A Quantitative Approach to Preposition-Pronoun Contraction in Polish

Beata Trawiński
University of Tübingen
SFB 441
Nauklerstraße 35
D-72074 Tübingen
trawinski@sfs.uni-tuebingen.de

Abstract

This paper presents the current results of an ongoing research project on corpus distribution of prepositions and pronouns within Polish preposition-pronoun contractions. The goal of the project is to provide a quantitative description of Polish preposition-pronoun contractions taking into consideration morphosyntactic properties of their components. It is expected that the results will provide a basis for a revision of the traditionally assumed inflectional paradigms of Polish pronouns and, thus, for a possible remodeling of these paradigms. The results of corpus-based investigations of the distribution of prepositions within preposition-pronoun contractions can be used for grammar-theoretical and lexicographic purposes.

1 Introduction

As (Świdziński and Derwojedowa, 2004) and (Trawiński, 2005) have observed, preposition-pronoun contraction (PPC) in Polish (cf. (1)) is a highly idiosyncratic phenomenon.

(1) a. *na niego* 'on him' → *nań* 'on_him'
b. *w niego* 'in him' → *weń* 'in_him'

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.[1]

The distribution of pronouns and prepositions within Polish PPCs has not yet been discussed in detail. There are, however, several traditional approaches to Polish third person personal pronouns (TPPPs) which provide some relevant information.[2] In the following, the approach to TPPPs of (Saloni, 1981), adopted in our research project, will be presented.

According to (Saloni, 1981), the inventory of Polish TPPPs 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). The inflectional paradigms of TPPPs proposed by (Saloni, 1981), and adopted in most Polish grammars, indicate that only genitive and accusative masculine human, masculine animate and masculine inanimate singular TPPPs possess unaccented postprepositional realizations, i.e., are able to contract with prepositions.[3] However, corpus evidence indicates that there may be many further possibilities of the realization of unaccented postprepositional pronouns, i.e., pronouns contractible with prepositions.

Corpus data also provide interesting information about the distribution of prepositions within PPCs. Only some PPCs found in the corpus correspond with respect to the form of prepositions contained in those PPCs, to dictionary data.

The goal of this research project is to characterize the corpus distribution of TPPPs and prepositions occurring within PPCs and to quantitatively analyze the results. While the first part of the

[1] For a discussion on prosodic, morphosyntactic and semantic properties of Polish PPC, see (Trawiński, 2005).

[2] Note that only third person personal pronouns can contract with prepositions in Polish.

[3] Note that (Doroszewski and Wieczorkiewicz, 1972) even claim that unaccented postprepositional pronouns are possible only in the accusative.

Proceedings of the Third ACL-SIGSEM Workshop on Prepositions, pages 17–22,
Trento, Italy, April 2006.

project has already been completed, the second one is still in progress. Section 2 presents the results of the corpus examination in regard to the distribution of pronouns and prepositions within PPCs, Section 3 outlines the proposal of a quantitative analysis of the results presented in Section 2, and Section 4 sums up the discussion and outlines future goals.

2 Corpus Distribution of Pronouns and Prepositions within PPCs

For the corpus-based investigation of the distribution of pronouns and prepositions within Polish PPCs, the IPI PAN Corpus of Polish was used.[4] Because of their very low frequency, the PPCs were searched for in the largest of the available IPI PAN subcorpora, i.e., the automatically annotated `wstepny corpus` (over 70 million segments).

PPCs had to be identified manually, as they were not recognized in the `wstepny corpus` as consisting of multiple segments, instead being identified as unknown forms (tagged by `ign`). Thus, in the first instance, a search was performed for all unknown forms ending in *-(e)ń*.[5] Next, a total of 1193 PPCs were manually extracted from 3308 result matches. Later, an interpretation in terms of grammatical features was assigned to each contracted pronoun by identifying its antecedent. The antecedent identification proceeded manually as well. Finally, the set of the acquired PPCs was verified by querying the corpus for all potential contractions of unaccented postprepositional pronouns with each particular Polish preposition.

As a result, genitive and accusative masculine human plural, locative masculine inanimate singular, genitive and accusative masculine inanimate plural, genitive and accusative neuter singular, genitive, accusative and locative neuter plural, genitive and accusative feminine singular, and genitive, accusative and locative feminine plural pronominal forms within PPCs were recorded in addition to the masculine human, masculine animate and masculine inanimate singular pronominal forms.

