A Quantitative Approach to Preposition-Pronoun Contraction in Polish

Beata Trawiński

SFB 441
University of Tübingen
trawinski@sfs.uni-tuebingen.de

EACL 2006 Workshop
Third ACL-SIGSEM Workshop on Prepositions
Trento, Italy
3 April 2006
Preposition-Pronoun Contraction (PPC)

na niego ‘on him’ → nań ‘on_him’

w niego ‘in him’ → weń ‘in_him’

do niego ‘to him’ → dön ‘to_him’
Motivation and Goal

(Trawiński 2005) has observed that PPC in Polish is a highly idiosyncratic phenomenon:

- On the one hand, not just any pronoun can occur in a PPC.
- On the other hand, the set of prepositions which are able to contract with pronouns involves a very limited number of elements.

Goal:

- to provide corpus evidence for PPC’s realizations by
  - the characterization of the distribution of pronouns and prepositions within PPCs found in the corpus,
  - providing a quantitative analysis of the results.
Outline

- The system of Polish TPPPs according to (Saloni 1981)

- Corpus distribution of pronouns within PPCs

- Corpus distribution of prepositions within PPCs

- Outline of Quantitative Analysis

- Summary and Outlook
The System of Polish TPPPs

According to (Saloni 1981), the inventory of TPPPs in Polish comprises

- masculine human,
- masculine animate,
- masculine inanimate,
- feminine, and
- neuter

pronouns, inflecting for

- case (nominative, genitive, dative, accusative, instrumental and locative),
- number (singular and plural),
- postprepositionality (yes or no) and
- accentability (yes or no).
**Inflectional Paradigm of the Third Person Masculine Human Pronoun on ‘he’**

<table>
<thead>
<tr>
<th>CASE</th>
<th>SINGULAR</th>
<th>PLURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominative</td>
<td>on (—) / on (—)</td>
<td>oni (—) / oni (—)</td>
</tr>
<tr>
<td>genitive</td>
<td>jegó (go) / niego (-ń)</td>
<td>ich (—) / nich (—)</td>
</tr>
<tr>
<td>dative</td>
<td>jemu (mu) / niemu (—)</td>
<td>im (—) / nim (—)</td>
</tr>
<tr>
<td>accusative</td>
<td>jegó (go) / niego (-ń)</td>
<td>ich (—) / nich (—)</td>
</tr>
<tr>
<td>instrumental</td>
<td>nim (—) / nim (—)</td>
<td>nimi (—) / nimi (—)</td>
</tr>
<tr>
<td>locative</td>
<td>— (—) / nim (—)</td>
<td>— (—) / nich (—)</td>
</tr>
</tbody>
</table>
### Inflectional Paradigm of the Third Person Masculine Animate Pronoun on ‘it’

<table>
<thead>
<tr>
<th>CASE</th>
<th>SINGULAR</th>
<th>PLURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominative</td>
<td>on (—) / on (—)</td>
<td>one (—) / one (—)</td>
</tr>
<tr>
<td>genitive</td>
<td>jegó (go) / niego (-ń)</td>
<td>ich (—) / nich (—)</td>
</tr>
<tr>
<td>dative</td>
<td>jemu (mu) / niemu (—)</td>
<td>im (—) / nim (—)</td>
</tr>
<tr>
<td>accusative</td>
<td>jegó (go) / niego (-ń)</td>
<td>je (—) / nie (—)</td>
</tr>
<tr>
<td>instrumental</td>
<td>nim (—) / nim (—)</td>
<td>nimi (—) / nimi (—)</td>
</tr>
<tr>
<td>locative</td>
<td>— (—) / nim (—)</td>
<td>— (—) / nich (—)</td>
</tr>
</tbody>
</table>
### Inflectional Paradigm of the Third Person Masculine Inanimate Pronoun "it"

<table>
<thead>
<tr>
<th>CASE</th>
<th>SINGULAR</th>
<th>PLURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominative</td>
<td>on (—) / on (—)</td>
<td>one (—) / one (—)</td>
</tr>
<tr>
<td>genitive</td>
<td>jegó (go) / niego (-ń)</td>
<td>ich (—) / nich (—)</td>
</tr>
<tr>
<td>dative</td>
<td>jemu (mu) / niemu (—)</td>
<td>im (—) / nim (—)</td>
</tr>
<tr>
<td>accusative</td>
<td>jegó (go) / niego (-ń)</td>
<td>je (—) / nie (—)</td>
</tr>
<tr>
<td>instrumental</td>
<td>nim (—) / nim (—)</td>
<td>nimi (—) / nimi (—)</td>
</tr>
<tr>
<td>locative</td>
<td>(—) / nim (—)</td>
<td>(—) / nich (—)</td>
</tr>
</tbody>
</table>
Inflectional Paradigm of the Third Person Feminine Pronoun *ona* ‘she’

