 0.3 | 44 | 0 | 0.2 | 56 | 0 | 0.1 | 68 | 0 | 0.1 | | | |
| 9 | 4 | 2.9 | 21 | 2 | 0.6 | 33 | 0 | 0.3 | 45 | 0 | 0.2 | 57 | 0 | 0.1 | 69 | 0 | 0.1 | | | |
| 10 | 5 | 2.39 | 22 | 1 | 0.6 | 34 | 1 | 0.3 | 46 | 0 | 0.2 | 58 | 0 | 0.1 | 70 | 0 | 0.1 | | | |
| 11 | 2 | 2.01 | 23 | 0 | 0.5 | 35 | 0 | 0.2 | 47 | 0 | 0.1 | 59 | 0 | 0.1 | 71 | 0 | 0.1 | | | |
| 12 | 0 | 1.71 | 24 | 2 | 0.5 | 36 | 0 | 0.2 | 48 | 0 | 0.1 | 60 | 0 | 0.1 | 72 | 0 | 0.1 | | | |
| | | **a = 1.83** | | | **R = 75** | | | ***X²* = 21.41** | | | | **DF = 23** | | | **P(*X²*) = 0.56** | | | | | |

Table 12: Apocalypse. Secondary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 118 | 116.6 | 11 | 0 | 0.58 | 21 | 0 | 0.14 | 31 | 0 | 0.06 | 41 | 0 | 0.03 | 51 | 0 | 0.02 |
| 2 | 17 | 25.12 | 12 | 2 | 0.48 | 22 | 0 | 0.12 | 32 | 0 | 0.05 | 42 | 0 | 0.03 | 52 | 0 | 0.02 |
| 3 | 10 | 10.24 | 13 | 1 | 0.4 | 23 | 0 | 0.11 | 33 | 1 | 0.05 | 43 | 0 | 0.03 | 53 | 0 | 0.02 |
| 4 | 5 | 5.41 | 14 | 3 | 0.34 | 24 | 0 | 0.1 | 34 | 0 | 0.05 | 44 | 0 | 0.03 | 54 | 0 | 0.02 |
| 5 | 1 | 3.3 | 15 | 0 | 0.29 | 25 | 0 | 0.09 | 35 | 0 | 0.04 | 45 | 0 | 0.03 | 55 | 0 | 0.02 |
| 6 | 5 | 2.21 | 16 | 0 | 0.25 | 26 | 0 | 0.09 | 36 | 0 | 0.04 | 46 | 0 | 0.02 | 56 | 1 | 0.02 |
| 7 | 2 | 1.57 | 17 | 0 | 0.22 | 27 | 0 | 0.08 | 37 | 0 | 0.04 | 47 | 0 | 0.02 | | | |
| 8 | 2 | 1.17 | 18 | 1 | 0.19 | 28 | 0 | 0.07 | 38 | 0 | 0.04 | 48 | 0 | 0.02 | | | |
| 9 | 1 | 0.9 | 19 | 0 | 0.17 | 29 | 0 | 0.07 | 39 | 0 | 0.04 | 49 | 0 | 0.02 | | | |
| 10 | 1 | 0.71 | 20 | 1 | 0.15 | 30 | 0 | 0.06 | 40 | 0 | 0.03 | 50 | 0 | 0.02 | | | |
| | | **a = 2.21** | | | **R = 56** | | | $X^2$ **= 14.94** | | | | **DF = 10** | | | **P(*X²*) = 0.13** | | |

Table 13: Romans. Primary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 113 | 113 | 21 | 0 | 0.3 | 41 | 0 | 0.1 | 61 | 0 | 0 | 81 | 0 | 0 | 101 | 0 |
| 2 | 36 | 28.8 | 22 | 0 | 0.3 | 42 | 0 | 0.1 | 62 | 0 | 0 | 82 | 0 | 0 | 102 | 0 |
| 3 | 11 | 12.9 | 23 | 0 | 0.2 | 43 | 0 | 0.1 | 63 | 0 | 0 | 83 | 0 | 0 | 103 | 0 |
| 4 | 8 | 7.33 | 24 | 0 | 0.2 | 44 | 0 | 0.1 | 64 | 0 | 0 | 84 | 0 | 0 | 104 | 0 |
| 5 | 6 | 4.72 | 25 | 0 | 0.2 | 45 | 0 | 0.1 | 65 | 0 | 0 | 85 | 0 | 0 | 105 | 0 |
| 6 | 4 | 3.29 | 26 | 2 | 0.2 | 46 | 0 | 0.1 | 66 | 0 | 0 | 86 | 0 | 0 | 106 | 0 |
| 7 | 1 | 2.43 | 27 | 0 | 0.2 | 47 | 0 | 0.1 | 67 | 0 | 0 | 87 | 0 | 0 | 107 | 0 |
| 8 | 2 | 1.86 | 28 | 0 | 0.2 | 48 | 0 | 0.1 | 68 | 0 | 0 | 88 | 0 | 0 | 108 | 0 |
| 9 | 0 | 1.48 | 29 | 0 | 0.2 | 49 | 0 | 0.1 | 69 | 0 | 0 | 89 | 0 | 0 | 109 | 0 |
| 10 | 0 | 1.2 | 30 | 0 | 0.1 | 50 | 0 | 0.1 | 70 | 0 | 0 | 90 | 0 | 0 | 110 | 0 |
| 11 | 0 | 0.99 | 31 | 0 | 0.1 | 51 | 0 | 0.1 | 71 | 0 | 0 | 91 | 0 | 0 | 111 | 0 |
| 12 | 1 | 0.84 | 32 | 0 | 0.1 | 52 | 0 | 0.1 | 72 | 0 | 0 | 92 | 0 | 0 | 112 | 0 |
| 13 | 1 | 0.71 | 33 | 0 | 0.1 | 53 | 0 | 0 | 73 | 0 | 0 | 93 | 0 | 0 | 113 | 0 |
| 14 | 0 | 0.62 | 34 | 0 | 0.1 | 54 | 0 | 0 | 74 | 0 | 0 | 94 | 0 | 0 | 114 | 0 |
| 15 | 0 | 0.54 | 35 | 0 | 0.1 | 55 | 0 | 0 | 75 | 0 | 0 | 95 | 0 | 0 | 115 | 0 |
| 16 | 0 | 0.47 | 36 | 0 | 0.1 | 56 | 0 | 0 | 76 | 0 | 0 | 96 | 0 | 0 | 116 | 1 |
| 17 | 0 | 0.42 | 37 | 0 | 0.1 | 57 | 0 | 0 | 77 | 0 | 0 | 97 | 0 | 0 | | |
| 18 | 1 | 0.38 | 38 | 0 | 0.1 | 58 | 1 | 0 | 78 | 0 | 0 | 98 | 0 | 0 | | |
| 19 | 0 | 0.34 | 39 | 0 | 0.1 | 59 | 0 | 0 | 79 | 0 | 0 | 99 | 0 | 0 | | |
| 20 | 0 | 0.3 | 40 | 0 | 0.1 | 60 | 0 | 0 | 80 | 0 | 0 | 100 | 0 | 0 | | |
| | **a = 1.98** | | | **R = 116** | | | **$X^2$ = 10.25** | | | | **DF = 16** | | | **P($X^2$) = 0.85** | | |

Table 14: Romans. Secondary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 197 | 190.9 | 13 | 2 | 0.71 | 25 | 0 | 0.17 | 37 | 1 | 0.07 | 49 | 0 | 0.04 | 61 | 0 | 0.02 | 73 | 0 | 0 |
| 2 | 40 | 42.13 | 14 | 0 | 0.61 | 26 | 0 | 0.16 | 38 | 0 | 0.07 | 50 | 0 | 0.04 | 62 | 0 | 0.02 | 74 | 0 | 0 |
| 3 | 11 | 17.41 | 15 | 2 | 0.52 | 27 | 1 | 0.14 | 39 | 0 | 0.07 | 51 | 0 | 0.04 | 63 | 0 | 0.02 | 75 | 1 | 0 |
| 4 | 9 | 9.3 | 16 | 2 | 0.45 | 28 | 0 | 0.13 | 40 | 0 | 0.06 | 52 | 0 | 0.03 | 64 | 0 | 0.02 | | | |
| 5 | 4 | 5.72 | 17 | 0 | 0.4 | 29 | 0 | 0.12 | 41 | 0 | 0.06 | 53 | 0 | 0.03 | 65 | 0 | 0.02 | | | |
| 6 | 4 | 3.84 | 18 | 1 | 0.35 | 30 | 0 | 0.12 | 42 | 0 | 0.06 | 54 | 0 | 0.03 | 66 | 0 | 0.02 | | | |
| 7 | 3 | 2.75 | 19 | 0 | 0.31 | 31 | 0 | 0.11 | 43 | 0 | 0.05 | 55 | 0 | 0.03 | 67 | 0 | 0.02 | | | |
| 8 | 4 | 2.05 | 20 | 0 | 0.28 | 32 | 0 | 0.1 | 44 | 0 | 0.05 | 56 | 0 | 0.03 | 68 | 0 | 0.02 | | | |
| 9 | 0 | 1.59 | 21 | 0 | 0.25 | 33 | 0 | 0.09 | 45 | 0 | 0.05 | 57 | 0 | 0.03 | 69 | 0 | 0.02 | | | |
| 10 | 0 | 1.26 | 22 | 1 | 0.23 | 34 | 1 | 0.09 | 46 | 0 | 0.05 | 58 | 0 | 0.03 | 70 | 0 | 0.02 | | | |
| 11 | 0 | 1.03 | 23 | 0 | 0.21 | 35 | 0 | 0.08 | 47 | 0 | 0.04 | 59 | 0 | 0.03 | 71 | 0 | 0.02 | | | |
| 12 | 2 | 0.85 | 24 | 0 | 0.19 | 36 | 0 | 0.08 | 48 | 0 | 0.04 | 60 | 0 | 0.03 | 72 | 0 | 0.02 | | | |
| | **a = 2.18** | | | **R = 75** | | | **$X^2$ = 14.54** | | | | **DF = 14** | | | **P($X^2$) = 0.41** | | | | | | |

Table 15: Herculano. Primary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 188 | 183 | 13 | 1 | 1 | 25 | 0 | 0.3 | 37 | 0 | 0.1 | 49 | 0 | 0.1 | 61 | 0 | 0.04 | 73 | 0 | 0 |
| 2 | 37 | 44.2 | 14 | 0 | 0.8 | 26 | 0 | 0.2 | 38 | 0 | 0.1 | 50 | 0 | 0.1 | 62 | 0 | 0.04 | 74 | 0 | 0 |
| 3 | 23 | 19.3 | 15 | 1 | 0.7 | 27 | 0 | 0.2 | 39 | 0 | 0.1 | 51 | 0 | 0.1 | 63 | 0 | 0.04 | 75 | 0 | 0 |
| 4 | 11 | 10.7 | 16 | 0 | 0.6 | 28 | 0 | 0.2 | 40 | 0 | 0.1 | 52 | 0 | 0.1 | 64 | 1 | 0.04 | 76 | 0 | 0 |
| 5 | 9 | 6.77 | 17 | 0 | 0.6 | 29 | 0 | 0.2 | 41 | 0 | 0.1 | 53 | 0 | 0.1 | 65 | 0 | 0.04 | 77 | 0 | 0 |
| 6 | 3 | 4.66 | 18 | 0 | 0.5 | 30 | 0 | 0.2 | 42 | 0 | 0.1 | 54 | 0 | 0.1 | 66 | 0 | 0.03 | 78 | 0 | 0 |
| 7 | 3 | 3.4 | 19 | 0 | 0.4 | 31 | 0 | 0.2 | 43 | 0 | 0.1 | 55 | 0 | 0.1 | 67 | 0 | 0.03 | 79 | 0 | 0 |
| 8 | 1 | 2.59 | 20 | 1 | 0.4 | 32 | 0 | 0.2 | 44 | 0 | 0.1 | 56 | 0 | 0.1 | 68 | 0 | 0.03 | 80 | 1 | 0 |
| 9 | 6 | 2.03 | 21 | 1 | 0.4 | 33 | 0 | 0.1 | 45 | 0 | 0.1 | 57 | 0 | 0.1 | 69 | 0 | 0.03 | | | |
| 10 | 0 | 1.64 | 22 | 0 | 0.3 | 34 | 0 | 0.1 | 46 | 0 | 0.1 | 58 | 0 | 0 | 70 | 0 | 0.03 | | | |
| 11 | 3 | 1.35 | 23 | 0 | 0.3 | 35 | 0 | 0.1 | 47 | 0 | 0.1 | 59 | 0 | 0 | 71 | 0 | 0.03 | | | |
| 12 | 1 | 1.13 | 24 | 0 | 0.3 | 36 | 0 | 0.1 | 48 | 0 | 0.1 | 60 | 0 | 0 | 72 | 0 | 0.03 | | | |
| | **a = 2.05** | | | **R = 80** | | | ***X²* = 21.49** | | | | **DF = 17** | | | **P(*X²*) = 0.21** | | | | | | |

Table 16: Herculano. Secondary Process

| x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx | x | g(x) | NPx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 81 | 75.2 | 6 | 1 | 1.9 | 11 | 2 | 0.5 | 16 | 0 | 0.3 | 21 | 2 | 0.1 | 26 | 0 | 0.09 | 31 | 0 | 0.1 |
| 2 | 13 | 18 | 7 | 2 | 1.4 | 12 | 1 | 0.5 | 17 | 0 | 0.2 | 22 | 0 | 0.1 | 27 | 0 | 0.08 | 32 | 0 | 0.1 |
| 3 | 6 | 7.81 | 8 | 0 | 1 | 13 | 0 | 0.4 | 18 | 0 | 0.2 | 23 | 1 | 0.1 | 28 | 0 | 0.08 | 33 | 0 | 0.1 |
| 4 | 2 | 4.32 | 9 | 2 | 0.8 | 14 | 1 | 0.3 | 19 | 0 | 0.2 | 24 | 0 | 0.1 | 29 | 0 | 0.07 | 34 | 0 | 0.1 |
| 5 | 2 | 2.73 | 10 | 0 | 0.7 | 15 | 0 | 0.3 | 20 | 1 | 0.2 | 25 | 0 | 0.1 | 30 | 0 | 0.07 | 35 | 1 | 0.1 |
| | **a = 2.06** | | | **R = 35** | | | ***X²* = 8.93** | | | | **DF = 9** | | | **P(*X²*) = 0.44** | | | | | | |

## *4 DISCUSSION*

The analysis has shown that the right-truncated Zeta distribution can be fitted satisfactorily to word-frequency spectra within the categories of primary and secondary process, as identified in texts using the Portuguese translation of the Regressive Imagery Dictionary. Acceptable fits were found for all six biblical books examined in the study, as well as for the romance by Herculano, with the majority showing a very good quality of fit. This would seem, therefore, to provide further support for the validity of the main categories within the RID. As well as showing good construct validity in previous studies, the categories have also now demonstrated formal quantitative behaviour that shows them to be correctly delimited entities in texts.