A further observation that was made on the basis of corpus data was that the set of prepositions detected in contractions with unaccented postprepositional pronouns involves a very limited number of elements, more precisely *dla* 'for', *do* 'to', *na* 'on', *od* 'from', *po* 'after', *przez* 'by', *w* 'in', *za* 'behind', *z* 'with', and *przed* 'in front of'. 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', *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', provided in Polish dictionaries such as (Dubisz, 2003) or (Bańko, 2000), does not.[6]

Figure 1 on the next page presents an overview of the distribution of all unaccented postprepositional pronouns and prepositions within PPCs found in the IPI PAN Corpus. For each pronoun form, the context in which it occurs is specified, i.e., the contraction of that form with a particular preposition, and the total number of times this form occurred together with the percentage of the total frequency of all unaccented postprepositional forms is recorded. In addition, the total of all occurrences of each contraction found in the corpus is indicated, as well as the percentage of the total frequency of all preposition-pronoun contractions occurring in the corpus.[7]

3 Quantitative Interpretation

To determine whether the distribution of the unaccented postprepositional pronouns and prepositions within PPCs found in the IPI PAN Corpus may be considered linguistically significant and, in consequence, may establish the basis for a revision of the traditionally assumed inflectional paradigms, a number of quantitative procedures must be performed.

First of all, it must be determined whether the frequency of each unaccented postprepositional

[4]The IPI PAN Corpus is a large (over 300 million segments), morphosyntactically annotated corpus of Polish, developed at the Institute of Computer Science at the Polish Academy of Sciences (cf. (Przepiórkowski, 2004)). The corpus web page is located at `http://korpus.pl`. For quantitative information about the corpus, see Przepiórkowski (to appear).

[5]Note that all TPPPs contracting with prepositions are realized by the syncretic form *-(e)ń*.

[6]Note, however, that in spite of the fact that contractions such as *oń* 'for_TPPP' or *weń* 'in_TPPP are included in dictionaries of contemporary Polish, these expressions are not accepted by all native speakers of Polish.

[7]The specifications *m1*, *m2* and *m3* refer to masculine human, masculine animate and masculine inanimate respectively. The minus signs indicate the absence of particular forms by means of the case government properties of the particular preposition.