<table>
<thead>
<tr>
<th>CASE</th>
<th>SINGULAR</th>
<th>PLURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominative</td>
<td><em>ona</em> (—) / <em>ona</em> (—)</td>
<td><em>one</em> (—) / <em>one</em> (—)</td>
</tr>
<tr>
<td>genitive</td>
<td><em>jej</em> (—) / <em>niej</em> (—)</td>
<td><em>ich</em> (—) / <em>nich</em> (—)</td>
</tr>
<tr>
<td>dative</td>
<td><em>jej</em> (—) / <em>niej</em> (—)</td>
<td><em>im</em> (—) / <em>nim</em> (—)</td>
</tr>
<tr>
<td>accusative</td>
<td><em>ją</em> (—) / <em>nią</em> (—)</td>
<td><em>je</em> (—) / <em>nie</em> (—)</td>
</tr>
<tr>
<td>instrumental</td>
<td><em>nią</em> (—) / <em>nią</em> (—)</td>
<td><em>nimi</em> (—) / <em>nimi</em> (—)</td>
</tr>
<tr>
<td>locative</td>
<td>(—) / <em>niej</em> (—)</td>
<td>(—) / <em>nich</em> (—)</td>
</tr>
</tbody>
</table>
Inflectional Paradigm of the Third Person Neuter Pronoun *ono* ‘it’

<table>
<thead>
<tr>
<th>CASE</th>
<th>SINGULAR</th>
<th>PLURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominative</td>
<td><em>ono</em> (—) / <em>ono</em> (—)</td>
<td><em>one</em> (—) / <em>one</em> (—)</td>
</tr>
<tr>
<td>genitive</td>
<td><em>jego</em> (<em>go</em>) / <em>niego</em> (—)</td>
<td><em>ich</em> (—) / <em>nich</em> (—)</td>
</tr>
<tr>
<td>dative</td>
<td><em>jemu</em> (<em>mu</em>) / <em>niemu</em> (—)</td>
<td><em>im</em> (—) / <em>nim</em> (—)</td>
</tr>
<tr>
<td>accusative</td>
<td><em>je</em> (—) / <em>nie</em> (—)</td>
<td><em>je</em> (—) / <em>nie</em> (—)</td>
</tr>
<tr>
<td>instrumental</td>
<td><em>nim</em> (—) / <em>nim</em> (—)</td>
<td><em>nimi</em> (—) / <em>nimi</em> (—)</td>
</tr>
<tr>
<td>locative</td>
<td>(—) / <em>nim</em> (—)</td>
<td>(—) / <em>nich</em> (—)</td>
</tr>
</tbody>
</table>
Observations

The above paradigms indicate that

- there are four possible nominative and instrumental realizations, i.e., singular and plural postprepositional and non-postprepositional accented forms;
- all locative TPPPs have only two, i.e., singular and plural postprepositional accented realizations;
- all masculine and neuter TPPPs can be realized in the dative by non-postprepositional accented and unaccented singular forms, a postprepositional accented singular form, as well as by postprepositional and non-postprepositional accented plural forms;
- feminine dative TPPPs have no non-postprepositional unaccented forms;
- feminine genitive and accusative TPPPs can be realized in the singular and plural by postprepositional and non-postprepositional accented forms;
- neuter TPPPs have in addition a non-prepositional unaccented forms;
- genitive and accusative masculine singular TPPPs can be realized in all possible forms, in the plural, they can only be realized by postprepositional and non-postprepositional accented forms.
Conclusion

According to the inflectional paradigms proposed by (Saloni 1981), only

- genitive and accusative masculine human singular TPPPs,
- genitive and accusative masculine animate singular TPPPs, and
- genitive and accusative masculine inanimate singular TPPPs

possess unaccented postprepositional realizations, able to contract with prepositions.

Corpus evidence seems to demonstrate many further possibilities!
The IPI PAN Corpus

For the corpus-based investigations, the IPI PAN Corpus of Polish was used.

- The IPI PAN Corpus of Polish is a large (over 250 million segments), morphosyntactically annotated, publicly available corpus of Polish, developed by the Linguistic Engineering Group at the Institute of Computer Science at the Polish Academy of Sciences (cf. \cite{Przepiorkowski2004}, \cite{Przepiorkowski2005}).

- The corpus web page is located at http://korpus.pl.
The Procedure

PPCs were identified:
- manual identification, as PPCs are not recognized in the corpus as consisting
  of multiple segments, instead being identified as unknown forms (ign);
- a search was performed for all unknown forms ending in -(e)n;
- a total of 1193 PPCs were manually extracted from 3308 result matches.

An interpretation in terms of grammatical features was assigned to each pronoun
involved in the extracted PPCs by identifying its antecedent (the antecedent
identification proceeded manually as well).