In all cases, the parameter $a$ of the right-truncated Zeta distribution approximated towards a value of 2 (mean = 2.00, min = 1.81, max = 2.21, sd = 0.13). However, in five out of the seven texts, the value of $a$ for the secondary process category was slightly larger than its value for the primary process category.[2] On the basis of these few preliminary data alone, it is difficult to hazard any very strong claims as to why the size of $a$ should tend to be slightly higher for the secondary process category. It certainly does not seem to be related to the dominance of one or the other category within a text, since the Epistle to the Romans, which has a significant dominance of secondary process words, shows exactly the same pattern of $a$-values as texts such as the Apocalypse, which has a primary process dominance (cf. Table 2). It seems more likely that the relative size of $a$ may be related to the *a priori* probability of occurrence of certain words that are contained within the pre-defined dictionary categories, since, in all cases, the value of R – i.e. the frequency of the most common word in the category – was higher for primary process than for secondary process: in other words, the most frequent primary process word always occurred more often than the most frequent secondary process word, even in a text with an overall dominance of secondary process words. However, this relationship requires further explicit investigation on a larger number of texts.

[2] In the case of two other texts – the Gospel according to John and the romance by Herculano – the values of $a$ for primary and secondary process were almost equal.

Further research should also aim to extend and consolidate these findings in a number of other ways. First, the Portuguese RID should be applied to a larger sample of texts, including other genres, in order to confirm the model suggested here. In such a study, any further systematic patterns of variation in the parameter *a* should be examined, as the parameters of frequency distributions have previously shown clear evidence of links to text typology and authorship. Second, the experiments should be replicated using other versions of the RID, in order to check that the findings of the present study are generally valid and not an artefact of the Portuguese version alone. For example, Hogenraad (2005) has investigated the behaviour of several versions of the RID on parallel translation texts and has found that the size of the correlations between category counts produced by different versions does vary. Third, the research should be extended to cover other kinds of psychoanalytic categories in texts – for example, the categories of body boundary definiteness developed in Wilson (2006) and the categories of anality and orality constructed by Vanheule, Desmet & Meganck (2008). Finally, the most challenging task would be to integrate these empirical findings into a synergetic theory of text production that incorporates cognitive elements such as consciousness states and concept-word mappings. Researchers such as Roy (2004) and Spivak (2004) have already made interesting contributions in this direction, but there is still much work that remains to be done in developing a comprehensive model.

## *REFERENCES*

Altmann, G. (2002): Zipfian linguistics, in: *Glottometrics,* 3, 19-26.

Aulagnier, P. (2001): *The violence of interpretation: From pictogram to statement* (A. Sheridan, Trans.). London: Brunner-Routledge. (Original work published 1975.)

Baayen, R.H. (2001): *Word frequency distributions.* Dordrecht: Kluwer.

Bucci, W. (1997): *Psychoanalysis and cognitive science: A multiple code theory.* New York: Guilford Press.

Elter-Nodvin, E. (2000): Computerized content analysis: a comparison of the verbal productions of high hypnotizable, low hypnotizable and simulating subjects. Ph.D. dissertation, University of Tennessee, Knoxville.

Goldstein, K. (1939): *The organism.* Boston: Beacon.

Freud, S. (1900): *Die Traumdeutung.* Leipzig & Vienna: Franz Deuticke.

Hogenraad, R. (2005): The Regressive Imagery Dictionary: a test of five versions (English, French, German, Portuguese, and Swedish). Paper presented

at the International Congress on Aesthetics, Creativity, and Psychology of the Arts, Perm, Russia, June 2005.

Hogenraad, R.; Daubies, C.; Bestgen, Y. & Mahau, P. (2003): Une théorie et une méthode générale d'analyse textuelle assistée par ordinateur. Le système PROTAN (PROTocol ANalyzer). 32-bits version of November 10, 2003 by Pierre Mahau. Psychology Department, Catholic University of Louvain, Louvain-la-Neuve.

Kristeva, J. (1996/1997): Freudian models of language: A conversation. In: *Psychomedia: Journal of European Psychoanalysis, 3/4.* [Retrieved December 8, 2006 from [http://www.psychomedia.it/jep/number3-4/kristeng.htm].

Lévy-Bruhl, L. (1966): *How natives think.* New York: Washington Square Press. (Original work published 1910.)

Martindale, C. (1975): *Romantic progression: the psychology of literary history.* Washington, DC: Hemisphere.

Martindale, C. (1976): Primitive mentality and the relationship between art and society, in: *Scientific Aesthetics,* 1, 5-18.

Martindale, C. & Fischer, R. (1977): The effects of psilocybin on primary process content in language. In: *Confinia Psychiatrica, 20,* 195-202.

Piaget, J. (1954): *The construction of reality in the child.* New York: Basic Books.

Popescu, I.-I.; Best, K.-H. & Altmann, G. (2007): On the dynamics of word classes in text, in: *Glottometrics,* 14, 58-71.

Prün, C. (1999): G.K. Zipf's conception of language as an early prototype of synergetic linguistics, in: *Journal of Quantitative Linguistics,* 6, 78-84.

Roy, P.K. (2004): Stochastic resonance as an emerging technique for neuronmodulation and pharmacolinguistics: using nonlinear dynamics to analyze drug-induced language transition and EEG, in: *Journal of Quantitative Linguistics,* 11, 49-77.

Sorokin, P. (1957): *Social and cultural dynamics: A study of change in major systems of art, truth, ethics, law and social relationships.* Boston, MA: Porter Sargent.

Spivak, D. (2004): Linguistics of altered states of consciousness: problems and prospects, in: *Journal of Quantitative Linguistics,* 11, 27-32.

Strauss, U.; Fan, F. & Altmann, G. (2008): *Problems in quantitative linguistics 1* (2nd ed.). Lüdenscheid: RAM-Verlag.

Uhlířová, L. (1995): On the generality of statistical laws and individuality of texts. A case of syllables, word forms, their length and frequencies, in: *Journal of Quantitative Linguistics,* 2, 238-247.

Vanheule, S., Desmet, M. & Meganck, R. (2008): Anal and oral word use in relation to dependency and self-criticism. Poster presentation at the Winter Meeting of the American Psychoanalytic Association, New York, 2008.

Werner, H. (1948): *Comparative psychology of mental development.* New York: International Universities Press.

West, A. (1991): Primary process content in the King James Bible: the five stages of Christian mysticism, in: *Computers and the Humanities,* 25, 227-238.

West, A. & Martindale, C. (1988): Primary process content in paranoid schizophrenic speech, in: *Journal of Genetic Psychology,* 149, 547-553.

West, A.; Martindale, C.; Hines, D. & Roth, W. (1983): Marijuana-induced primary process content in the TAT, in: *Journal of Personality Assessment,* 47, 466-467.

West, A.; Martindale, C. & Sutton-Smith, B. (1985): Age trends in the content of children's spontaneous fantasy narratives, in: *Genetic, Social, and General Psychology Monographs,* 111, 391-405.

Wilson, A. (2002): The application of computer content analysis in sexology: a case study of primary process content in fictional fetishistic narratives, in: *Electronic Journal of Human Sexuality,* 5. [Retrieved October 31, 2006, from: http://www.ejhs.org/volume5/wilson.html].

Wilson, A. (2006): Development and application of a content analysis dictionary for body boundary research, in: *Literary and Linguistic Computing,* 21, 105-110.

Wilson, A. (2007): Barrier and penetration imagery as a supplementary measure of altered states of consciousness discourse: replicating the five-stage model of Christian mysticism in the Bible. Forthcoming.

Ziegler, A., Best, K.-H. & Altmann, G. (2002): Nominalstil, in: *Empirical Text and Culture Research,* 2, 72-85.

Zipf, G.K. (1935): *The psycho-biology of language: an introduction to dynamic philology.* Boston: Houghton Mifflin.

# An integral qualitative-quantitative approach to the study of concept realization in the text

*Nadia Yesypenko (Chernivtsi, Ukraine)*

## *1. INTRODUCTION*

A vast array of accounts attempt to explain the nature of concepts. According to classical accounts, a concept denotes all the entities, phenomena and relations in a given category or class by using their definition. Concepts are discursive and result from reason. Concepts are expected to be useful in dealing with reality. Generally speaking, concepts are taken to be acquired dispositions to recognize perceived objects, to understand what this or that kind of object is like (Langacke 1990). Mastery of symbolic thought (in particular, language) is a prerequisite for conceptual thought. Concepts are bearers of meaning.

Cognitive semantics gives a treatment of issues in the construction of meaning both at the level of the sentence and the level of the lexeme in terms of the structure of concepts (Schwarz 1992). If we aim to look how linguistic strings convey different semantic content, we must first understand what cognitive processes are being used to do it. Researchers start by investigating the ways how people structure their experience through language (Langacke 1990; Wierzbicka 1985, 1992). Language is full of conventions which allow for subtle and nuanced conveyances of experience. According to linguists William Croft and D. Alan Cruse (2004), there are four broad cognitive abilities that play an active part in the construction of a concept. They are: attention, judgement, situatedness and constitution/gestalt. They explain the ways we encode experience into language in some unique way. Concepts are a glimpse of a different world, one which contains timeless truths (James 1975), that is why language and cognition mutually influence one another, and are both embedded in the experiences and environments of the users. This can be considered a moderate offshoot of the Sapir-Whorf hypothesis (Whorf 1956). A concept in language describes a certain state of affairs in the world, namely, the way that some object is presented.

Cognitive linguistics challenges us to move from seeing language as an abstract entity to seeing our words as having meaning in a particular condition. Even more significant, writer's words are used to convey a broad sense of meanings and the meaning he conveys with those words is identified by his immediate experience. Concepts are reflected images that the author reproduces in language. Concepts presented in a literary text are shaped and constrained by the words selected by the author. The semantic layer of the text represents several general ways of the verbal presentation of objects and emotional colouring of the given world.

In the article we aim to explore and understand how different concepts are mediated by mainstream wordstock in adventure novels "The Adventures of Tom Sawyer" by M. Twain, "Gulliver's Travels" by J. Swift and "A Handful of Dust" by E. Waugh. Starting with the full text, working down to the high-frequency vocabulary, we peel back the dominant lexical semantic classes of nouns, verbs, adjective and adverbs to reveal verbalized concepts that build up a language world view in the adventure novel.

## *2. METHODS OF ANALYSES*

Concept communication through language is an important activity of human thinking which is basic for interpreting realities. There is a wide spectrum of methods to perform concept structure and lexicalization analysis, but the interest is to develop further methods going beyond logical understanding of concepts and to advocate logic-mathematical proofs of concept verbal realization. Especially, there is a strong demand for objective methods deprived of subjective interpretation of cognitive processes. This has stimulated an integrative approach based on qualitative and cognitive methods of text analysis. Such an approach, which defines concepts verbalized in text through presice statistical calculations, overcomes intuitional scaling of concepts. The goal is to reach a well-founded explanatory approach to modeling a language world view of a written text that closely approximates and reflects the author's world view. Methodologically, this aims at statistical analysis of the word use in the author's vocabulary and then cognitive interpretation of the dominant lexical semantic classes that allows to construct concept hierarchy in the given text.

Thus, based on qualitative and quantitative analyses of a very large corpus this article introduces a conceptual and lexical-semantic

framework for the linguistic description of concept lexicalization and modeling a language world view. Lexical semantic analysis of the wordstock helps to divide four notional parts of speech into lexical semantic word groups. Statistical analysis involves counting particular features of the textual data and then applying some mathematical transformations.

The simplest type of analysis produces frequency list of word-forms, arranged from the most to the least frequent. More powerful and complex types of statistical analysis are used in building a model of language world view in the novels under study. In the case of conceptual analysis, basic means are mainly translations of statistically defined high-frequent lexical semantic classes. Those translations interpret the quantitative calculations with respect to actual word use in text, so that prevailing classes become understandable as the bearers of principal concepts.

The analysis of realization of lexical semantic classes in the text with the help of statistical methods (Altmann 1996; Levickij 2004) has resulted in a wide variety of algorithms that use the distributional hypothesis to discover many aspects of concept verbalization by applying statistical techniques to large corpora of word-forms selected from literary texts. The chi-square criterion of independence made it possible to define the accordance and the difference of the frequency of lexical semantic classes.

The present research is characterized by the application of quantitative methods complementing the qualitative view in the field of concept realization in the text. It is one way to overcome the subjective approach of a researcher to concept analysis. Statistics in this case is an appropriate meta-language for linguistic studies.

## *3. HIGH-FREQUENCY WORDS*

In our research we are interested in words rather than word-forms. Hence a quite crude but useful technique is to look through a list of the most frequent words for anything that is unusual or particularly characteristic of the texts in question, especially if we are focused on concepts, represented by words. We need to find all the relevant synonyms and combine the frequency of all the inflected forms. When we find a potentially interesting word (reflecting a certain concept), the next step is to run a concordance on it, then see what concept pattern can we

spot. Such sorting tends to bring out patterns since a certain lexical semantic class applies to one generalized concept having a full set of words semantically bound with the concept. Each generalized concept comprises a sequence of sub-concepts, represented by every word of a certain lexical semantic class. Our intension of grouping such sub-concepts is to show a broader verbal implementation of the represented generalized concept. For example, a generalized concept of *feelings/ emotions* includes sub-concepts of *love, hatered, anger, tenderness, sympathy, fright, despair, dismay, dread, faith, fury, fear, happiness, harmony.*

The calculation of actual realization of lexical semantic word groups of nouns, verbs, adjectives and adverbs in the texts by three authors shows uneven frequency of their use in the probed works (Tables 1, 2, 3, 4).

Nouns denoting *people/mythical characters*; *devices/articles of furniture*; *building/premises* are most frequent in the novels by three writers. There we find word groups that are high-frequently used in one novel only: nouns describing *appearance*; *proper names/nick-names* in M. Twain's novel; nouns presenting *wildlife*; *actions/changes/ movements* in J. Swift's novel; nouns showing *time* in E. Waugh's novel.

The realization of verb group in the three novels shows that the authors tend to have a dynamic storyline, as the lexical semantic group of *motion* is the most frequent. The verbs denoting *existence* are prevailing in all texts under study. In E. Waugh and M. Twain's novels verbs presenting *physical action* and *communication* are the most numerous. J. Swift uses verbs denoting *ownership /loss* in greater number.

