E-Dictionaries: the user's perspective

Stela Manova

ÖLT, WU Vienna, December 6, 2014

The research reported herein is funded by the **Laboratorio di Linguistica, Scuola Normale Superiore di Pisa** and carried out in cooperation with New Bulgarian University in Sofia and University of Trento.

E-dictionaries

- **Frequency e-dictionaries**
  - Frequency e-dictionaries provide information about the frequency of occurrence of lemmas and tokens in electronic corpora.

There are also frequency dictionaries of the language of an author, of a book, of a journal or newspaper, etc.

The importance of frequency

- Can be used to establish corpus similarity and homogeneity (Kilgarriff & Salkie 1996)
- Central notion in usage-based linguistics which is currently one of the dominant research paradigms in theoretical linguistics.
- Of particular importance to psycholinguistic research that involves lexical decision task, i.e. measuring how quickly people classify stimuli as words or non-words.

Frequency effects in word recognition

- **Frequent words are easier to recognize** than words which appear less frequently. Thus, recognition of frequent words is faster and more accurate than recognition of less frequent words.
- The **word frequency effect** is one of the most **robust** and most commonly reported effects in the literature on word recognition.

Outline of the paper

- A note on terminology
- Frequency information in use: an Italo-Bulgarian project on verb inflection
- The frequency dictionary of Italian used in this research
- Frequency dictionaries of Bulgarian
- Frequency dictionaries of other languages
- Small frequency dictionary vs large frequency dictionary
- Conclusions

A note on terminology

Corpus linguistics vs Theoretical morphology

- Lemma vs lexeme

(On the notion of lemma, Knowles & Zuraidah 2004)

  – lemma frequency

- Token vs word (form)

  – token frequency

An Italo-Bulgarian project on verb inflection

- **Our goal is to find out whether the citation form of the verb has an advantage over any other possible candidate to the status of base form**,

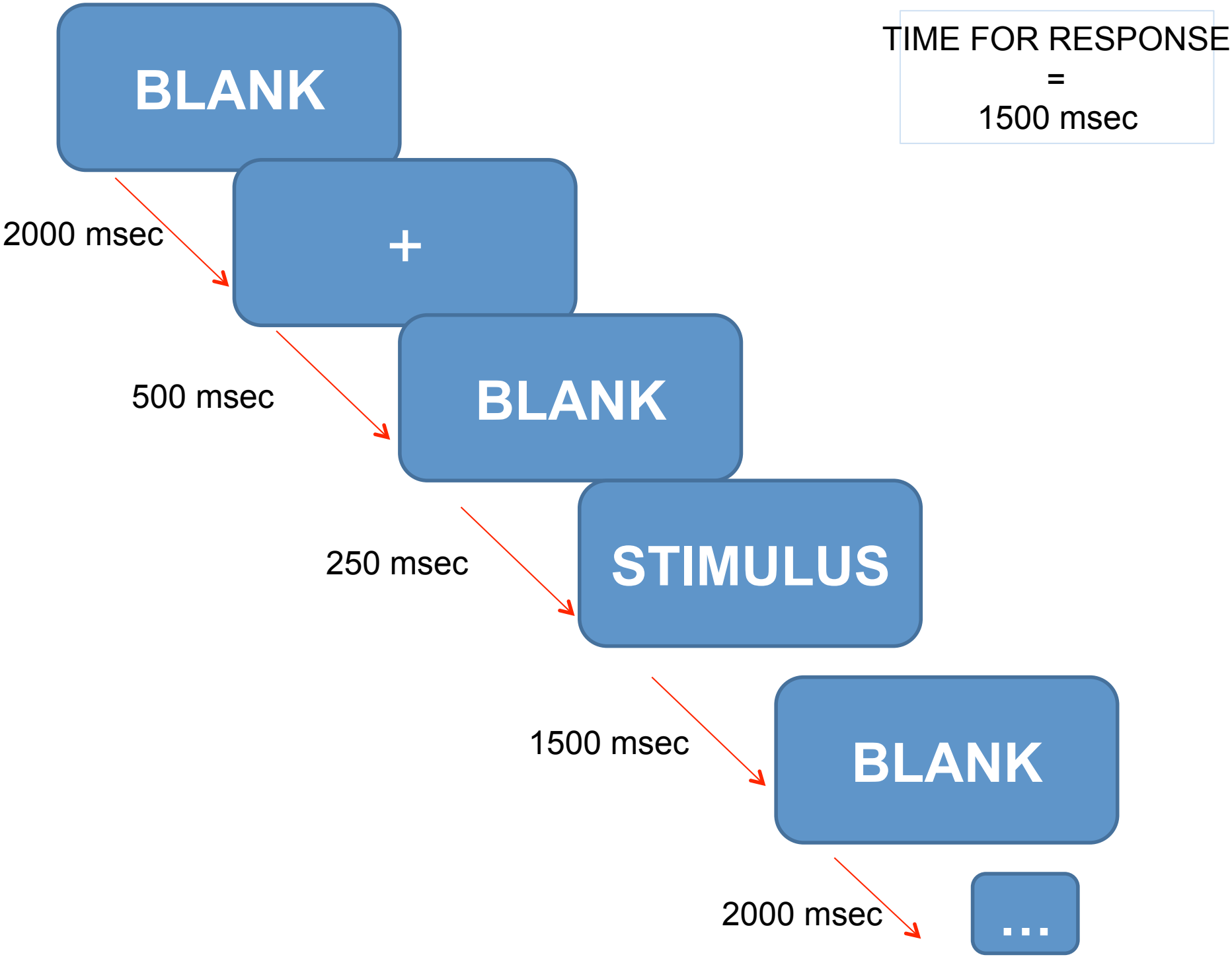
  or whether an alternative form should be considered the real base form of the verb. This would have implications for the process of lexical retrieval, and therefore we are interested in the matter.