The set of the acquired PPCs was verified by querying the corpus for all potential
contractions of unaccented postprepositional pronouns including each particular
Polish preposition.
Results I

As a result

- genitive and accusative masculine human plural unaccented postprepositional pronouns,
- locative masculine inanimate singular unaccented postprepositional pronouns,
- genitive and accusative masculine plural unaccented postprepositional pronouns,
- genitive and accusative neuter singular unaccented postprepositional pronouns,
- genitive, accusative and locative neuter singular unaccented postprepositional pronouns,
- genitive and accusative feminine singular unaccented postprepositional pronouns,
- genitive, accusative and locative feminine plural unaccented postprepositional pronouns

were recorded in addition to

- masculine human singular unaccented postprepositional pronouns,
- masculine animate signular unaccented postprepositional pronouns,
- masculine inanimate singular unaccented postprepositional pronouns.
Example 1

(1) Skoro zatem państwo pozbawiło fundację,
since therefore state robbed foundation.ACC.FEM.SG
ustanowionych dlań,
set for_TPPP.GEN.FEM.SG sources financing
zrozumiałe jest, że także państwo będzie przeznaczało ze
justifiable is that also state will assign from
swego budżetu środki na jej utrzymanie.
its budget means for its subsistence
‘Therefore, since the state robbed the foundation of the
financing sources set for it, it is justifiable that also the state
will assign means for its subsistence from its budget.’
Example 2

(2) Udana rola może co najwyżej spowodować, że successful role can what at most cause that zaproszą cię na przesłuchanie, a nie will invite.3RD.PL you for casting.ACC.NEUT.SG and not musisz się sam nań zgłaszać. have.2ND.SG RM alone for_TPPP.ACC.NEUT.SG enter ‘A successful role can at most cause them to invite you for a casting and you won’t have to come to the casting by yourself.’
Example 3

(3) Po ciężkiej operacji nogi i wyczerpującym okresie rekonwalescencji zawodnik powrócił na rajdowe trasy, by rywalizować nań o punkty w grupie N.

‘After a complicated leg operation and an exhaustive convalescence period, the contestant returned to the race tracks to compete there for points in the group N.'
Example 4

(4) także przed wschodem słońca wychodzili na swoje pola, gospodarze i wbijali w nich małe krzyżyki zrobione z gałązek tarniny.

‘Also before sunrise, farmers went to their fields and hammered small crosses made from the blackthorn twigs into them.’
Example 5

(5) [Polskie przedsiębiorstwa przemysłowe]i najlepiej prosperują w Mazowieckiem i Śląskiem. Pod koniec sierpnia br., uzyskało weńi zatrudnienie. August of this year gained in Loc.Neut.Pl employment łącznie ponad 1 mln osób [...]
altogether over 1 million persons [...] ‘Polish industrial companies are prospering the most in Mazowieckie and Śląskie. By late August of this year, a total of over 1 million people had gained employment there.’
A further observation that can be made on the basis of the corpus data is that the set of prepositions detected in contractions with unaccented postprepositional pronouns involves a very limited number of elements:

- dla ‘for’,
- do ‘to’,
- na ‘on’,
- od ‘from’,
- po ‘after’,
- przez ‘by’,
- w ‘in’,
- za ‘behind’,
- z ‘with’, and
- przed ‘in front of’.
Results III

No occurrences of contractions containing other prepositions were found in the corpus. While the absence of contractions involving secondary prepositions, such as

- *ponad* ‘above’,
- *poprzez* ‘through’,
- *zza* ‘from behind’,
- *między* ‘between’, etc.

corresponds to dictionary data, the non-appearance of contractions containing prepositions such as

- *bez* ‘without’,
- *o* ‘about’,
- *nad* ‘above’, or
- *pod* ‘under’,

being provided in Polish dictionaries such as *(Dubisz 2003)* or *(Bańko 2000)*, does not.
Table 1

Table 1 (see handout) presents an overview of the distribution of all unaccented postprepositional pronouns and prepositions within PPCs found in the IPI PAN Corpus. For each pronominal form,

- the context in which it occurs is specified, i.e., the contraction of that form with a particular preposition,
- the total number of times it occurred in the corpus is given,
- the percentage of the total frequency of all unaccented postprepositional forms is recorded,
- the total of all occurrences of each contraction found in the corpus is indicated,
- the percentage of the total frequency of all preposition-pronoun contractions occurring in the corpus is calculated.

The corpus contains considerably more unaccented postprepositional pronoun forms than the traditionally assumed paradigms!
To determine whether the distribution of the unaccented postprepositional pronouns in the corpus may be considered linguistically significant and, in consequence, may establish a basis for the revision of the inflectional paradigms, a number of quantitative procedures must be performed:

- It must be determined whether the frequency of each unaccented postprepositional pronoun form in the corpus is statistically significant.