Adjectives denoting *shape/size, action done to the object, evaluation of value/function of the object* dominate among lexical semantic group in the novels by J. Swift and M. Twain. High-frequent use of adjectives that show *degree/intensity* is characteristic for E. Waugh and J. Swift. *Physical/natural condition, positive evaluation* adjectives are very frequent in E. Waugh's novel only.

The most frequent use of adverb lexical semantic groups is characteristic for adverbs of *time*; *manner*; *degree and quantity* in the three probated novels.

Table 1
Frequency of the lexical semantic groups of nouns

| | Lexical semantic word group | "A Handful of Dust" by E. Waugh | "Gulliver's Travels" by J. Swift | "The Adventures of Tom Sawyer" by M. Twain |
|---|---|---|---|---|
| 1. | Appearance / parts of the body | 45 | 86 | 100 |
| 2. | Feelings/ emotions | 35 | 62 | 50 |
| 3. | Proper names / nicknames | 281 | 6 | 142 |
| 4. | Establishments /groupings | 10 | 32 | 18 |
| 5. | Diseases / defects | 8 | 18 | 6 |
| 6. | General notions of people / mythical characters | 133 | 138 | 128 |
| 7. | Devices / articles of furniture | 102 | 92 | 106 |
| 8. | Abstract notions | 88 | 154 | 122 |
| 9. | Food / meals | 30 | 26 | 10 |
| 10. | Weigh / length / volume | 12 | 54 | 24 |
| 11. | Sound / fragrance / tempera-ture / light | 7 | 2 | 22 |
| 12. | Wildlife / celestial objects | 70 | 102 | 80 |
| 13. | Actions / changes /movement | 70 | 102 | 70 |
| 14. | Time | 119 | 62 | 86 |
| 15. | Clothes | 24 | 20 | 28 |
| 16. | Shape/ structure | 3 | 4 | 4 |
| 17. | Speech | 45 | 32 | 54 |
| 18. | Building / premises | 118 | 222 | 112 |
| 19. | Profession | 21 | 36 | 24 |
| 20. | Materials /liquids | 12 | 12 | 20 |
| 21. | Vehicles | 24 | 34 | 10 |
| 22. | Geographical notions | 33 | 68 | 20 |
| 23. | Weapons | 8 | 4 | 2 |
| 24. | Events / holidays | 12 | 10 | 2 |
| 25. | Other notions | 15 | 6 | 6 |
| | Total | 1325 | 1384 | 1246 |

Table 2
Frequency of the lexical semantic groups of verbs

| | Lexical semantic word group | "A Handful of Dust" by E. Waugh | "Gulliver's Travels" by J. Swift | "The Adventures of Tom Sawyer" by M. Twain |
|---|---|---|---|---|
| 1. | Verbs of Motion/Removing | 148 | 148 | 202 |
| 2. | Verbs of Process, Change, Development | 24 | 28 | 26 |
| 3. | Verbs of Beginning/End of Action | 30 | 18 | 36 |
| 4. | Verbs of Physical Action | 113 | 84 | 132 |
| 5. | Engender Verbs | 31 | 48 | 30 |
| 6. | Destroy Verbs | 15 | 22 | 20 |
| 7. | Verbs of Successful/Unsuccessful Action Implementation | 2 | 0 | 6 |
| 8. | Verbs of Attempt | 3 | 4 | 8 |
| 9. | Verbs of Sound Emission | 3 | 0 | 6 |
| 10. | Verbs of Light Phenomena | 3 | 2 | 2 |
| 11. | Verbs of Temperature Phenomena | 0 | 2 | 0 |
| 12. | Verbs of Nature Phenomena | 4 | 0 | 4 |
| 13. | Verbs of Communication | 131 | 30 | 122 |
| 14. | Verbs of Moral Impact/Effect | 41 | 24 | 20 |
| 15. | Verbs of Social Activity | 23 | 22 | 30 |
| 16. | Position Verbs | 16 | 30 | 44 |
| 17. | Verbs of Existence | 271 | 174 | 184 |
| 18. | Modality Verbs | 91 | 70 | 42 |
| 19. | Verbs of Human Relations | 17 | 18 | 20 |
| 20. | Verbs of Reference | 23 | 50 | 30 |
| 21. | Verbs of Emotional Psychological Impact | 9 | 6 | 26 |
| 22. | Verbs of Ownership/Loss | 58 | 104 | 58 |
| 23. | Verbs of Physiological State | 7 | 32 | 12 |
| 24. | Verbs of Perception | 50 | 52 | 78 |
| 25. | Verbs of Mental Activity | 66 | 66 | 68 |
| 26. | Verbs of Subjective Assessment | 18 | 8 | 8 |
| 27. | Verbs of Emotional Psychological State | 39 | 22 | 52 |
| | Total | 1236 | 1064 | 1266 |

Table 3
Frequency of the lexical semantic groups of adjectives

| | Lexical semantic word group | "A Handful of Dust" by E.Waugh | "Gulliver's Travels" by J.Swift | "The Adventures of Tom Sawyer" by M.Twain |
|---|---|---|---|---|
| 1. | Traits of character/emotions | 59 | 57 | 75 |
| 2. | Physical/natural condition | 122 | 66 | 48 |
| 3. | Intellectual capacity | 7 | 0 | 0 |
| 4. | Appearance | 22 | 12 | 12 |
| 5. | Senses | 2 | 0 | 0 |
| 6. | Age/time | 42 | 33 | 42 |
| 7. | Temperature/sound | 6 | 3 | 18 |
| 8. | Shape/size | 48 | 120 | 87 |
| 9. | Flavour | 2 | 0 | 0 |
| 10. | Weight | 0 | 0 | 0 |
| 11. | Degree/intensity | 100 | 99 | 42 |
| 12. | Color | 22 | 18 | 12 |
| 13. | Actions done to the object | 67 | 75 | 93 |
| 14. | Positive evaluation | 118 | 63 | 63 |
| 15. | Evaluation of length/distance/ position of the object | 44 | 60 | 54 |
| 16. | Evaluation of value/function of the object | 68 | 93 | 81 |
| 17. | Material | 19 | 6 | 21 |
| 18. | Negative evaluation | 45 | 54 | 69 |
| | Total | 793 | 759 | 717 |

However, the determination of the usage frequency does not constitute the complete statistical analysis of the subject-matter. It remains unrevealed whether the usage frequency of the lexical semantic groups of nouns, verbs, adjectives and adverbs in the novels substantially exceeds some theoretically expected quantity. Therefore for a more reliable quantitative analysis of data presented in Tables 1, 2, 3, 4, the normal test for individual cells of a contingency table can be applied. The most widespread formula for the calculation of the z criterion is as follows:

$$z = \frac{n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}}{\sqrt{\frac{n_{i\cdot} n_{\cdot j} \left(n - n_{i\cdot}\right)\left(n - n_{\cdot j}\right)}{n^2 \left(n - 1\right)}}}$$

where $n_{ij}$ is the frequency in the i, j cell;
$n_i$ is a the sum of the $i^{th}$ row;
$n_j$ is a the sum of the $j^{th}$ column;
$n$ is a the sum of all frequencies in the given Table (1, 2, 3, 4).

Table 4
Frequency of the lexical semantic groups of adverbs

| | Lexical semantic word group | "A Handful of Dust" by E. Waugh | "Gulliver's Travels" by J. Swift | "The Adventures of Tom Sawyer" by M. Twain |
|---|---|---|---|---|
| 1. | Adverbs of time | 141 | 108 | 96 |
| 2. | Adverbs of repetition and frequency | 66 | 18 | 54 |
| 3. | Adverbs of place and direction | 66 | 63 | 75 |
| 4. | Adverbs of condition and consequence | 5 | 3 | 15 |
| 5. | Adverbs of manner | 153 | 135 | 111 |
| 6. | Adverbs of degree and quantity | 215 | 111 | 111 |
| 7. | Question adverbs | 49 | 0 | 12 |
| | Total | 695 | 438 | 474 |

The result of the calculation shows that the authors avoid some lexical semantic classes (if $z < -1.96$) and prefer another word groups in their novels (if $z > 1.96$). The rest of lexical semantic classes are considered to be neutral. We present the data indicating the calculation by P (preferred), A (avoided), N (neutral) lexical semantic groups.

Table 5
The realization of lexical semantic groups

| | **Lexical semantic word groups of verbs** | W* | S | T | **Lexical semantic word groups of nouns** | W | S | T |
|---|---|---|---|---|---|---|---|---|
| 1. | Verbs of Motion/Removing | A | N | P | Appearance /parts of the body | A | N | P |
| 2. | Verbs of Process, Change, Development | N | N | N | Feelings / emotions | A | N | N |
| 3. | Verbs of Beginning/End of Action: | N | N | N | Proper names / nicknames | P | A | N |
| 4. | Verbs of Physical Action | N | N | N | Establishments /groupings | A | P | N |
| 5. | Engender Verbs | N | P | N | Diseases / defects | N | P | N |
| 6. | Destroy Verbs | N | N | N | General notions of people / mythical characters | N | N | N |
| 7. | Verbs of Successful/ Unsuccessful Action Implementation | N | N | P | Devices / articles of furniture | N | N | N |
| 8. | Verbs of Attempt | N | N | N | Abstract notions | A | P | N |
| 9. | Verbs of Sound Emission | N | N | N | Food / meals | P | N | A |
| 10. | Verbs of Light Phenomena | N | N | N | Weigh / length / volume | A | P | N |
| 11. | Verbs of Temperature Phenomena | N | P | N | Sound / fragrance / temperature / light | N | A | P |
| 12. | Verbs of Nature Phenomena | N | N | N | Wildlife / celestial objects | A | N | N |
| 13. | Verbs of Com-munication | P | A | P | Actions / changes /movement | N | P | N |
| 14. | Verbs of Moral Impact/ Effect | P | N | A | Time | P | A | N |
| 15. | Verbs of Social Activity | N | N | N | Clothes | N | N | N |
| 16. | Position Verbs | A | N | P | Shape/ structure | N | N | N |
| 17. | Verbs of Existence | P | N | A | Speech | N | A | P |
| 18. | Modality Verbs | P | N | A | Building / premises | A | P | A |
| 19. | Verbs of Human Relations | N | N | N | Profession | N | N | N |
| 20. | Verbs of Reference | A | P | N | Materials /liquids | N | N | P |
| 21. | Verbs of Emotional Psychological Impact | N | A | P | Vehicles | N | P | A |
| 22. | Verbs of Ownership/Loss | A | P | A | Geographical notions | N | P | A |
| 23. | Verbs of Physiological State | A | P | N | Weapons | N | N | N |
| 24. | Verbs of Perception | A | N | P | Events / holidays | N | N | A |
| 25. | Verbs of Mental Activity | N | N | N | Other notions | P | N | N |
| 26. | Verbs of Subjective Assessment | P | N | N | | | | |
| 27. | Verbs of Emotional Psychological State | N | A | P | | | | |

Continuation (Table 5)

| | **Lexical semantic word groups of adjectives** | W | S | T | **Lexical semantic word groups of adverbs** | W | S | T |
|---|---|---|---|---|---|---|---|---|
| 1. | Traits of character/emotions | N | N | P | Adverbs of time | N | N | N |
| 2. | Physical/natural condition | P | N | A | Adverbs of repetition and frequency | N | A | P |
| 3. | Intellectual capacity | P | N | N | Adverbs of place and direction | A | N | P |
| 4. | Appearance | N | N | N | Adverbs of condition and consequence | A | N | P |
| 5. | Senses | N | N | N | Adverbs of manner | A | P | N |
| 6. | Age/time | N | N | N | Adverbs of degree and quantity | P | N | A |
| 7. | Temperature/sound | N | A | P | Question adverbs | P | A | N |
| 8. | Shape/size | A | P | N | | | | |
| 9. | Flavour | N | N | N | | | | |
| 10. | Weight | | | | | | | |
| 11. | Degree/intensity | P | P | A | | | | |
| 12. | Color | N | N | N | | | | |
| 13. | Actions done to the object | A | N | P | | | | |
| 14. | Positive evaluation | P | A | A | | | | |
| 15. | Evaluation of length/ distance/position of the object | N | N | N | | | | |
| 16. | Evaluation of value/ function of the object | A | N | N | | | | |
| 17. | Material | N | A | P | | | | |
| 18. | Negative evaluation | A | N | P | | | | |

* W = Waugh, S = Swift, T = Twain

Every author tends to display a particular way of his writing and concept verbalization. He tends to shape his vocabulary using high frequency lexical semantic word groups to create his language world view. Different authors may interpret the world differently. The language world view in every novel under study will appear to be different. This difference can consist in the authors' ability to expose concepts behind the words, their codified practices and linguistic habits. Table 5 shows the preferred word groups in the novels of the three writers. We state that there is a distinctive, rational language means by which writers get the readers recognize their intention to reflect their understanding of the world.

J. Swift's nouns apparently communicate a great deal to the readers. We may conclude that 8 dominant lexical semantic groups of nouns represent diverse concepts that correspond to the common theme of the adventure novels: *geographical notions, vehicles, building/premises, actions/changes/movement, weigh/length/volume, abstract notions, diseases/defects, establishments/groupings.* Awareness of his historically and to certain extend socially defined vocabulary gives hints to J. Swift's exposing the deep social meaning behind words. With the detailed description of the adventures the author tries to choose the appropriate words for the exact context of the concepts in the text. It is crucial to understand how widely J. Swift describes adventures in his book preferably using lexical semantic groups of verbs of *engender, temperature phenomena, reference, ownership/loss, physiological state.* Lexical se-mantic groups of adjectives help to realize how important are *shape/size* and *degree/intensity* for the author.

The investigation reveals that E. Waugh and M. Twain have dominant lexical semantic verb groups denoting different concepts of human relations. In E. Waugh's novel they are: *verbs of communication, verbs of moral impact/effect, verbs of subjective assessment.* In M. Twain's book they are*: verbs of communication, verbs of emotional psychological impact, verbs of emotional psychological state.* Perhaps such word groups are common in literary works, but they also seem to reflect in adventure novels the conventions of sentimental literature with the focus on the emotions and embodied feelings. The tone of the texts is set with the use of specific groups of adjectives conveying *physical/natural condition, intellectual capacity, degree/intensity, positive evaluation* in E. Waugh's novel; *traits of character/emotions, negative evaluation, actions done to the object, temperature/sound* in M. Twain's novel.

E. Waugh and M. Twain's language gives us a new perspective of adventure theme and allows us to look at it with social conditioning. Although the authors may still use the same exhausted words and vague terms like *love, hatred, friendship* to refer to human relations, to place these words in a refreshing context of adventure allows the authors to represent their socially inclined world view.

