Statistical analysis of the interaction between word order and definiteness in Polish

Adrian Czardybon, Oliver Hellwig, and Wiebke Petersen

SFB 991, University of Düsseldorf, Germany

Abstract. Although (in-)definiteness is semantically relevant in Polish, the language lacks explicit linguistic features for marking it. The paper presents the first quantitative, statistical evaluation of the correlation between word order and definiteness. Our results support previous qualitative theories about the influence of the verb-relative position on definiteness in Polish.

1 Introduction

The paper presents the first quantitative assessment of linguistic strategies for expressing definiteness in Polish using statistical evaluation of an annotated corpus.[1] We define definiteness as referential uniqueness of a noun or noun phrase (NP; details in Section 2). In contrast to languages such as English or German, Polish lacks definite and indefinite articles. Therefore, definiteness is usually not marked explicitly at the sentence level. This contrast between Polish and English is illustrated by example (1) which represents the first sentence of the Polish translation of George Orwell's novel *Nineteen Eighty-Four* [1, p. 7]. While no explicit markers of definiteness are found with the nouns *dzień* 'day' and *zegary* 'clocks' in the Polish sentence, articles mark the definiteness of the words '*day*' and '*clocks*' in the English translation.

(1) Był jasny, zimny dzień kwietniowy i zegary biły trzynastą.
was bright cold day April and clocks struck thirteen
"It was a bright day in April, and the clocks were striking thirteen."

Although Polish lacks articles, previous research leaves no doubt that definiteness is a relevant semantic feature in Polish. Szwedek states that "[a]lthough there is no article in Polish we seldom have doubts whether a noun in a text is definite or indefinite" [2, p. 203]. Researchers have discussed several linguistic structures that may be used for expressing definiteness in Polish, one of the most frequently mentioned being the position of an NP in relation to the position of the main verb [3,4,5,6].

While previous studies have dealt with definiteness in Polish mainly from a qualitative perspective, the present paper is, to our best knowledge, the first

[1] We would like to thank our annotators Joanna Strzępek and Helena Zamolska. The work on this topic is financed by the Deutsche Forschungsgemeinschaft (DFG) through CRC 991.

quantitative evaluation of definiteness strategies. Following the ideas found in previous research, the paper focusses on the verb-relative positions of NPs as an indicator of definiteness. Apart from submitting existing scientific hypotheses to a statistical assessment, the computational and statistical framework developed for this paper serves a more far-reaching purpose. If we are able to validate strategies that have an influence on the definiteness of the Polish NP, these strategies can also be used for developing machine learning algorithms that determine the definiteness of an NP in unannotated Polish corpora automatically. Such algorithms are a major building block for assessing Löbner's theory of concept types and determination [7,8] for computationally under-resourced languages such as Polish.

The following sections introduce the concept of definiteness and the linguistic features used to express it (2), describe the corpus (3) and its evaluation (4), and summarize our results (5).

2 Theoretical Background and Linguistic Features

This section formalizes the notion of definiteness for NPs. In addition, we present a short survey of previous research on linguistic factors that are said to influence definiteness in Polish, with a special focus on word order.

2.1 Definiteness of NPs

For this study, we follow Löbner [7,8] in assuming that uniqueness is the underlying concept of definiteness: If a noun is definite, there is only one referent that fits the definite NP in the given linguistic context. Löbner distinguishes between semantic and pragmatic uniqueness. Individual nouns such as *John*, *Pope* or *moon* are semantically, i.e. inherently unique, because they have only one referent in their contexts of utterance. This is also true for functional nouns such as *father*, *head* or *difference* which are two- or more-place predicates in contrast to the individual nouns. Functional nouns are inherently unique since each person can only have one father. Thus, they express a one-to-one relation between two entities (for example the father and the person who he is the father of). In contrast, sortal (*dog, book, chair*) and relational nouns (*brother, finger, uncle*) are not inherently unique. They require (extra-)linguistic context in order to achieve unique reference. Since a person can have more than one brother or none, relational nouns are not inherently unique expressing a one-to-many relation in contrast to the functional nouns.

We annotated NPs as definite if they refer uniquely. In this context, it was not relevant whether unique reference was due to the semantics of the noun (individual and functional nouns) or whether the unique reference was established from the context (pragmatic uniqueness).