- If one determines that the frequency of unaccented postprepositional pronouns in the corpus is statistically significant, ratios of the total number of particular accented postprepositional pronouns to the total number of their unaccented counterparts can be ascertained. These ratios can then be compared.

- If there are any significant differences between particular ratios, an attempt will be made to ascertain possible reasons for these differences (e.g., ungrammaticality, production errors, metadata restrictions, etc.).

- If the ratios of accented postprepositional pronouns to their unaccented counterparts not included in the inflectional paradigms correlate with the ratios of accented postprepositional pronouns to their unaccented counterparts contained in the inflectional paradigms, the distribution of the unaccented postprepositional pronouns in the corpus may be considered linguistically important.
Accented Postprepositional Pronouns

In our ongoing study, the distribution of accented postprepositional pronouns combining with the prepositions

- *dla* ‘for’,
- *do* ‘to’,
- *na* ‘on’,
- *w* ‘in’,
- *z* ‘with’,
- *od* ‘from’,
- *przez* ‘by’,
- *po* ‘after’,
- *za* ‘behind’, and
- *przed* ‘in front of’

has been investigating (these pronouns correspond to their unaccented counterparts occurring as parts of preposition-pronoun contractions).
Notes on the Interpretation

Assigning interpretations to pronouns must proceed manually on the basis of their antecedents, as most of the pronouns in the IPI PAN Corpus are resolved incorrectly.

In some cases, assigning an interpretation to a given pronoun was impossible for one the following reasons:

- the identification of an antecedent was not possible;
- multiple antecedent candidates bearing different features came into question;
- some features provided by an antecedent and a given pronoun were inconsistent with one another, e.g.,

\[
\text{(6) Możemy oczywiście dać Europie, także słowo honoru, ale obawiam się, że nie byłoby to oferta zbyt dla nich interesująca.}
\]

\[
\text{RM that not would be this offer too for them.PL interesting}
\]

‘We can, of course, also give Europe our word of honor but I am afraid that this would not be an interesting offer for them.’
Table 2 (see handout) provides the current results. For each form,

- the context corresponding to that of its unaccented counterpart is specified, i.e., a particular preposition with which the pronoun occurs,
- the total of all occurrences of the PPs headed by the particular prepositions is indicated,
- the percentage of the total frequency of the PPs in question is calculated.

For genitive and accusative feminine singular posprepositional forms:

- the total number of times it occurs in the corpus is given,
- the percentage of the total frequency of the accented postprepositional forms in question is recorded,
At the current stage of research, it has been ascertained that:

- genitive unaccented postprepositional feminine pronouns are used significantly less frequently in the IPI PAN Corpus than are genitive accented postprepositional feminine pronouns ($\chi^2 = 101.76$ (df=1), $p<0.001$);

- accusative unaccented postprepositional feminine pronouns are used significantly less frequently in the IPI PAN Corpus than are genitive accented postprepositional feminine pronouns ($\chi^2 = 36.95$ (df=1), $p<0.001$).

- the ratios of the totals of genitive and accusative accented postprepositional feminine singular pronouns to the totals of their unaccented counterparts amount to 226.56 and 148.42 respectively.
The Next Step

In the next step,

- the remaining accented postprepositional pronoun forms will be identified in the corpus and totaled,
- the ratios of the totals of these pronouns to the totals of their unaccented forms will be calculated,
- all ratios will be compared.

→ If there are any significant differences between particular ratios, an attempt will be made to ascertain possible reasons for these differences (e.g., ungrammaticality, production errors, meta data, etc.).

→ If there are no significant differences between the particular ratios, it will be concluded that the distribution patterns of pronouns and prepositions within PPCs found in the corpus are also linguistically significant and that the traditionally assumed inflectional paradigms of TPPPs, as well as previous dictionary specifications of PPCs, may have to be revised.
Summary

-The current results of our ongoing corpus-based study on the distribution of prepositions and pronouns within Polish PPCs were presented.

-At this point, conclusions can be drawn that, according to corpus evidence,

  - there seem to exist more pronominal forms being able to contract with prepositions than traditionally assumed,
  - there are fewer prepositions contracting with pronouns than assume Polish dictionaries.

-A quantitative analysis was proposed which draws on the calculation and comparison of ratios of the total frequency of all accented postprepositional forms to the total frequency of their unaccented counterparts.
Outlook

In future work,

- other corpora of Polish, such as the PWN Corpus of Polish (http://korpus.pwn.pl) or the PELCRA Corpus (http://korpus.ia.uni.lodz.pl) will be examined with respect to the distribution of PPCs, and the results will be compared with those achieved using the IPI PAN Corpus,

- meta data will be analyzed with respect to the distribution of PPCs,

- all results will be evaluated by human judges.
References