Inflectional forms under investigation

- **Bulgarian** is a language **without infinitive**.
- **Italian** is a language **with infinitive**.
- The possible candidates for Bulgarian are: 1SG PRES (citation form) vs 3 SG PRES (most frequent form).
- The possible candidates for Italian are: Infinitive (citation form) vs 3 SG PRES.

The experiment

- Lexical decision protocol
  - **visual word recognition**
  - a set of **targets** (the forms under investigation: 1SG, 3SG PRES for BG and INF and 3 SG PRES for IT)
  - a set of **control forms** (other inflectional forms such as participles and gerunds)
  - a set of **fillers** (non-words). The non-words are obtained by modifying the root of real verb forms.


BLANK
2000 msec
+
500 msec
BLANK
250 msec
STIMULUS
1500 msec
BLANK
2000 msec
...
TIME FOR RESPONSE
=
1500 msec

Experiments 1 & 2 (Italian verbs)

- Exp. 1: We contrasted **high frequency** vs **low frequency** verbs. **Higher overall frequency of the 3 SG PRES** as compared with the INFINITIVE in Italian.
- Exp. 2: We selected **very rare verbs**, so that the difference between 3 SG PRES and INFINITIVE might be regarded as irrelevant, we selected **forms with frequency 1** in the Italian reference frequency e-dictionary.

Experiment 3 (Italian verbs)

- Exp. 3: Slightly more frequent verbs, although still in the lower range, i.e. since in exp. 2 we selected forms **with frequency 1, we had no control over the actual frequency** of the competing form. Thus, by selecting 3 SG PRES and INFINITIVE forms **with frequency 4 or 5, we could have a better control over the actual frequency** of each base-form candidate.

The frequency dictionary of Italian used in this research

- Based on Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS), Bertinetto et al. (2005).

CoLFIS is a small corpus, less than 4 million tokens; includes data from the most read Italian newspapers, books and journals. Thus, CoLFIS aims to represent the language perceived by the average Italian reader.

Demonstration

- The frequency of the verb *brindare* ‘to toast’ according to CoLFIS:

- http://esploracolfis.sns.it/EsploraCoLFIS/

Forme **Lemmi** **Concordanze** **Coricorrenze** **Lista** **Opzioni**

Forma: Azzera

Numero di lettere: minimo: Azzera massimo: Azzera

Dizionario

Lemma: Azzera

Parte del discorso: Azzera

Sintagmaticità: Azzera

Sintagmaticità della forma: Azzera

Statistiche

Risultati per pagina: 10

Cerca Azzera tutto

Forme | Lemmi | Concordanze | Coricorrenze | Lista | Opzioni

Forma: brindare Azzera

Numero di lettere: minimo: Azzera massimo: Azzera

Dizionario

Lemma: Azzera

Parte del discorso: Azzera

Sintagmaticità: Azzera

Sintagmaticità della forma: Azzera

Statistiche

Risultati per pagina: 10

Cerca | Azzera tutto

1 forma trovata

Forma	Lemma	Codice parte del discorso	Contesto sintagmatico: codice della parte del discorso	Rango	Freq. ass. totale	Log. freq. ass. totale	Freq. rel. totale	Dispersione totale	Ricerche correlate
brindare	BRINDARE	V	-	41666	4	0.60	0.42	0.36	Concordanze Lemma

Forme Lemmi Concordanze Coricorrenze Lista Opzioni

Lemma: brindare Azzera

Parte del discorso: Azzera

Numero di lettere: minimo: Azzera massimo: Azzera

Sintagmaticità: Azzera

Statistiche

Risultati per pagina: 10

Cerca Azzera tutto

1 lemma trovato

Lemma	Codice parte del discorso	Rango	Freq. ass. totale	Log. freq. ass. totale	Freq. rel. totale	Dispersione totale	Ricerche correlate
BRINDARE	V	12860	11	1.04	1.69	0.56	Concordanze Forme Forme dei tempi composti

Frequency dictionary of spoken Bulgarian (FDSB)

(100.000 tokens) Nikolova (1987)

```
аванс [м.] (пари)                6/4
          аванс    0 3 1 0 1
          аванси   0 0 0 1 0
аванта [ж.]        3/1
          аванта   0 3 0 0 0
август [м.]        2/2
          август   0 0 0 1 1
авиация [ж.]       1/1
          авиация  0 0 1 0 0
австрийски [прил.]          1/1
          австрийска        0 0 0 0 1
```