## *4. MODELLING OF LANGUAGE WORLD VIEW IN ADVENTURE NOVELS*

Using the representation of the most frequent preferred word groups in each book enables us to get a sense of totality of language world view

of an adventure novel. We drill down and compare the significance of every particular lexical semantic group by its correlation ties. The correlation analysis adapts the spectrum of techniques described by V. Levickij (2004: 80-83). Correlation indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation refers to the departure of two random variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of the data.

When comparing the correlation between two items, one item is called the "dependent" item and the other the "independent" item. The goal is to see if a change in the independent item will result in a change in the dependent item. The correlation coefficient can range between ±1.0. A coefficient of +1.0, a "perfect positive correlation," means that changes in the independent item will result in an identical change in the dependent item. A coefficient of -1.0, a "perfect negative correlation," means that changes in the independent item will result in an identical change in the dependent item, but the change will be in the opposite direction. A coefficient of zero means there is no relationship between the two items and that a change in the independent item will have no effect in the dependent item. A low correlation coefficient (less than ±0.10) suggests that the relationship between two items is weak or non-existent. A high correlation coefficient (closer to plus or minus one) indicates that the dependent variable will usually change when the independent variable changes. The direction of the dependent variable's change depends on the sign of the coefficient. If the coefficient is a positive number, then the dependent variable will move in the same direction as the independent variable; if the coefficient is negative, then the dependent variable will move in the opposite direction of the independent variable. The most widespread formula for the calculation of linear correlation is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where r – the coefficient linear correlation,

$x_i$ – the means of the first variable,

$y_i$ – the means of the second variable,

$\overline{x}$ – the average of the first variable,

$\overline{y}$ – the average of the second variable.

Table 6 shows the computation results. A critical index for r is 0,98. For the analysis we take only positive correlation index.

The results of the calculation allow us to determine linear relationship between lexical semantic word groups. Some lexical semantic groups appear to have several correlation ties with other word groups of the four parts of speech. Computing the correlation between all preferred high frequent lexical semantic groups of the three novels enables us to trace concept dependence in the conceptual world view of the author reflected in his language world view. Author's language world view assigns to verbalized broad concept patterns giving a full-scale lexical presentation of a concept in text.

The resulting structure of the author's language world view can be visualized by a nested graph. This graph is a labelled graph that represents concepts verbalized by lexical semantic classes. It shows the concepts, inscribed in the ovals, and the connection among them, represented by arrows. The ovals contain a name of the lexical semantic class that denotes a concept. The arrows link the ovals showing a positive correlation of lexical semantic classes.

*__Verbs__: I. Verbs of Motion/Removing. II. Engender Verbs. III. Verbs of Successful/Unsuccessful Action Implementation. IV. Verbs of Temperature Phenomena. V. Verbs of Communication. VI. Verbs of Moral Impact/Effect. VII. Position Verbs. VIII. Verbs of Existence. IX. Modality Verbs. X. Verbs of Reference. XI. Verbs of Emotional Psychological Impact. XII. Verbs of Ownership/ Loss. XIII. Verbs of Physiological State. XIV. Verbs of Perception. XV. Verbs of Subjective Assessment. XVI. Verbs of Emotional Psychological State.

__Nouns__: I. Appearance/parts of the body. II. Proper names/nicknames. III. Establishments/groupings. IV. Diseases/defects. V. Abstract notions. VI. Food/ meals. VII. Weigh/length/volume. VIII. Sound/fragrance/temperature/light. IX. Actions/changes/movement. X. Time. XI. Speech. XII. Building/premises. XIII. Materials/liquids. XIV. Vehicles. XV. Geographical notions.

__Adjectives__: I. Traits of character/emotions. II. Physical/natural condition. III. Intellectual capacity. IV. Temperature/sound. V. Shape/size. VI. Degree/intensity. VII. Actions done to the object. VIII. Positive evaluation. IX. Material. X. Negative evaluation

__Adverbs__: I. Repetition and frequency. II. Place and direction. III. Condition and consequence. IV. Manner. V. Degree and quantity. VI. Question adverbs.

Table 6

Coefficient of linear correlation of lexical semantic word classes

| | | Verbs | | | | | | | | | | | | | | | | Nouns | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | I* | II | III | IV | V | VI | VII | VIII | IX | X | XI | XII | XIII | XIV | XV | XVI | I | II | III | IV | V | VI |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| Verbs | I | 1 | | | | | | | | | | | | | | | | | | | | | |
| | II | -0.54 | 1 | | | | | | | | | | | | | | | | | | | | |
| | III | 0.94 | -0.79 | 1 | | | | | | | | | | | | | | | | | | | |
| | IV | -0.5 | 1 | -0.76 | 1 | | | | | | | | | | | | | | | | | | |
| | V | 0.43 | -0.99 | 0.7 | -1 | 1 | | | | | | | | | | | | | | | | | |
| | VI | -0.65 | -0.29 | -0.36 | -0.34 | 0.41 | 1 | | | | | | | | | | | | | | | | |
| | VII | 0.87 | -0.05 | 0.65 | 0 | -0.08 | -0.94 | 1 | | | | | | | | | | | | | | | |
| | VIII | -0.42 | -0.54 | -0.1 | -0.58 | 0.64 | 0.96 | -0.82 | 1 | | | | | | | | | | | | | | |
| | IX | -0.9 | 0.13 | -0.71 | 0.08 | 0 | 0.91 | -1 | 0.77 | 1 | | | | | | | | | | | | | |
| | X | -0.27 | 0.95 | -0.57 | 0.97 | -0.99 | -0.56 | 0.25 | -0.76 | -0.17 | 1 | | | | | | | | | | | | |
| | XI | 0.99 | -0.65 | 0.98 | -0.62 | 0.55 | -0.53 | 0.79 | -0.29 | -0.84 | -0.4 | 1 | | | | | | | | | | | |
| | XII | -0.5 | 1 | -0.76 | 1 | -1 | -0.34 | 0 | -0.58 | 0.08 | 0.97 | -0.62 | 1 | | | | | | | | | | |
| | XIII | -0.33 | 0.97 | -0.62 | 0.98 | -0.99 | -0.51 | 0.19 | -0.72 | -0.11 | 1 | -0.46 | 0.98 | 1 | | | | | | | | | |
| | XIV | 1 | -0.49 | 0.92 | -0.44 | 0.37 | -0.69 | 0.9 | -0.47 | -0.93 | -0.21 | 0.98 | -0.44 | -0.27 | 1 | | | | | | | | |
| | XV | -0.5 | -0.46 | -0.19 | -0.5 | 0.57 | 0.98 | -0.87 | 1 | 0.82 | -0.7 | -0.37 | -0.5 | -0.65 | -0.55 | 1 | | | | | | | |
| | XVI | 0.83 | -0.92 | 0.96 | -0.9 | 0.86 | -0.1 | 0.43 | 0.17 | -0.5 | -0.77 | 0.9 | -0.9 | -0.8 | 0.79 | 0.08 | 1 | | | | | | |
| Nouns | I | 0.7 | 0.22 | 0.42 | 0.27 | -0.35 | -1 | 0.96 | -0.94 | -0.94 | 0.5 | 0.59 | 0.27 | 0.45 | 0.74 | -0.97 | 0.17 | 1 | | | | | |
| | II | -0.01 | -0.84 | 0.32 | -0.86 | 0.9 | 0.77 | -0.51 | 0.91 | 0.43 | -0.96 | 0.13 | -0.86 | -0.94 | -0.07 | 0.87 | 0.56 | -0.72 | 1 | | | | |
| | III | -0.16 | 0.91 | -0.47 | 0.93 | -0.96 | -0.65 | 0.36 | -0.83 | -0.28 | 0.99 | -0.29 | 0.93 | 0.98 | -0.09 | -0.78 | -0.69 | 0.6 | -0.99 | 1 | | | |
| | IV | -0.63 | 0.99 | -0.85 | 0.99 | -0.97 | -0.19 | -0.16 | -0.45 | 0.24 | 0.92 | -0.73 | 0.99 | 0.94 | -0.58 | -0.36 | -0.96 | 0.12 | -0.77 | 0.87 | 1 | | |
| | V | 0.02 | 0.83 | -0.31 | 0.86 | -0.9 | -0.77 | 0.52 | -0.92 | -0.44 | 0.96 | -0.12 | 0.86 | 0.94 | 0.08 | -0.87 | -0.55 | 0.73 | -1 | 0.98 | 0.77 | 1 | |
| | VI | -0.98 | 0.37 | -0.87 | 0.33 | -0.25 | 0.78 | -0.94 | 0.58 | 0.97 | 0.08 | -0.95 | 0.33 | 0.14 | -0.99 | 0.65 | -0.7 | -0.82 | 0.2 | -0.03 | 0.47 | -0.21 | 1 |
| | VII | -0.24 | 0.95 | -0.54 | 0.96 | -0.98 | -0.58 | 0.28 | -0.78 | -0.2 | 1 | -0.37 | 0.96 | 1 | -0.18 | -0.72 | -0.75 | 0.53 | -0.97 | 1 | 0.91 | 0.97 | 0.05 |
| | VIII | 0.97 | -0.73 | 1 | -0.69 | 0.63 | -0.45 | 0.72 | -0.19 | -0.78 | -0.49 | 0.99 | -0.69 | -0.54 | 0.95 | -0.28 | 0.94 | 0.5 | 0.23 | -0.39 | -0.8 | -0.22 | -0.91 |

Continuation (Table 6)

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nouns | IX | -0.5 | 1 | -0.76 | 1 | -1 | -0.34 | 0 | -0.58 | 0.08 | 0.97 | -0.62 | 1 | 0.98 | -0.44 | -0.5 | -0.9 | 0.27 | -0.86 | 0.93 | 0.99 | 0.86 | 0.33 |
| | X | -0.09 | -0.79 | 0.24 | -0.82 | 0.86 | 0.82 | -0.58 | 0.94 | 0.51 | -0.94 | 0.05 | -0.82 | -0.91 | -0.15 | 0.91 | 0.49 | -0.78 | 1 | -0.97 | -0.72 | -1 | 0.28 |
| | XI | 0.81 | -0.93 | 0.96 | -0.91 | 0.88 | -0.08 | 0.41 | 0.2 | -0.48 | -0.78 | 0.88 | -0.91 | -0.82 | 0.77 | 0.1 | 1 | 0.14 | 0.58 | -0.71 | -0.97 | -0.57 | -0.68 |
| | XII | -0.54 | 1 | -0.79 | 1 | -0.99 | -0.29 | -0.05 | -0.54 | 0.13 | 0.96 | -0.65 | 1 | 0.97 | -0.49 | -0.46 | -0.92 | 0.23 | -0.84 | 0.91 | 0.99 | 0.83 | 0.37 |
| | XIII | 1 | -0.54 | 0.94 | -0.5 | 0.43 | -0.65 | 0.87 | -0.42 | -0.9 | -0.27 | 0.99 | -0.5 | -0.33 | 1 | -0.5 | 0.83 | 0.7 | -0.01 | -0.16 | -0.63 | 0.02 | -0.98 |
| | XIV | -0.91 | 0.84 | -1 | 0.81 | -0.76 | 0.27 | -0.58 | 0 | 0.65 | 0.64 | -0.96 | 0.81 | 0.69 | -0.88 | 0.1 | -0.99 | -0.34 | -0.41 | 0.55 | 0.89 | 0.4 | 0.82 |
| | XV | -0.71 | 0.98 | -0.9 | 0.97 | -0.94 | -0.08 | -0.26 | -0.35 | 0.34 | 0.87 | -0.8 | 0.97 | 0.9 | -0.66 | -0.26 | -0.98 | 0.01 | -0.7 | 0.81 | 0.99 | 0.69 | 0.56 |
| Adjectives | I | 0.99 | -0.62 | 0.97 | -0.59 | 0.52 | -0.57 | 0.81 | -0.32 | -0.86 | -0.36 | 1 | -0.59 | -0.42 | 0.99 | -0.41 | 0.88 | 0.62 | 0.1 | -0.25 | -0.7 | -0.08 | -0.96 |
| | II | -0.69 | -0.24 | -0.41 | -0.28 | 0.36 | 1 | -0.96 | 0.95 | 0.93 | -0.51 | -0.58 | -0.28 | -0.46 | -0.73 | 0.97 | -0.16 | -1 | 0.73 | -0.61 | -0.13 | -0.74 | 0.81 |
| | III | -0.5 | -0.46 | -0.19 | -0.5 | 0.57 | 0.98 | -0.87 | 1 | 0.82 | -0.7 | -0.37 | -0.5 | -0.65 | -0.55 | 1 | 0.08 | -0.97 | 0.87 | -0.78 | -0.36 | -0.87 | 0.65 |
| | IV | 0.98 | -0.69 | 0.99 | -0.65 | 0.59 | -0.49 | 0.76 | -0.24 | -0.81 | -0.45 | 1 | -0.65 | -0.5 | 0.97 | -0.33 | 0.92 | 0.55 | 0.18 | -0.34 | -0.76 | -0.17 | -0.93 |
| | V | 0.05 | 0.81 | -0.28 | 0.84 | -0.88 | -0.79 | 0.54 | -0.93 | -0.47 | 0.95 | -0.09 | 0.84 | 0.93 | 0.11 | -0.89 | -0.52 | 0.75 | -1 | 0.98 | 0.75 | 1 | -0.24 |
| | VI | -1 | 0.53 | -0.94 | 0.49 | -0.42 | 0.66 | -0.87 | 0.43 | 0.91 | 0.25 | -0.99 | 0.49 | 0.31 | -1 | 0.51 | -0.82 | -0.71 | 0.02 | 0.14 | 0.62 | -0.03 | 0.98 |
| | VII | 0.95 | -0.26 | 0.8 | -0.22 | 0.14 | -0.85 | 0.98 | -0.67 | -0.99 | 0.03 | 0.9 | -0.22 | -0.03 | 0.97 | -0.74 | 0.62 | 0.88 | -0.31 | 0.15 | -0.37 | 0.32 | -0.99 |
| | VIII | -0.5 | -0.46 | -0.19 | -0.5 | 0.57 | 0.98 | -0.87 | 1 | 0.82 | -0.7 | -0.37 | -0.5 | -0.65 | -0.55 | 1 | 0.08 | -0.97 | 0.87 | -0.78 | -0.36 | -0.87 | 0.65 |
| | IX | 0.6 | -1 | 0.83 | -0.99 | 0.98 | 0.22 | 0.12 | 0.47 | -0.2 | -0.93 | 0.71 | -0.99 | -0.95 | 0.55 | 0.39 | 0.95 | -0.15 | 0.79 | -0.88 | -1 | -0.79 | -0.44 |
| | X | 0.93 | -0.19 | 0.76 | -0.14 | 0.06 | -0.88 | 0.99 | -0.72 | -1 | 0.11 | 0.87 | -0.14 | 0.05 | 0.95 | -0.79 | 0.56 | 0.91 | -0.38 | 0.22 | -0.3 | 0.39 | -0.98 |
| Adverbs | I | 0.28 | -0.96 | 0.58 | -0.97 | 0.99 | 0.55 | -0.24 | 0.76 | 0.16 | -1 | 0.41 | -0.97 | -1 | 0.22 | 0.69 | 0.77 | -0.5 | 0.96 | -0.99 | -0.92 | -0.96 | -0.09 |
| | II | 0.97 | -0.73 | 1 | -0.69 | 0.63 | -0.45 | 0.72 | -0.19 | -0.78 | -0.49 | 0.99 | -0.69 | -0.54 | 0.95 | -0.28 | 0.94 | 0.5 | 0.23 | -0.39 | -0.8 | -0.22 | -0.91 |
| | III | 0.99 | -0.67 | 0.98 | -0.63 | 0.56 | -0.52 | 0.78 | -0.27 | -0.83 | -0.41 | 1 | -0.63 | -0.47 | 0.98 | -0.36 | 0.9 | 0.58 | 0.15 | -0.31 | -0.74 | -0.14 | -0.94 |
| | IV | -0.9 | 0.13 | -0.71 | 0.08 | 0 | 0.91 | -1 | 0.77 | 1 | -0.17 | -0.84 | 0.08 | -0.11 | -0.93 | 0.82 | -0.5 | -0.94 | 0.43 | -0.28 | 0.24 | -0.44 | 0.97 |
| | V | -0.5 | -0.46 | -0.19 | -0.5 | 0.57 | 0.98 | -0.87 | 1 | 0.82 | -0.7 | -0.37 | -0.5 | -0.65 | -0.55 | 1 | 0.08 | -0.97 | 0.87 | -0.78 | -0.36 | -0.87 | 0.65 |
| | VI | -0.28 | -0.65 | 0.05 | -0.69 | 0.75 | 0.91 | -0.72 | 0.99 | 0.67 | -0.85 | -0.15 | -0.69 | -0.81 | -0.34 | 0.97 | 0.31 | -0.88 | 0.96 | -0.9 | -0.57 | -0.96 | 0.46 |