	dlań 'for_TPPP'	doń 'to_TPPP'	nań 'on_TPPP'	weń 'in_TPPP'	zeń 'with_TPPP' / 'from_TPPP'	odeń 'from_TPPP'	przezeń 'by_TPPP'	poń 'after_TPPP'	zań 'behind_TPPP'	przedeń 'in front of_TPPP	Total, Percentage	
nom, m1, sg	—	—	—	—	—	—	—	—	—	—	0	0.00 %
gen, m1, sg	74	72	—	—	17	12	—	—	0	—	175	14.68 %
dat, m1, sg	—	—	—	—	—	—	—	0	—	—	0	0.00 %
acc, m1, sg	—	—	207	39	—	—	140	0	4	0	390	32.70 %
instr, m1, sg	—	—	—	—	0	—	—	–	0	0	0	0.00 %
loc, m1, sg	—	—	0	0	—	—	—	0	—	—	0	0.00 %
nom, m1, pl	—	—	—	—	—	—	—	—	—	—	0	0.00 %
gen, m1, pl	2	1	—	—	0	0	—	—	0	—	3	0.25 %
dat, m1, pl	—	—	—	—	—	—	—	0	—	—	0	0.00 %
acc, m1, pl	—	—	3	0	—	—	2	0	0	0	5	0.42 %
instr, m1, pl	—	—	—	—	0	—	—	—	0	0	0	0.00 %
loc, m1, pl	—	—	0	0	—	—	—	0	—	—	0	0.00 %
nom, m2, sg	—	—	—	—	—	—	—	—	—	—	0	0.00 %
gen, m2, sg	2	2	—	—	1	0	—	—	0	—	5	0.42 %
dat, m2, sg	—	—	—	—	—	—	—	0	—	—	0	0.00 %
acc, m2, sg	—	—	10	0	—	—	0	0	0	0	10	0.84 %
instr, m2, sg	—	—	—	—	0	—	—	—	0	0	0	0.00 %
loc, m2, sg	—	—	0	0	—	—	—	0	—	—	0	0.00 %
nom, m2, pl	—	—	—	—	—	—	—	—	—	—	0	0.00 %
gen, m2, pl	0	0	—	—	0	0	—	—	0	—	0	0.00 %
dat, m2, pl	—	—	—	—	—	—	—	0	—	—	0	0.00 %
acc, m2, pl	—	—	0	0	—	—	0	0	0	0	0	0.00 %
instr, m2, pl	—	—	—	—	0	—	—	—	0	0	0	0.00 %
loc, m2, pl	—	—	0	0	—	—	—	0	—	—	0	0.00 %
nom, m3, sg	—	—	—	—	—	—	—	—	—	—	0	0.00 %
gen, m3, sg	14	102	—	—	49	8	—	—	0	—	173	14.51 %
dat, m3, sg	—	—	—	—	—	—	—	0	—	—	0	00.0 %
acc, m3, sg	—	—	134	48	—	—	62	1	20	1	266	22.31 %
instr, m3, sg	—	—	—	—	0	—	—	—	0	0	0	0.00 %
loc, m3, sg	—	—	1	0	—	—	—	0	—	—	1	0.08 %
nom, m3, pl	—	—	—	—	—	—	—	—	—	—	0	00.0 %
gen, m3, pl	0	5	—	—	4	0	—	—	0	—	9	0.75 %
dat, m3, pl	—	—	—	—	—	—	1	0	—	—	1	0.08 %
acc, m3, pl	—	—	1	2	—	—	1	0	1	0	5	0.42 %
instr, m3, pl	—	—	—	—	0	—	—	—	0	0	0	0.00 %
loc, m3, pl	—	—	0	0	—	—	—	0	—	—	0	0.00 %
nom, neut, sg	—	—	—	—	—	—	—	—	—	—	0	0.00 %
gen, neut, sg	3	16	—	—	16	1	—	—	0	—	36	3.02 %
dat, neut, sg	—	—	—	—	—	—	—	0	—	—	0	0.00 %
acc, neut, sg	—	—	13	6	—	—	32	0	2	0	53	4.45 %
instr, neut, sg	—	—	—	—	0	—	—	—	0	0	0	0.00 %
loc, neut, sg	—	—	0	0	—	—	—	0	—	—	0	0.00 %
nom, neut, pl	—	—	—	—	—	—	—	—	—	—	0	0.00 %
gen, neut, pl	0	5	—	—	0	0	—	—	0	—	5	0.42 %
dat, neut, pl	—	—	—	—	—	—	—	0	—	—	0	0.00 %
acc, neut, pl	—	—	0	1	—	—	1	0	0	0	2	0.17 %
instr, neut, pl	—	—	—	—	0	—	—	—	0	0	0	0.00 %
loc, neut, pl	—	—	0	1	—	—	—	0	—	—	1	0.08 %
nom, fem, sg	—	—	—	—	—	—	—	—	—	—	0	0.00 %
gen, fem, sg	5	15	—	—	4	1	—	—	0	—	25	2.06 %
dat, fem, sg	—	—	—	—	—	—	—	0	—	—	0	0.00 %
acc, fem, sg	—	—	5	4	—	—	10	0	0	0	19	1.59 %
instr, fem, sg	—	—	—	—	0	—	—	—	0	0	0	0.00 %
loc, fem, sg	—	—	0	0	—	—	—	0	—	—	0	0.00 %
nom, fem, pl	—	—	—	—	—	—	—	—	—	—	0	0.00 %
gen, fem, pl	1	1	—	—	2	1	—	—	0	—	5	0.42 %
dat, fem, pl	—	—	—	—	—	—	—	0	—	—	0	0.00 %
acc, fem, pl	—	—	2	0	—	—	1	0	0	0	3	0.25 %
instr, fem, pl	—	—	—	—	0	—	—	—	0	0	0	0.00 %
loc, fem, pl	—	—	1	0	—	—	—	0	—	—	1	0.08 %
Total	101	219	377	101	93	23	250	1	27	1	1193	
Percentage	8.47%	18.36%	31.60%	8.47%	7.80%	1.93%	20.96%	0.08%	2.26%	0.08%		100%

Figure 1: The distribution of unaccented postprepositional pronouns and prepositions within the PPCs occurring in the IPI PAN Corpus

pronoun form in the corpus is statistically significant. For this purpose, the distribution of all accented postprepositional pronouns must be compiled. On the basis of the total frequency of accented and unaccented postprepositional pronouns, the statistical significance can be calculated using the χ^2 test, for instance. 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.[8] If the ratios of accented postprepositional pronouns to their unaccented counterparts not included in the traditionally assumed inflectional paradigms correlate with the ratios of accented postprepositional pronouns to their unaccented counterparts contained in the traditionally assumed inflectional paradigms, the distribution of the unaccented postprepositional pronouns in the corpus may be considered linguistically important.