BulTreeBank Frequency List

http://www.bultreebank.org/Resources.html

- **based on 72 Mio tokens** (2.055 pp)
- **only token frequency; frequency considered > 22**

време	96081
една	94958
години	94634
защото	89561
според	89083
преди	87499
обаче	86992
него	86402
бяха	83786

Frequency Dictionary of Bulgarian National Corpus (BNC)
http://dcl.bas.bg/en/frequency_en.html

- **is a frequency list** (15.385 pp)

- **based on BNC version: December 2011**

Style	By frequency	In alphabetical order
Administrative	A-Administrative0001_byFreq	A-Administrative0001
Science	B-Science0001_byFreq	B-Science0001
Journalism	C-MassMedia_byFreq	C-MassMedia
Fiction	D-Fiction0001_byFreq	D-Fiction0001
Popular Science	G-PopularScience_byFreq	G-PopulatScience
Informal/Fiction	F-InformalFiction0001_byFreq	F-InformalFiction0001
GENERAL	General_byFreq	General

The problem: 1 SG PRES in BG

	verb	translation	Form freq. FDSB	Form freq. BTB	Form freq. BNC	Form freq.≈K Google
1	peja	sing	0	363	31465	423
2	svirja	wistle	0	265	19903	134
3	păxam	insert	0	31	2803	83,4
4	vărža	tie	0	37	3	131
5	pomolja	ask	1	369	42	921
6	pokrija	cover	3	42	35046	44,1
7	vnasjam	import	0	48	18549	64,8
8	proumeja	realize	2	118	6722	74,7
9	nakaram	make sb do sth	1	351	66358	380

The problem: 3 SG PRES in BG

	verb	translation	Form freq. FDSB	Form freq. BTB	Form freq. BNC	Form freq.≈K Google
1	maže	spread on	0	62	1	265
2	pljue	spit	2	112	0	336
3	vikne	call, cry	1	56	0	73,7
4	butne	push	0	92	1	115
5	opiše	describe	1	370	2	319
6	pukne	die	1	77	5	71,2
7	opăne	stretch	0	56	1	53,8
8	sxvane	grasp the meaning	1	158	7080	223
9	popadne	appear	0	904	5	659

Frequency dictionaries of other languages

- English

http://www.natcorp.ox.ac.uk/tools/xaira_search.xml?ID=frequency
http://www.wordcount.org/main.php

- German

http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html
http://wortschatz.informatik.uni-leipzig.de/

- Russian

http://bokrcorpora.narod.ru/frqlist/frqlist-en.html
http://masterrussian.com/vocabulary/most_common_words.htm

- etc.

See also http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists

The British National Corpus

Results of your search

Your query was

`table`

Here is a random selection of 50 solutions from the 19289 found.

A0U 2142 We settle ourselves down in a First Class cabin, lay our delicacies out on the table, open some wine and champagne, set the crayfish on to plates that don't look paper, and eat, drink and devour vast quantities of pâté, hors d'oeuvre and champagne.

AA1 98 For instance: ‘Man Explodes On Operating Table’(The Globe, California).

AEA 739 Instead of laying the dining-room table formally, Fru Gertlinger served supper on trays in the music room, so that the friends could listen to the gramophone while they ate.

AMT 1265 An analogy which would better express Aquinas' view compares the universe to a number of objects, say a pile of books on a table.

Conclusions

- Frequency dictionaries are an **important resource**.
- Frequency dictionaries should provide not only information on **token-frequency** but also such on **lemma-frequency**.
- **The size of the corpus that serves/d as basis** for preparation of the frequency dictionary must be mentioned in the dictionary explicitly.
- **Smaller, manually-annotated corpora** (CoLFIS < 4 Mio, BNC Sampler (two subcorpora, each 1 Mio tokens)) are, as a rule, more homogeneous (well-balanced) and **more reliable** as sources of information on frequency than huge corpora. Smaller resources are also easier to use.
- **Unification of frequency dictionaries**.
- **Software for creation of frequency lists** (of lemmas and tokens) is relatively inexpensive and should be **part of every serious electronic corpus**.

BNC

- The BNC Sampler is a subset of the full BNC. It comprises two samples of written and spoken material of one million words each, compiled to mirror the composition of the full BNC as far as possible. The word-class annotation of the BNC Sampler texts has been carefully checked and manually corrected.

Thank you!

Email: stela.manova@univie.ac.at
Homepage:
http://homepage.univie.ac.at/stela.manova/

Selected references

Bertinetto PM, Burani C, Laudanna A, Marconi L, Ratti D, Rolando C, Thornton AM (2005) Corpus e Lessico di Frequenza dell'Italiano Scritto.

Kilgarriff, A & Salkie, R (1996) Corpus similarity and homogeneity via word frequency. Euralex 1996, 121-130.

Knowles, G & Zuraidah, MD (2004) The notion of a "lemma": Headwords, roots and lexical sets. International Journal of Corpus Linguistics 9(1): 69-81.

Manova, S & Talamo, L (in press). On the significance of the corpus size in affix-order research. SKASE Journal of Theoretical Linguistics.