Continuation (Table 6)

| | | Nouns | | | | | | | | | Adjectives | | | | | | | | | | Adverbs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | VII | VIII | IX | X | XI | XII | XIII | XIV | XV | I | II | III | IV | V | VI | VII | VIII | IX | X | I | II | III | IV | V | VI |
| 1 | 2 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 |
| Verbs | I | | | | | | | | | | | | | | | | | | | | | | | | | |
| | II | | | | | | | | | | | | | | | | | | | | | | | | | |
| | III | | | | | | | | | | | | | | | | | | | | | | | | | |
| | IV | | | | | | | | | | | | | | | | | | | | | | | | | |
| | V | | | | | | | | | | | | | | | | | | | | | | | | | |
| | VI | | | | | | | | | | | | | | | | | | | | | | | | | |
| | VII | | | | | | | | | | | | | | | | | | | | | | | | | |
| | VIII | | | | | | | | | | | | | | | | | | | | | | | | | |
| | IX | | | | | | | | | | | | | | | | | | | | | | | | | |
| | X | | | | | | | | | | | | | | | | | | | | | | | | | |
| | XI | | | | | | | | | | | | | | | | | | | | | | | | | |
| | XII | | | | | | | | | | | | | | | | | | | | | | | | | |
| | XIII | | | | | | | | | | | | | | | | | | | | | | | | | |
| | XIV | | | | | | | | | | | | | | | | | | | | | | | | | |
| | XV | | | | | | | | | | | | | | | | | | | | | | | | | |
| | XVI | | | | | | | | | | | | | | | | | | | | | | | | | |
| Nouns | I | | | | | | | | | | | | | | | | | | | | | | | | | |
| | II | | | | | | | | | | | | | | | | | | | | | | | | | |
| | III | | | | | | | | | | | | | | | | | | | | | | | | | |
| | IV | | | | | | | | | | | | | | | | | | | | | | | | | |
| | V | | | | | | | | | | | | | | | | | | | | | | | | | |
| | VI | | | | | | | | | | | | | | | | | | | | | | | | | |
| | VII | 1 | | | | | | | | | | | | | | | | | | | | | | | | |
| | VIII | -0.47 | 1 | | | | | | | | | | | | | | | | | | | | | | | |

Continuation (Table 6)

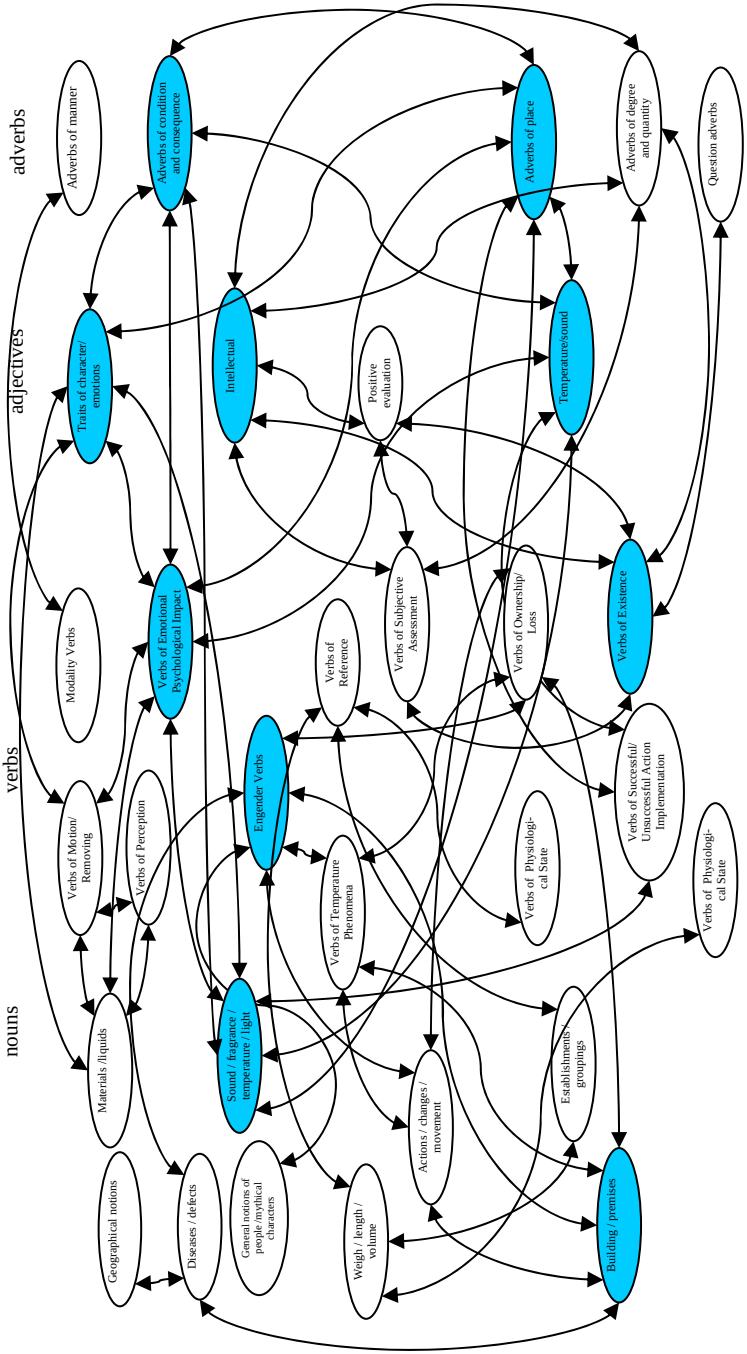
| 1 | 2 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nouns | IX | 0.96 | -0.69 | 1 | | | | | | | | | | | | | | | | | | | | | | |
| | X | -0.94 | 0.15 | -0.82 | 1 | | | | | | | | | | | | | | | | | | | | | |
| | XI | -0.76 | 0.93 | -0.91 | 0.51 | 1 | | | | | | | | | | | | | | | | | | | | |
| | XII | 0.95 | -0.73 | 1 | -0.79 | -0.93 | 1 | | | | | | | | | | | | | | | | | | | |
| | XIII | -0.24 | 0.97 | -0.5 | -0.09 | 0.81 | -0.54 | 1 | | | | | | | | | | | | | | | | | | |
| | XIV | 0.62 | -0.98 | 0.81 | -0.33 | -0.98 | 0.84 | -0.91 | 1 | | | | | | | | | | | | | | | | | |
| | XV | 0.85 | -0.86 | 0.97 | -0.64 | -0.99 | 0.98 | -0.71 | 0.94 | 1 | | | | | | | | | | | | | | | | |
| Adjectives | I | -0.34 | 0.99 | -0.59 | 0.01 | 0.86 | -0.62 | 0.99 | -0.95 | -0.78 | 1 | | | | | | | | | | | | | | | |
| | II | -0.54 | -0.49 | -0.28 | 0.79 | -0.13 | -0.24 | -0.69 | 0.33 | -0.02 | -0.61 | 1 | | | | | | | | | | | | | | |
| | III | -0.72 | -0.28 | -0.5 | 0.91 | 0.1 | -0.46 | -0.5 | 0.1 | -0.26 | -0.41 | 0.97 | 1 | | | | | | | | | | | | | |
| | IV | -0.42 | 1 | -0.65 | 0.1 | 0.91 | -0.69 | 0.98 | -0.97 | -0.83 | 1 | -0.54 | -0.33 | 1 | | | | | | | | | | | | |
| | V | 0.96 | -0.19 | 0.84 | -1 | -0.55 | 0.81 | 0.05 | 0.37 | 0.67 | -0.05 | -0.76 | -0.89 | -0.14 | 1 | | | | | | | | | | | |
| | VI | 0.23 | -0.97 | 0.49 | 0.11 | -0.8 | 0.53 | -1 | 0.9 | 0.7 | -0.99 | 0.7 | 0.51 | -0.98 | -0.06 | 1 | | | | | | | | | | |
| | VII | 0.06 | 0.85 | -0.22 | -0.39 | 0.6 | -0.26 | 0.95 | -0.74 | -0.46 | 0.92 | -0.87 | -0.74 | 0.88 | 0.35 | -0.96 | 1 | | | | | | | | | |
| | VIII | -0.72 | -0.28 | -0.5 | 0.91 | 0.1 | -0.46 | -0.5 | 0.1 | -0.26 | -0.41 | 0.97 | 1 | -0.33 | -0.89 | 0.51 | -0.74 | 1 | | | | | | | | |
| | IX | -0.92 | 0.78 | -0.99 | 0.74 | 0.96 | -1 | 0.6 | -0.88 | -0.99 | 0.68 | 0.16 | 0.39 | 0.74 | -0.77 | -0.59 | 0.33 | 0.39 | 1 | | | | | | | |
| | X | 0.14 | 0.81 | -0.14 | -0.45 | 0.53 | -0.19 | 0.93 | -0.69 | -0.4 | 0.89 | -0.91 | -0.79 | 0.84 | 0.42 | -0.93 | 1 | -0.79 | 0.26 | 1 | | | | | | |
| Adverbs | I | -1 | 0.5 | -0.97 | 0.93 | 0.79 | -0.96 | 0.28 | -0.65 | -0.87 | 0.37 | 0.51 | 0.69 | 0.45 | -0.95 | -0.26 | -0.02 | 0.69 | 0.93 | -0.1 | 1 | | | | | |
| | II | -0.47 | 1 | -0.69 | 0.15 | 0.93 | -0.73 | 0.97 | -0.98 | -0.86 | 0.99 | -0.49 | -0.28 | 1 | -0.19 | -0.97 | 0.85 | -0.28 | 0.78 | 0.81 | 0.5 | 1 | | | | |
| | III | -0.39 | 1 | -0.63 | 0.07 | 0.89 | -0.67 | 0.99 | -0.96 | -0.81 | 1 | -0.57 | -0.36 | 1 | -0.11 | -0.99 | 0.9 | -0.36 | 0.72 | 0.86 | 0.42 | 1 | 1 | | | |
| | IV | -0.2 | -0.78 | 0.08 | 0.51 | -0.48 | 0.13 | -0.9 | 0.65 | 0.34 | -0.86 | 0.93 | 0.82 | -0.81 | -0.47 | 0.91 | -0.99 | 0.82 | -0.2 | -1 | 0.16 | -0.78 | -0.83 | 1 | | |
| | V | -0.72 | -0.28 | -0.5 | 0.91 | 0.1 | -0.46 | -0.5 | 0.1 | -0.26 | -0.41 | 0.97 | 1 | -0.33 | -0.89 | 0.51 | -0.74 | 1 | 0.39 | -0.79 | 0.69 | -0.28 | -0.36 | 0.82 | 1 | |
| | VI | -0.86 | -0.04 | -0.69 | 0.98 | 0.34 | -0.65 | -0.28 | -0.14 | -0.48 | -0.18 | 0.89 | 0.97 | -0.1 | -0.97 | 0.3 | -0.56 | 0.97 | 0.6 | -0.62 | 0.84 | -0.04 | -0.13 | 0.67 | 0.97 | 1 |

As we see from the graph, a certain lexical semantic class is marked by correlation ties with several lexical semantic classes. The dependence of other lexical semantic classes on this very class permits to consider it a dominant verbalized concept in the language world view. For example, the biggest cluster of correlated classes highlights a concept of *emotional psychological impact.* The lexical semantic class, representing this concept, is found in linear relationship with 6 other classes: nouns denoting sound/fragrance/temperature/light; materials/ liquids; the verbs of motion/removing; adjectives of temperature/sound; traits of character and the adverbs of condition/consequence; place/ direction. Another dominant concept of *sound/fragrance/temperature/ light* has the same number of connections. The lexical semantic class that reflects this concept shows correlation with the nouns of action/ change/movement, the verbs of engender; temperature phenomena, the adverns of condition/consequence; the adjectives of temperature/sound.

Such an approach to the dominant concept determination makes explicit in which way the authors' concept hierarchy in the adventure novels is built. Thus, the concept hierarchy (including first ten dominant concepts) of the novels under study appears to be as follows:

- a concept of *emotional psychological impact* (6 connections);
- concept of *sound/fragrance/temperature/light* (6 connections);
- a concept of *place and direction* (6 connections);
- a concept of *engender* (6 connections);
- a concept of *traits of character/emotions* (6 connections);
- a concept of *condition and consequence* (5 connection);
- a concept of *temperature/sound* (5 connections);
- a concept of *existence* (5 connections);
- a concept of *building/premises* (5 connections);
- a concept of *intellectual capacity* (5 connections).