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 ascertained. These pronouns correspond to their unaccented counterparts occurring as parts of the contractions *dlań* 'for_TPPP', *doń* 'to_TPPP', *nań* 'on_TPPP', *weń* 'in_TPPP', *zeń* 'with_TPPP' / 'from_TPPP', *odeń* 'from_TPPP', *przezeń* 'by_TPPP', *poń* 'after_TPPP', *zań* 'behind_TPPP', and *przedeń* 'in front of_TPPP respectively. Note that assigning interpretations to pronouns must proceed manually on the basis of their antecedents, as a vast number of pronouns in the IPI PAN Corpus are resolved incorrectly. Figure 2 on the next page provides the current results.[9]

[8]Alternatively, the percentage of occurrences of each unaccented postprepositional pronoun of the total number of occurrences of unaccented postprepositional pronouns and the percentage of occurrences of each accented postprepositional pronoun of the total number of occurrences of accented postprepositional pronouns can be ascertained and the results compared.

[9]Note that in some cases, assigning an interpretation to a given pronoun was impossible, which is indicated in Figure 2 by the question mark (*?*). In some cases, identification of an antecedent was not possible, more than one antecedent candidate bearing different features came into question, or some features provided by an antecedent and a given pronoun were inconsistent with one another. In the majority of cases, morphosyntactic features clashed with contextual / pragmatic / natural features.

Currently, only the distributional characterization of genitive and accusative feminine singular postprepositional pronouns is available for analysis. 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 (χ^2=101.76 (df=1), p<0.001), and accusative unaccented postprepositional feminine pronouns are used significantly less frequently in the IPI PAN Corpus than are accusative accented postprepositional feminine pronouns (χ^2=36.95 (df=1), p<0.001). The percentage of genitive unaccented postprepositional feminine singular pronouns of the total of all unaccented postprepositional pronouns amounted to 2.06%, while the percentage of genitive accented postprepositional feminine singular pronouns amounted to 11.41%. The percentage of accusative unaccented postprepositional feminine singular pronouns of the total of all unaccented postprepositional pronouns was 1.59%, while the percentage of accusative accented postprepositional feminine singular pronouns was 5.68%. The ratios of the totals of genitive and accusative accented postprepositional feminine singular pronouns to the totals of their unaccented counterparts are given in Figure 3. Additionally, Figure 3 provides the ratio of the total of all accented plural pronouns occurring in the contexts indicated in Figure 2, to the total of the unaccented forms. For the final conclusions, however, the distribution patterns of particular plural pronouns must be described.

	Ratio
gen, fem, sg	226.56
acc, fem, sg	148.42
pl	759.60

Figure 3: Ratios of accented postprepositional pronouns to their unaccented counterparts

In the next step, the remaining accented postprepositional pronoun forms will be identified in the corpus and totaled.[10] Then, the ratios of the totals of these pronouns to the totals of their unaccented forms will be calculated. Finally, all ra-

[10]Note that the total frequency of accented postprepositional forms corresponding to unaccented forms with zero frequency will, in fact, not affect the analysis.

	dla TPPP 'for TPPP'	do TPPP 'to TPPP'	na TPPP 'on TPPP'	w TPPP 'in TPPP'	z TPPP 'with TPPP' / 'from TPPP'	od TPPP 'from TPPP'	przez TPPP 'by TPPP'	po TPPP 'after TPPP'	za TPPP 'behind TPPP'	przed TPPP 'in front of TPPP'	Total, Percentage
nom, m1, sg											
gen, m1, sg	1141	1902									
dat, m1, sg											
acc, m1, sg									192		
instr, m1, sg									699		
loc, m1, sg											
nom, m1, pl											
gen, m1, pl	1207	987									
dat, m1, pl											
acc, m1, pl									126		
instr, m1, pl									310		
loc, m1, pl											
nom, m2, sg											
gen, m2, sg	8	24									
dat, m2, sg											
acc, m2, sg									1		
instr, m2, sg									25		
loc, m2, sg											
nom, m2, pl											
gen, m2, pl	14	12									
dat, m2, pl											
acc, m2, pl											
instr, m2, pl									9		
loc, m2, pl											
nom, m3, sg											
gen, m3, sg	128	1066									
dat, m3, sg											
acc, m3, sg									99		
instr, m3, sg									183		
loc, m3, sg											
nom, m3, pl											
gen, m3, pl	166	808									
dat, m3, pl											
acc, m3, pl									16		
instr, m3, pl									75		
loc, m3, pl											
nom, neut, sg											
gen, neut, sg	80	336									
dat, neut, sg											
acc, neut, sg									14		
instr, neut, sg									41		
loc, neut, sg											
nom, neut, pl											
gen, neut, pl	170	429									
dat, neut, pl											
acc, neut, pl									7		
instr, neut, pl									29		
loc, neut, pl											
nom, fem, sg											
gen, fem, sg	872	2619	0	0	1514	659	0	0	0	0	5664 11.41%
dat, fem, sg											
acc, fem, sg	0	0	1401	264	0	0	830	74	251	0	2820 5.68%
instr, fem, sg									580		
loc, fem, sg											
nom, fem, pl											
gen, fem, pl	319	914									
dat, fem, pl											
acc, fem, pl									9		
instr, fem, pl									123		
loc, fem, pl											
?	350								26		
Total	4455	9097	4853	4652	15143	2582	3661	591	2815	1773	49622
Percentage	8.98%	18.33%	9.78%	9.37%	30.52%	5.20%	7.38%	1.19%	5.67%	3.57%	100%