The resulting concept hierarchy reveals the importance of human relations, emotions, intellect as the basic notions of conceptual world view of the adventure story writers alongside the concepts of nature phenomena that is stipulated by the plot of the novels.




Graph 1. Model of the language world view in adventure novels

## *CONCLUSION*

The concept analysis based on objective statistical data of lexical semantic word class realization in the text explores how concept-word relations are a factor of language world view creation. We conclude that the word meaning corresponds to the concept held in the mind of the author, which is based on his personal understanding of the world.

For applying and elaborating the discussed method of concept verbalization analysis, it is worthwhile not only to work in the frame of philosophical logic, but also on the level of factual data activating quantitative methods.

## *REFERENCES*

Altmann, G. (1996): The nature of linguistic units, in: *Journal of Quantitative Linguistics* 3, 1-7.

Croft, W.; Cruse, D. Alan (2004): *Cognitive linguistics*. New York: Cambridge University Press.

James, W. (1975): *The Works of William James*. (eds.) Frederick H. Burkhardt; Fredson Bowers and Ignas K. Skrupskelis, Harvard University Press, Cambridge, MA.

Langacke, R. (1990): *Concept, image and symbol: The cognitive basis grammar*. Berlin: Mouton de Gruyter.

Levickij, V. (2004): *Quantitative methods in linguistics*. Chernivtsi: Ruta.

Schwarz, M. (1992): Kognitive *Semantiktheorie und neuropsychologische Realitaet*. Tuebingen: Niemeyer.

Whorf, B. (1956): *Language, thought and reality*. New-York: Cambridge University Press.

Wierzbicka, A. (1992): *Semantics, culture and cognition. Universal human concepts in culture specific configurations*. N.Y., Oxford.

Wierzbicka, A. (1985): *Lexicography and conceptual analysis*. Ann Arbor.

# Linguistics as a diagnostic tool for textual ineptness: Narrative perspective in Zamjatin's *Navodnenie*[*]

*Olga T. Yokoyama (Los Angeles, USA)*

*Здесь интегральный образ наводнения я попытался провести через рассказ в двух планах: реальное петербургское наводнение*[1] *отражено в наводнении душевном.*
*Е. Замятин*

I attempted to integrate the image of flooding throughout this story on two levels, mirroring the real St. Petersburg flood in a psychological flood.
E. Zamjatin

## *0. INTRODUCTION*

The power of linguistics as a tool for literary analysis was demonstrated by the Formalists[2] and the Structuralists[3]. It was often pointed out, however, that even when linguistics provided convincing analytical approaches for revealing the structure of a given text and "laying bare" its poetic methods, it never succeeded in establishing linguistic criteria for judging the quality of a given literary text. In fact, evaluation has never been one of the stated goals of linguistic approaches to literary texts. If value judgments regarding a piece of literature are made, these rely on "macro" features, such as the historical, cultural, psychological, or

[*] I thank Allen MacDuffie (University of Texas, Austin) for helpful discussion of some of the literary aspects of this paper.
[1] The worst St. Petersburg flood in a century occurred in 1924; the water level reached 112 feet above normal.
[2] Among the classic works that laid the foundations of formalist literary analysis, see in particular Ejxenbaum 1927.
[3] See numerous works by Jakobson, e.g., those collected in 1987a.

ideological significance of the work, or on its ability to grip, amuse, or move the reader. Literary value judgments do not rest on the examination of a text's linguistic or poetic structure. I will argue, however, that linguistics does possess the tools needed for detecting at least some problem spots in literary texts and for explaining their structure and the causes of their infelicity.

Linguists generally operate under the assumption that all utterances naturally produced by native speakers — whether felicitous or not, and even those including spoonerisms, attraction errors, and other such problems — constitute legitimate linguistic data worth analyzing, because they are the product of speakers' linguistic faculty.[4] Apart from performance errors like spoonerisms, however, descriptive linguists[5] expect no unacceptable utterances to be produced in natural native discourse. The fact that an utterance was produced in the first place is usually taken as evidence of its acceptability in a given language or dialect. This is, however, more true of spoken language (with its reliance on the control of phonology) than of written language, and of sentence grammar (with its reliance on the control of case, tense, agreement, government, etc.) rather than the grammar of text (with its reliance on discourse pragmatic communicative skills). If native speakers produce infelicities in their native language, the place to look for them, as every teacher of writing knows, is in written texts. These texts are both less spontaneous and more extensive than oral "texts". They require the speaker's/writer's well developed communicative ability to sustain the assessment[6] of the addressee's/reader's knowledge sets and to maintain an appropriate degree of continuous attention to the cognitive process of communicative transactions that take place between the two interlocutors. Moreover, since the addressee/reader can neither acknowledge[7] the transaction nor adjust the speaker's/writer's misassessments[8], written texts do not benefit from the kind of corrective input

[4] Even non-native interlanguage production, or non-natural so-called "starred" sentences, are used in linguistic analysis in many important ways.

[5] As opposed to normative linguists, whose role is to check native production against normative rules, which not infrequently lack linguistic justification; cf., e.g., Yokoyama 2006.

[6] In the sense of Yokoyama 1986, 44 ff., or Jokojama 2005, 80 ff.

[7] For the concept of acknowledgement, see Yokoyama 1986, 52-53 or Jokojama 2005, 89-91.

[8] For more on misassessment and its adjustment, see Yokoyama 1986, 53-59 or Jokojama 2005, 91-97.

available in synchronic interaction. It is in the production of such texts that a native speaker's communicative competence frequently fails.

Literary texts belong among the least spontaneous of linguistic productions, generated as they are on the basis of considerable deliberation. In contrast with deliberately produced written texts oriented towards conveying information or building an argument, literary texts are also special in that they focus heavily on the form of the message.[9] It may, for these reasons, be tempting to expect them (with the exception of embedded mimetic non-native speech or *skaz*-type narration) to lack the infelicities found in informational or non-artistic written texts. This expectation, however, rests on the tacit assumption that literary texts represent final, ultimate products, and that their every feature flows from considered authorial judgment, is intentional, and results from a highly developed communicative skill. It is easy to see that such an assumption is amply contradicted by the sheer existence of multiple authorial drafts[10] and from biographical facts surrounding authors working under the pressure of deadlines and under other conditions not conducive to perfect performance. Then there is the question of skill, especially of the discourse pragmatic competence of maximal assessment and of imposition[11] within the limits of the author/reader contract.[12] A search for grammatical errors in the areas of phonology (insofar as it can be decoded from writing), case marking, government, agreement, tense, or aspect is

[9] Cf. Jakobson's definition of the poetic function as *Einstellung* 'orientation' towards the message itself (Jakobson 1987, 69 ff.). This orientation may be more or less focused on certain linguistic components (e.g., poetry is more concerned with phonetics and phonology than prose) and may be more or less pronounced in different literary genres; but in all literary production the orientation towards form and structure is greater than in other kinds of linguistic production.

[10] See Zaitseva 1993 for a sensitive analysis of the changes undertaken by Dostoevsky between his 1st person draft of *Crime and Punishment* and its ultimate 3rd person narrative.

[11] On imposition, see Yokoyama 1986, 59-66, or Jokojama 2005, 97-104.

[12] On various kinds of interlocutor contracts, see Yokoyama 1986, 144 ff. and passim or Jokojama 2005, 188 ff. and passim. Contracts between the author and the reader of literary creations are a function of time and culture, and the ability to follow these contracts belongs to the literary and not simply communicative competence of the "interlocutors". Innovative authors implicitly proclaim new author-reader contracts by creating literature that practices new contract types as yet unknown to the culture, which then for a time result in what Barthe calls "unreadable" texts (Barthes 1970, 10).

not likely to yield results in literary texts; yet, for the reason just mentioned, non-optimal authorial decisions in the area of discourse pragmatics are always a possibility. As such, these non-optimal decisions can affect both the selection and the combination process of literary creation.[13]

Infelicities are perceptually detectable by the "competent" reader.[14] By studying loci where a jarring effect is encountered in a literary work and by explaining them in systematic terms using independently existing discourse linguistic categories, linguistics can make explicit just what it is that the reader perceives as jarring. Even though Structuralist poetics never aspired to the production of value judgments, I will show below that by adopting some of the methods arising from two developments in linguistics in recent decades, it is possible to begin to speak of linguistic methods doing precisely that, i.e., making it possible to pass judgment on a text's quality.

That said, the evaluation of artistic literary creation cannot be equated with the evaluation of a literary text's discourse-pragmatic features. Other characteristics peculiar to deliberate literary texts must also be taken into consideration, if the goal is to elevate textual evaluation to the higher plane of literary criticism. I return to this point in the final section of this paper.

The two linguistic tools shown in action here are Discourse Grammar and Genderlinguistics. Specifically, I will argue that Discourse Grammar allows us to identify point of view in a narrative and, by doing so, to detect infelicitous shifts in point of view — a potential sign of compromised quality in writing. I will also argue that active metaphors, as one type of poetic device capable of reflecting a gendered perspective, may also be inept when used in violation of that perspective, thus rendering a text unconvincing, and thereby compromising its quality in a different way.

The text chosen for this qualitative study is E.I. Zamjatin's little known short story *Navodnenie* (*The Flood*, 1929). It tells the story of a married woman named Sofya, taking her from infertility through pregnancy and childbirth, while her rival is eliminated along the way. Most of the narrative is presented from Sofya's point of view, often a limited one, although shifts to her husband Trofim Ivanovich's equally limited

[13] The reference here is to Jakobson's often-quoted description of the poetic function: "The poetic function projects the principle of equivalence from the axis of selection into the axis of combination" (Jakobson 1987, 71).

[14] Cf. Culler's characterization of such a reader (Culler 1975, 113-130, and passim).

perspective occasionally occur as well. The first kind of infelicity, I argue, appears in a few cases of inept point of view shifts, and the second in passages featuring the author's ambitious attempt to capture female sensibility and physical sensations through two metaphors.[15] On the whole, the story is "readable" in the Barthesian sense. It contains considerable mimetic detail, is chronologically structured and tightly crafted, and is rich in symbolism and thematic code. Problematic point of view shifts and metaphors will be described in sections 1 and 2, respectively, followed by discussion in section 3.

## *1. POINT OF VIEW SHIFTS REVEALED THROUGH KNOWLEDGE AND PERCEPTION OWNERSHIP*

Point of view in a text can be established on the basis of the "ownership" of the knowledge provided in the text and the identity of the experiencer of the perceptions provided in it. Thus in (1), the perceptions, the thoughts, and the judgments belong to Sofya (S):

(1) […] ona uvidala v pustom okne svet. Ona ostanovilas': ne možet byt'!
she saw in empty window light she stopped not can be
Vernulas' nazad, zagljanula v dyru okna. Vnutri, sredi oblomkov kirpiča,
returned back peeked in hole of-window inside amidst fragments of-brick
gorel koster, vokrug nego sidelo četvero otrepyšej-mal'čišek. […] Odin, licom k
burned bonfire around it sat four ragged boys one facing to
Sof'e, černoglazyj, dolžno byt' cyganenok, pripljasyval, na goloj grudi u nego
S. black-eyed must be Gypsy-kid danced on bare chest at him
prygal serebranyj krestik, zuby blesteli. Pustoj dom stal živym. Cyganenok
bounced silver cross teeth glistened empty house came alive Gypsy-kid
čem-to poxodil na Trofima Ivanyča. Sof'ja vdrug
somehow resembled at T. I. S. suddenly
počuvstvovala, čto […] (sec. 2)
felt that

'[…] she saw light in the paneless window. She stopped — it can't be! She backed up and peeked into the window hole. Inside, amidst brick fragments,

[15] Some of the presentation in this paper refers to things that would ordinarily be unmentionable in polite society; but as the Russian proverb has it, *Iz pesni slova ne vykineš* 'You can't toss a word out of a poem': in order to analyze the material of this story and to make the argument presented here, explicit language cannot be avoided.

a bonfire was burning and four boys clad in rags sat around it, […] One boy, who was facing Sofya, with black eyes — probably a Gypsy kid — was dancing, a little silver crucifix bouncing on his bare chest, his teeth shining. The empty house had come alive. The Gypsy kid looked a little like Trofim Ivanych. Sofya suddenly felt that […]'

The unity of the perspective in this segment is unambiguous. *Ona* 'she' is the subject of the verbs *uvidala* 'saw', *ostanovilas'* 'stopped', *vernulas'* 'returned', *zagljanula* 'peeked', and *počuvstvovala* 'felt'; the first four describe the steps that led S to observing the scene in the empty house, followed by the description of what she saw, and the fifth concludes the experience by describing S's reaction to it. Following *ostanovilas'* 'stopped', her initial reaction to the first glimpse of the scene she had caught is rendered through *erlebte Rede* (*ne možet byt'!*), verbalizing S's internal monologue. The longer description of the scene that follows *zagljanula* shows S's perspective not only physically (*licom k Sof'e*) but also cognitively: the conjecture that the homeless boy must be a Gypsy and the judgment that he looked a little like TI both clearly reside in S's "knowledge sets".[16] The whole passage is consistent and convincing.

This fragment contrasts with the following, where the shift in point of view that occurs towards the end of (2) catches the reader by surprise, causing disorientation:

(2) Krugom Vasil'evskogo Ostrova dalekim morem ležal mir: tam byla
around V. Island as-far-away sea lay world there was
vojna, potom revoljucija. A v kotel'noj u Trofima Ivanyča kotel gudel vse
war then revolution & in boiler-room at T. I. boiler buzzed all
tak že, manometer pokazyval vse te že devjat' atmosfer. Tol'ko ugol'
the-same manometer showed all the-same nine atmospheres only coal
pošel drugoj: byl kardif, teper' – doneckij. Ètot krošilsja, černaja pyl'
came different was Cardiff now Don this was-crumby black dust
zalezala vsjudu, ee bylo ne otmyt' ničem. Vot budto èta že černaja
got everywhere it was not to-wash-off by-anything just as-if the-same black
pyl' neprimetno obvolokla vse i doma. Tak, snaruži,
dust imperceptibly wrapped-around all too at-home like-that from-outside
ničego ne izmenilos'. Poprežnemu žili vdvoem, bez detej. Sof'ja,
nothing not changed as-before they-lived twosome without children S.