Figure 2: The distribution of accented postprepositional pronouns in the IPI PAN Corpus

tios 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.) and conclusions will be made. 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.

4 Summary and Outlook

In this paper, 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. On the other hand, corpus data provide fewer prepositions contracting with pronouns than do Polish dictionaries. To verify these results for the purpose of a possible revision of the traditionally assumed inflectional paradigms of TPPPs, as well as for lexicographic purposes, 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. The analysis will be completed within the next project phase.

In future work, other corpora of Polish, such as the PWN Corpus of Polish[11] or the PELCRA Corpus[12] will be examined with respect to the distribution of pronouns and prepositions within PPCs, and the results will be compared with those achieved using the IPI PAN Corpus.[13] Further on, meta data will be analyzed with respect to the distribution of TPPPs. Finally, all results will be evaluated by human judges.

[11] `http://korpus.pwn.pl`

[12] `http://korpus.ia.uni.lodz.pl`

[13] A preliminary list of PPCs occurring in the PWN Corpus has been provided to us by Magdalena Derwojedowa (personal communication). According to this list, the following PPCs appear in the PWN Corpus: *dlań* 'for_TPPP', *doń* 'to_TPPP', *nadeń* 'above_TPPP', *nań* 'on_TPPP', *odeń* 'from_TPPP', *oń* 'above_TPPP', *poń* 'after_TPPP', *przedeń* 'behind_TPPP', *przezeń* 'by_TPPP', *weń* 'in_TPPP', *zeń* 'with_TPPP' / 'from_TPPP'.

This set of PPCs does not fully correspond to that found of the IPI PAN Corpus. Thus, such a comparison seems to be reasonable.

Acknowledgments

We would like to thank Magdalena Derwojedowa, Elżbieta Hajnicz, Timm Lichte, Adam Przepiórkowski, Janina Radó, Zygmunt Saloni, Marek Świdziński and Marcin Woliński, as well as the reviewers of the Third ACL-SIGSEM Workshop on Prepositions held at the EACL 2006 in Trento for their helpful comments. We are also grateful to Janah Putnam for proofreading this paper.

References

Mirosław Bańko. 2000. *Inny słownik języka polskiego [Different Polish Dictionary]*. Wydawnictwo Naukowe PWN, Warszawa.

Witold Doroszewski and Bolesław Wieczorkiewicz. 1972. *Gramatyka opisowa języka polskiego z ćwiczeniami [A Descriptive Grammar of Polish with Exercises]*, volume II: Fleksja. Składnia [Inflection. Syntax.]. Państwowe Zakłady Wydawnictw Szkolnych, Warszawa.

Stanisław Dubisz. 2003. *Uniwersalny słownik języka polskiego [The Universal Polish Dictionary]*. Wydawnictwo Naukowe PWN, Warszawa.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus. Preliminary Version*. Institute of Computer Science PAS, Warsaw.

Adam Przepiórkowski. to appear, . The Potential of the IPI PAN Corpus. *Poznań Studies in Contemporary Linguistics*, 41:–.

Zygmunt Saloni. 1981. Uwagi o opisie fleksyjnym tzw. zaimków rzeczownych [Some Remarks on the Inflexional Description of Polish Pronouns]. In *Acta Universitatis Lodziensis*, volume 2 of *Folia Linguistica*, pages 243–253. Uniwersytet Łódzki.

Marek Świdziński and Magdalena Derwojedowa. 2004. Idiosynkrazja na przecięciu idiosynkrazyj, czyli o poprzyimkowości i liczebnikach [Idiosyncrasy at the Interface of Idiosynrasies. About Postprepositionality and Numerals]. In Andrzej Moroz and Marek Wiśniewski, editors, *Studia z gramatyki i semantyki języka polskiego*, pages 33–42. Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.

Beata Trawiński. 2005. Preposition-Pronoun Contraction in Polish. In *Proceedings of the Second ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 20–29, University of Essex, Colchester, United Kingdom.