[16] For a cognitive model of knowledge, see Yokoyama 1986, 3-42 or Jokojama 2005, 27-77.

xot' bylo ej uže pod sorok, byla vse tak že legka, stroga vsem
although was to-her already close-to 40 was all the-same light prim in-all

telom, kak ptica, ee budto dlja vsex navsegda sžatye guby poprežnemu
body like bird her as-if for all forever pressed-tight lips as-before
raskryvalis' Trofimu Ivanyču noč'ju – i vse-taki bylo ne to. Čto "ne to" –
opened to-T. I. at-night & yet was wrong what wrong
bylo ešče nejasno, ešče ne otverdelo v slovax. Slovami èto v pervyj raz
was yet unclear yet not set in words by-words this for 1st time
skazalos' tol'ko pozže, osen'ju, i Sof'ja zapomnila: èto bylo
verbalized only later in-fall & S. remembered it was
noč'ju v subbotu, byl veter, voda v Neve podnimalas'. (sec. 1)
at-night on Saturday was wind water in N. was-rising

'Far around Vasilyev Island was the sea of the world. There was war there, then the revolution. But in Trofim Ivanych's boiler room, the boiler buzzed as ever, and the manometer showed the same nine atmospheres of pressure, as always. Only the coal was different now: it used to be Cardiff, now it was Don. This one was crumby, its black dust would get into everything, it was impossible to wash it off with anything. It was just as if this same black dust had sneakily covered everything at home as well. From outside, nothing had changed. They still lived as before, just the two of them, without kids. Sofya, although she was already almost forty, was still just as light and prim in her body as a bird, her lips, which seemed shut tight for anybody, would still open to Trofim Ivanych at night; and yet something was wrong. What exactly was wrong was still unclear, it had not yet been set in words. It found its way into words for the first time only later, in the fall, and Sofya remembered: it was on a Saturday night, it was windy, and the water in the Neva River was rising.'

In this long passage, with which the story begins, the perspective belongs to TI until the appearance of S halfway through the last sentence. The conjunction *a* at the beginning of the second sentence shows that the following coordinated sentence is based on a two member set of referential knowledge associated by contrast[17] and consisting of: the world out there (*tam*) and TI's boiler room (*v kotel'noj u Trofima Ivanyča*). This opposition immediately and unambiguously places the reader into TI's cognitive sphere, forming a Chekhovian-style *erlebte Rede*, where the perceptions and thought content belong to the not-necessarily-literate character, but the language is that of the literate narrator.[18] The details

[17] See Yokoyama 1986, 312-326 or Jokojama 2005, 374-393 on discourse features of contrast in Russian.
[18] See Mikhaychuk 1994.

about the quality of the coal continue this narrative perspective, since burning coal is what TI's job was all about. The associative link from coal to the coal dust that fills TI's daily life and which, frustratingly, refuses to be scrubbed off (note the impersonal construction with deleted experiencer[19] in the dative case: *ee bylo ne otmyt' ničem*) shows that the following metaphorical grey film is part of TI's perception of his family life as well: the parallelism between the black dust at work (*černaja pyl' zalezala vsjudu*) and the black dust at home (*èta že černaja pyl' neprimetno obvolokla vse i doma*) is built on a two-member contrastive set {(everywhere at) work, at home}, clearly part of TI's experience residing in his knowledge set. The knowledge that to an outsider, nothing had changed (*ničego ne izmenilos'*) can only belong to TI, an insider who knows what things were like at home before, and how they have or have not changed. The deletion of the subject of the verb *žili* 'lived' shows that the subject is assumed by the narrator to be highly recoverable. A first person subject is the most likely candidate for such deletions. In this case, it is the first person plural subject coreferent with TI and S, although the referential knowledge[20] of S, who has not yet been introduced into the narrative, is impositional. In fact, at this juncture the *erlebte Rede* becomes more evident than before. It continues to build up as the text provides TI's knowledge of his wife S's body now and before, while comparing it to a bird's (*byla vse tak že legka, stroga vsem telom, kak ptica*). Only TI can have this knowledge and it is his perception of S as a bird that is expressed here. The next phrase about S's lips now and before (*guby poprežnemu raskryvalis' Trofimu Ivanyču*) reflects TI's intimate experience of past and present. This series of TI's perceptions and experiences culminates in TI's subjective reaction to them, i.e., that something was wrong, despite the seeming permanency (*i vse-taki bylo ne to*). The very inability to find the right words to express what was wrong, an inability overcome only later in the autumn, constitutes TI's experi-

[19] It could be argued that the deleted dative constituent is the underlying subject of the infinitive. The impersonal phrase, however, is strongly subjective and implies that the person who experiences the frustration is unable to wash off the dust. The choice between calling the dative constituent the experiencer or the underlying subject is in any case irrelevant to the reference of this constituent, which is unambiguously TI.

[20] For the concept of referential knowledge, see Yokoyama 1986, 7-11 and passim or Jokojama 2005, 31-36 and passim.

ence, as the impersonal *skazalos'* 'got verbalized' indicates: the most easily deleted experiencer is first person singular, coreferent with TI.[21]

The sudden shift to S's knowledge set in the next clause takes the reader by surprise. The verb of mental activity *zapomnila* 'remembered' coming after the subject *Sof'ja* and the following specification of the content of her memories resides in S's cognitive sphere and creates a starkly disjunctive effect. To make things worse, immediately following this passage, the next paragraph begins with a description of a burst pipe on TI's boiler, leading the reader back to TI's experience in the shop, as described in (3) below. Surrounded in this way by extensive pieces of text narrated from TI's perspective, mostly expressed in a very strong form of *erlebte Rede* perspective-building, the short-lived switch to S's knowledge set is highly disconcerting and clearly infelicitous.

Inept shifts like this one in (2) are rare in *The Flood*. Besides the very pronounced case of (2), as just examined, there is one more passage in section 6 of the story, where several mildly confusing shifts between TI's and S's points of view occur within a span of four paragraphs.

## *2. POINT OF VIEW REVEALED THROUGH ACTIVE GENDERED METAPHORS*

Metaphors substitute one concept with another based on perceived similarity. I will call the substituted concept the *designatum* and the concept that is used in its place the *designans*.[22] An active metaphor is an individualized substitution of one concept with another, where, unlike

[21] The description of that moment when words were finally found appears three paragraphs later in the same section:

On ležal, steklo ot vetra pozvjakivalo odnoobrazno. Vdrug vspomnilos': […] "Ono samoe", vslux skazal T.I. […] "Detej ty ne rožaeš', vot čto". 'He was lying down, the window glass was monotonously clanking from the wind. Suddenly it came back to him. […] "I know what", said T.I. out loud. […] "You don't have kids, that's what".

Note the impersonal *vspomnilos'* 'it came back (to him)', echoing the impersonal *skazalos'* 'got verbalized' in (2).

[22] Cf. Richards' corresponding terms *tenor* and *vehicle.* My terms, at the risk of multiplying near-synonymous terminology, are motivated by the consideration that the *designans* is nothing other than a new lexical designation for the *designatum*. Active metaphors thus create ad hoc *signifier — signified* pairs.

"dead" metaphors, the *designatum* and the *designans* (pl. *designantia*) are not regularly related to one another in a given speech community. Active metaphors are thus not only limited to a single speaker (i.e., are "idiolectal"), but they are also usually limited within the scope of an individual's usage to a single speech event (including extended "speech events", like a long conversation or a piece of literature). Zemskaja, Kitajgorodskaja, and Rozanova (1993, 127-129) have noted that male and female speakers of Russian tend to employ different (active) metaphors: "male" metaphors abound in *designantia* from the "male" semantic fields of sports, hunting, the military, and technology, while female *designantia* are taken from household-related and common encyclopedic fields such as animals and nature. The *designantia* evoked by males and females are thus associated with their respective life experiences and standing concerns.

*The Flood* abounds in extended active metaphors, the most overarching being the one embodied in its very title (witness Zamjatin's own statement to that effect in the epigraph[23]). The frequency of the metaphors, their recurrence, and the elaborate way in which some of them are developed in *The Flood* — all of this suggests that the author used this poetic device extensively and with great deliberation.

Most metaphors in the story transpose concrete mundane concepts onto the complex emotional and psychological states of the characters. Thus in (3), the *designatum* — the almost subconscious and unarticulated but clearly disturbing and unpleasant feeling TI experiences in the shop at his workplace — acquires the *designans* "empty pit" (*pustaja jama*), which begins as a simile in (3a), and develops into a metaphor two paragraphs later in (3b):

(3a) Teper' v tom lesu byla osen', remni transmissii xlopali vxolostuju,
now in this forest was fall belts transmission flapped emptily
sonno voročalas' kakaja-to šajba. Trofimu Ivanyču stalo nexorošo,
sleepily turned some puck T. I. became sick
kak byvaet, esli stoiš' nad pustoj, neizvestno dlja čego vyrytoj jamoj. (sec. 1)
as happens if stands above empty unknown for what dug pit

[23] Zamjatin talks in this quote about the integrated image (*интегральный образ*) of flooding. The distinction between *образ* 'image' and metaphor is ignored here as inessential for the purpose of this study and probably not principal for Zamjatin himself. On the complex relationship of image, symbol, metaphor, and other signs, see Arutjunova 1998, pp. 313 ff.

'Now it was fall in this forest, disengaged transmission belts flapped emptily, some puck was turning sleepily. Trofim Ivanych felt sick, as one feels when one is standing above an empty pit dug out for some unknown purpose.'

The perspective in this passage belongs to TI: the chilly autumnal atmosphere in the barely operating shop is part of his experience, and his reaction to it is explicitly stated as such, using an impersonal construction of which he is the experiencer (*Trofimu Ivanyču stalo nexorošo*). This feeling of sickness is described by means of the "empty pit" simile. The sense of unpleasant emptiness returns to TI that same night:

(3b) Trofim Ivanyč v temnote našel rukoju ee koleni, dolgo byl vmeste
T. I. in dark found by-hand her knees long was together
s neju. I opjat' bylo ne to, byla kakaja-to jama. (sec. 1)
with her & again was wrong was some pit

'Trofim Ivanych found her knees in the dark, and stayed with her for a long time. And again something was wrong, there was some sort of pit.'

In (3b), the point of view is again unambiguously TI's. He is the grammatical subject of the first sentence, acting in the dark of night and initiating an act of sexual intimacy with S. In contrast to (3a), however, the metaphor of the "empty pit" acquires a gendered meaning in this passage. The pit dug out for no purpose is now used to represent S's barren sexual organs. The receptacle-like physical shape of the *designans* parallels the concept of a vagina or womb. To the extent that these organs do receive the male organ and its semen and hold it, they are indeed receptacles or vessels.[24] In many traditions, female sexual organs, and specifically the womb, are viewed as the vessel central to the following stages of the reproductive process, i.e., pregnancy and childbirth.[25] In TI's case, both notions are present. The "pit" is not just a receptacle. It is its failure to live up to its *raison d'être* that causes TI's feeling of sickness (*nexorošo*) and his sense of something being wrong (*bylo ne to*). The transition from his economically depressed semi-idle shop to his wife's barren womb and his reaction to both is managed quite consistently, thus convincingly representing TI's (male) perspective.

[24] Onan's options were to go "in unto his brother's wife" or to "spill" his semen "on the ground" (Genesis 38:9), a choice between putting his semen into a womb or spilling it outside. The frame of reference is clearly that of a vessel and its utilization.

[25] Conceptualization of the womb as a carrier but not the owner of its content,

The text is considerably less persuasive when the "pit" is mentioned from S's (female) perspective several paragraphs later, with reference to her feelings after the arrivals of menstruation:

(4) Kogda približalsja srok, ona ne spala, ona bojalas' – i xotela, čtob
when approached time she not slept she feared & wanted that
poskoree: a vdrug na ètot raz ne budet – vdrug okažetsja, čto ona … No
quicker what-if for this time not will-be what-if turns-out that she but
ničego ne okazyvalos', vnutri byla jama, pusto. (sec. 1)
nothing not turned-out inside was pit empty

'When the time approached, she wouldn't sleep, she would fear — and at the same time wish that the day would come sooner: what if this time the period wouldn't come — what if it turns out that she … But nothing would turn out, there was a pit inside, it was empty there.'

This passage attempts to represent S's point of view, and up to a point this works. The verbs *bojalas'* 'feared' and *xotela* 'wished' represent S's private fears and wishes, and the iterative verb *okazyvalos'* 'turned out to be' echoes her private expectations, which, in turn, are rendered in *erlebte Rede* (*a vdrug na utot raz ne budet — vdrug okažetsja, čto ona ...*), a sure indicator of S's point of view. Coming after this unambiguous evidence of S's perspective, the reference to S's womb as an empty three-dimensional receptacle (*vnutri byla jama, pusto*) is jarring. It not only adopts TI's "empty pit" *designans* for a *designatum* that is limited to a physical reality, but it also implies that S felt a three-dimensional empty space inside her body, a physical sensation that arises with certain digestive tract problems, but not when females are not pregnant. Reassigning to a woman both the *designans* and the *designatum* that presuppose a man's perception of the alleged three-dimensionality of the empty space inside a woman

---

the child, has found its way into the Japanese referential term *ofukuro* 'mother, lit. honorable sack' and into the saying *Hara wa karimono* 'The belly is something leased'; cf. also the testimony of grandmother Gujarati reported by Nina Mehta: "My grandmother was also not allowed to speak of her six children as *her* children. […] she'd have to avoid the possessive pronoun and note, simply, that *this* boy needs medicine […]. Her mother-in-law […] said that just because a sack contains the grain, does that mean the sack owns the grain?" (1998, 232). English expressions like *she bore him children* also imply a similar train of thought.

renders this "male" metaphor infelicitous in the context of a female perspective.[26]

Zamjatin ascribes another "male" metaphor to S in the following rendition of post-coital thoughts:

(5) I ponjala: esli ne budet rebenka, Trofim Ivanyč ujdet iz nee,
& understood-FEM if not will-be child T. I. will-go-out of her
nezametno vytečet iz nee ves' po kapljam, kak voda iz rassoxšejsja bočki.
imperceptibly flow-out of her all by drops as water from dried-up barrel
Èta bočka stojala u nix v senjax za dver'ju. Trofim Ivanyč uže davno
this barrel stood at them in hallway behind door T. I. already long
sobiralsja perebit' na nej obruči, i vse bylo nekogda. (sec. 1)
meant to-break on it hoops & always was busy

'And she understood: if she didn't bear him a child, Trofim Ivanych would go out of her, he would imperceptibly flow out of her drop by drop, like water out of a dried-up leaky barrel. The barrel stood in their hallway behind the door. Trofim Ivanych kept meaning to break its hoops, but never got to it.'

This is an extended and complex metaphor where the *designans* is an old, warped, and leaky *bočka* 'barrel', and the *designatum* is S's womb. The barrel's loss of the liquid inside it is a simile of Sofya losing TI's semen, which in turn is a metaphor of her losing him altogether. The perspective is clearly meant to be S's: the clauses containing the metaphors are introduced by the verb *ponjala* 'understood', with the thematic subject *Sof'ja* deleted. The identification of the real-life source of the 'barrel' metaphor as the barrel that stood in their hallway waiting to be dismantled by TI is also part of S's experience as a housewife. What is jarring here is the choice of the *designans*. A barrel

[26] This metaphor is later developed and reified, combining TI's first simile of a pit dug out for no obvious reason and a real pit that S dug out in order to bury the bisected body of her competitor, the teenager Gan'ka:

Ona vykopala jamu i svalila tuda vse, čto bylo v meške. Kogda bylo uže sovsem temno, ona prinesla polnyj mešok ešče raz, zaryla jamu i pošla domoj. 'She dug out a pit and dumped in it everything that was in the sack. When it was already completely dark, she brought another full sack, covered up the pit, and went home.'

This development is quite artful, showing that S irrationally perceived her pregnancy as a reincarnation of the girl she murdered. This reification of the metaphor, however, goes beyond the gendered and into the pathological — insofar as the psychology of a murder can be called pathological. The evaluation of metaphors based on pathological perception and experience goes beyond the scope of this paper.

is a somewhat oblong three-dimensional vessel that retains its shape even when empty. The metaphor again implies that S felt a three-dimensional empty space inside her body, not a sensation that is common to the introspective perception of a woman who is not pregnant.

The two metaphors discussed in this section are infelicitous. They occur in passages that are narrated from a woman's point of view; and yet, I suggest, a woman is not likely to relate to them. It is the "male" nature of these metaphors that brings about the jarring effect.

## *3. BEYOND DISCOURSE-PRAGMATIC INFELICITY*

There must be a reason why some and not other literary texts are selected for poetic analysis. But the criteria for selecting them are often taken for granted, with the choice only vaguely understood to be based on the analyst's or the literary critic's literary taste. I chose to analyze *The Flood* because I stumbled upon several jarring spots in it, after having accidentally obtained a copy of this story, over a decade ago — a story by an author far more famous for his other work. It was discourse pragmatic analysis and genderlinguistic analysis that finally made it possible for me to see what was wrong, and I have offered my diagnosis in sections 1 and 2 of this paper. Linguistic analysis enabled me to make explicit the causes for my own perception that something was wrong with the passages in (2), (4), and (5) analyzed above.

The analysis presented here so far, however, is incomplete. There are two problems with it, one on the linguistic side and the other literary. To begin with the linguistic side, the discourse analysis of example (2) was prompted by this analyst's sense of incoherence. Although, to be sure, a great deal of linguistic analysis is and has been driven by idiolectal native acceptability judgments, this kind of intuition, strictly speaking, needs to be confirmed, and we now have instrumental and statistical tools capable of doing that. In the case at hand, an experiment[27] could be devised using multiple native readers, in order to objectively measure readers' responses to loci where point of view shifts occur. Only after it is statistically verified that subjects react differently to the shift in (2) (as well as in the few other infelicitous passages of this sort in section 6 of the story), as opposed to all of the story's felicitous

[27] Perhaps on the basis of eye tracking (cf. Tanenhaus et al. 1995), or by measuring temporal responses to text processing.

shift sites, could the jarring effect be considered objectively established, at which point the analyst could then follow up on the intuition that was uncritically taken as the point of departure in this paper. Similarly, men and unpregnant women of child-bearing age could be tested about all of the metaphors in *The Flood,* so as to establish whether the two groups react differently to the two metaphors of the pit and the barrel, as compared with the other metaphors in the story.[28] If female subjects rate the two metaphors in question lower than the other metaphors in the story, but male subjects' ratings do not show such a pattern, then the analyst's original intuition would be confirmed, and the analytical enterprise would be justified.

Even if the phenomena themselves are confirmed by testing conducted on multiple subjects, and even if the discourse pragmatic and genderlinguistic analyses presented here are found to explain them satisfactorily, the analysis would still remain limited. It stops at explaining the infelicitous passages using criteria for coherent and communicatively competent narration, and does not yet go beyond, say, a genderlinguistic explanation as to why these metaphors are unconvincing in the eyes of female readers. Here we are dealing, however, with a deliberately created artistic text, which may, in principle, be intentionally "unreadable" in the Barthesian sense. Ostensibly inept locutions, even if confirmed to be perceived as such by multiple subjects and even if explainable using independent discourse pragmatic or genderlinguistic mechanisms, might in theory be a product of authorial intention. They could then be explained by a variety of reasons beyond sheer communicative or genderlinguistic infelicity. As exceptions to readers' expectation of felicitous narration, *skaz* and embedded nonnative messages were already mentioned in section 1, but there could be other literary rationales as well. One must look at the piece as a whole in order to determine whether any justification for infelicity can be found in it that may validate textual infelicity from a poetic standpoint, even if linguistically the text is unequivocally "inept". And besides the possibility of story-internal validation, if one is to seek possible justifications for apparent "ineptness" on even broader grounds, one must also consider the nature of author-reader contracts prevalent at the time in a given literary tradition.

[28] It would be interesting to test native speakers of different languages, to determine to what extent the results are culture-specific or universal.

As it happens, in the case of *The Flood*, there is no evidence of any literary justification for the few confusing shifts in point of view. The text as a whole is quite "readable" and predates the appearance of "unreadable" prose in Russia by over half a century. There is considerable textual evidence that the heroine is an independently feeling, thinking, and acting person capable of putting distance between herself and her husband. Furthermore, there is no evidence of the author's treating her as a woman whose psyche is prone to adopt her husband's views and sensibilities, as if in some sort of literary Stockholm syndrome. Finally, the absence of structural or content-based motivation for resorting to inept shifts just in those few spots where they occur militates in favor of treating them as aberrations rather than as a matter of conscious authorial choice. *Mutatis mutandis,* the same can be said about the two problematic gendered metaphors as well.

Thus, to fully evaluate these passages, one must go beyond a purely linguistic analysis. Only after the extralinguistic literary aspects of these problematic passages are examined and any poetic justifications for them are ruled out can one return to the linguistic analysis, and safely conclude that these jarring bits indeed reveal unintended infelicities and thereby compromise the *literary* quality of the text. This is what distinguishes the role of linguistic analysis in literary as opposed to non-literary texts that may ostensibly exhibit the same infelicities at the level of discourse. The linguist's diagnosis[29] of non-literary texts is the final word in those texts' evaluation; but in literary texts, literary analysis is required in addition, to determine whether the attested infelicities result from authorial choice or from the author's unintended and unaddressed inadequate command of discourse pragmatics, genderlinguistics, or other higher-level aspects of communicative competence. What is clear, however, as I have attempted to show here, is that a perfect command of communicational competence cannot be assumed in the production of a literary text, and that modern linguistics can play a significant role in diagnosing and explaining linguistic failures at that level of text analysis.

[29] Assuming, of course, that the diagnosis is accurate in linguistic terms.

*REFERENCES*

Arutjunova, N.D. (1998): *Jazyk i mir čeloveka.* Moscow: Jazyki russkoj kul'tury.

Barthes, R. (1970): *S/Z.* Paris: Seuil.

Culler, J. (1975): *Structuralist Poetics: Structuralism, Linguistics, and the Study of Literature.* Ithaca, NY: Cornell UP.

Ejxenbaum, B.M. (1978 [1927]): The theory of the formal method. In: Ladislav Matejka and Krystyna Pomorska, eds., *Readings in Russian Poetics: Formalist and Structuralist Views*, 3-37. Ann Arbor, MI: Michigan Slavic Publications.

Jakobson, R. (1987a): *Language in Literature.* Krystyna Pomorska and Stephen Rudy, eds. London, England & Cambridge, MA: Harvard UP.

_____ (1987b): Linguistics and poetics. In: Jakobson 1987a, 62-94.

Jokojama, O. (2005): *Kognitivnaja model' diskursa i russkij porjadok slov.* Moskva: Jazyki slavjanskoj kul'tury.

Mehta, N. (1998): From here to Poland. In: C. O'Hearn, ed., *Half and Half.* NY: Pantheon Books.

Mikhaychuk, G. (1994): The thread of consciousness in Chekhov's 'Steppe': the relevance of discourse features. *Slavic and East European Journal* 38, 4, 574-590.

Richards, I.A. (1936): *The Philosophy of Rhetoric*. Oxford: Oxford UP.

Tanenhaus, M. K.; Spivey-Knowlton, M. J.; Eberhard, K. M. & Sedivy, J. E. (1995): Integration of visual and linguistic information in spoken language comprehension. *Science* 268: 1632-1634.

Yokoyama, O.T. (1986): *Discourse and Word Order.* Amsterdam: John Benjamins.

_____ (2006): Discourse grammar vs. prescriptive stylistics. In: *Proceedings of the 27th Berkeley Linguistic Society Meeting*, 327-337. Berkeley, CA: Berkeley Linguistic Society.

Zaitseva, V. (1993): Discourse theory and the author-reader contract: the first person drafts of *Crime and Punishment.* In: Olga T. Yokoyama, ed., *Harvard Studies in Slavic Linguistics*, II, 243-265. Cambridge, MA: Harvard U. Slavic Linguistics Colloquium.

Zemskaja, E.A.; M.V. Kitajgorodskaja, and N.N. Rozanova (1993): Osobennosti mužskoj i ženskoj reči. In E.A. Zemskaja and D.N. Šmelev, eds., *Russkij jazyk v ego funkcionirovanii*, 90-136. Moscow: Nauka.

АВТОРИ/CONTRIBUTING AUTHORS

АЛЕФИРЕНКО, Николай; доктор филол. наук, профессор кафедры русского языка и методики преподавания Белгородского университета, Россия. Электронный адрес: Alefirenko@bsu.edu.ru

ALTMANN, Gabriel; Prof. em. Dr., editor Glottometrics. RAM-Verlag. Lüdenscheid, Germany. e-mail: ram-verlag@t-online.de

BEST, Karl-Heinz; Dr. Akad. Oberrat, Seminar für deutsche Philologie, Universität Göttingen, Germany. e-mail: kbest@gwdg.de

БЕЗУГЛАЯ, Лилия; кандидат филол. наук, доцент кафедры немецкой филологии и перевода, Харьковский национальный университет имени В.Н. Каразина, Украина. Электронный адрес: bezugla@daad-alumni.de

БІЛИНСЬКИЙ, Михайло; кандидат філол. наук, доцент кафедри англійської філології, Львівський національний університет імені І. Франка, Україна. Електронна адреса: bislo@ukrpost.ua

BUK, Solomija; Dr., Department for General Linguistics, University Lviv, Ukraine. e-mail: solomija@gmail.com

FENGXIANG, Fan; Prof. Dr., School of Foreign Languages, Dalian Martime University, Dalian, China. e-mail: fanfengxiang@yahoo.com

GRZYBEK, Peter; Prof. Dr., Institut for Slavic Studies, University Graz, Austria. e-mail: peter.grzybek@uni-graz.at

KELIH, Emmerich; Mag. Dr. phil., Institut for Slavic Studies, University Graz, Austria. e-mail: emmerich.kelih@uni-graz.at

KOHLSCHÜTTER, Christian; Dipl.-Inf. FH, Junior Researcher/PhD student at L3S Research Center, University Hannover, Germany. e-mail: kohlschuetter@L3S.de

КОРОЛЬ, Антоніна; кандидат філол. наук, доц. кафедри теорії і практики перекладу, Чернівецький національний університет імені Юрія Федьковича, Україна. Електронна адреса: konig@sacura.net

LÁSZLÓ, János; Prof., PhD, DSc., Institute for Psychology of the Hungarian Academy of Sciences and Professor and Director of the Institute of Psychology of the University of Pecs, Hungary. e-mail: laszlo@mtapi.hu

ЛЕВИЦКИЙ, Виктор; доктор филол. наук, профессор, зав. кафедрой германского, общего и сравнительного языкознания, Черновицкий национальный университет имени Юрия Федьковича, Черновцы. Электронный адрес: lessja@gmail.com

MAČUTEK, Jan; PhD, Institut for for Slavic Studies, University Graz, Austria. e-mail: jan.macutek@uni-graz.at

NEMCOVÁ, Emília; Prof. Dr., Department of Slovak Language and Literature, University of Saints Cyril and Methodius Trnava, Slovakia. e-mail: milka@stonline.sk

PÉLEY, Bernadette; Dr., Institute of Psychology University of Pécs, Hungary. e-mail: peley@btk.pte.hu

ПЕРЕЙМИБІДА, Андрій; кандидат фіз.-мат. наук, доцент кафедри математичного моделювання соціально-економічних процесів, Львівський національний університет імені І. Франка, Україна.

POPESCU, Ioan-Iovitz; Prof. Dr., Romanian Academy, Physical Sciences, Romania. e-mail: iovitzu@gmail.com

ПРОКОФЬЕВА, Лариса; доктор филол. наук, зав. кафедрой русского языка, Саратовский медицинский университет, Россия. Электронный адрес: prokofievalp@mail.ru

ROVENCHAK, Andrij; Dr., Department for Theoretical Physics, University Lviv, Ukraine. e-mail: andrij.rovenchak@gmail.com

САМОХИНА (ДМИТРЕНКО), Виктория; канд. филол. наук, доцент, зав. кафедрой английской филологии, Харьковский национальный университет имени В.Н. Каразина, Украина. Электронный адрес: germphil@univer.kharkov.ua

VENITOVÁ, Zuzana; Msc., Department of Slovak Language and Literature, University of Saints Cyril and Methodius Trnava, Slovakia. e-mail: zuzkave@zoznam.sk

WILSON, Andrew; PhD, Linguistics and English Language, Lancaster University, United Kingdom. e-mail: eiaaw@exchange.lancs.ac.uk

YESYPENKO, Nadia; PhD, Department of Theory and Practice of Translation, Chernivtsi National University, Ukraine. e-mail: jargar@ukr.net

YOKOYAMA, Olga T.; Prof. Dr., Chair of the Department of Applied Linguistics, UCLA, USA. e-mail: olga@humnet.ucla.edu

# ЗМІСТ/CONTENTS

*Наукове видання*

Методи аналізу тексту: Збірник наукових праць

Літературні редактори *Колодій О.В.*
Комп'ютерний набір та верстка *Мацкуляк Ю.Й.*

Методи аналізу тексту: збірник наукових праць. – Чернівці: ЧНУ, 2009. – 350 с.

*Реєстраційне свідоцтво ДК №891 від 08.04.2002 р.*
Підписано до друку . Формат 60 х 84/16.
Папір офсетний. Друк офсетний. Ум. друк. арк.
Обл.-вид. арк. . Зам. . Наклад

Друкарня Чернівецького національного університету
58012, Чернівці, вул. Коцюбинського, 2.