2.2 Features

Word order has been mentioned frequently as one of the most important strategies for expressing definiteness in Polish [3,4,5,6]. Błaszczak claims that "in a postverbal position ... a nominal phrase not accompanied by any determiner ... is in principle ambiguous (definite or indefinite)" [5, p. 11]. Furthermore, she writes that "[i]n a preverbal position a nominal is normally interpreted as definite" [5, p. 15]. The theory that preverbal bare NPs are mainly interpreted as definite, whereas postverbal bare NPs can be definite or indefinite will be assessed in Section 4.

Apart from the verb-relative position of NPs, other strategies for expressing definiteness in Polish including perfective and imperfective aspect ([9], [10]) as well as case marking[2] ([10, p. 35], [14, pp. 30, 48–49, 86]). We are planning to examine the influence of these features in follow-up studies, along with the roles of pronouns such as possessive, demonstrative (*ten, tamten, ów, taki*), and indefinite pronouns (*jakiś* 'some', *jakikolwiek* 'any', *niektóry* 'some', *niejaki* 'some', *żaden* 'none', *pewien* 'certain', *inny* '(an)other', *jeden* 'one', numerals, quantifiers (*wszystek* 'all', *wiele* 'many/much', *kilka* 'a few/several', *parę* 'a few', *oba* 'both'), restrictive linguistic structures such as relative clauses or prepositional phrases, and NPs with ordinals and superlatives.

3 Data and Annotation

We based our study on the first 479 sentences of a Polish translation of George Orwell's novel *Nineteen Eighty-Four* [15], which is annotated with morphosyntactical information according to the TEI standard. Frequently, the 1-million-word subcorpus of the National Corpus of Polish ("NKJP") [16] is used for such annotation tasks. However, the fact that the NKJP does not consist of coherent text passages of more than 40 to 70 words [16, p. 54] would have been a major drawback in our case, because context plays a crucial role when it comes to deciding whether an NP is definite or not in Polish.

We used MMAX2 [17] for annotating the data. Annotation was carried out independently by two native speakers of Polish. Since we had to develop annotation guidelines while performing this initial study, guidelines were adapted during the process of annotation.

For each noun the three main categories (1) "part of an idiom/proverb", (2) "multiword lexeme", and (3) "(in-)definite noun" were assigned. Furthermore, there was always the option to choose "don't know". The nouns contained in the category "part of an idiom/proverb" (*w końcu* 'finally', *na czas* 'in time', *zdać sobie sprawę z czegoś* 'to realize sth') were excluded from the further analysis because they are normally not referential. The monolingual dictionary of Polish [18] was consulted in unclear cases of idioms/proverbs and multiword lexemes

[2] It is argued that verbs such as *kupić* 'buy', *dać* 'give', and *pożyczyć* 'lend/borrow' allow for a case alternation of the direct object ([11, p. 83], [12, pp. 316–317], [13, p. 72]).

(*klatka piersiowa* ‘chest’, *hokej na trawie* ‘field hockey’). The definiteness of the NPs assigned to categories (2) and (3) was chosen from among the subcategories (i) generic, (ii) indefinite, (iii) definite, explicitly marked by a demonstrative, (iv) definite due to other reasons, and (v) ambiguous between definite and indefinite reading. Generic NPs were excluded from the further evaluation because we were only interested in referential NPs. Option (iii) was included as a preparatory step for a follow-up study in which the role of demonstratives in marking definiteness in Polish will be investigated.

The annotation produced a total of 8664 word tokens, including 2447 nouns. Out of these nouns, 2059 were annotated with definiteness information, while the remaining ones belonged to the category “part of an idiom/proverb” and “don’t know”. Nouns having definiteness information were derived from 1079 different lemmata, out of which 696 were hapax legomena, 306 occurred 2-5 times, and the remaining 77 more frequently. Annotation yielded a $\kappa = 0.985$ according to [19]. This high value is certainly due to the fact that the guidelines were developed along with the initial annotation, and a clear drop of κ may be expected in follow-up annotations.

4 Statistical Evaluation

For assessing the sentence-position hypothesis that is examined in this paper (refer to Section 2), we needed to determine the positions of nouns in relation to the positions of the main verbs in each sentence. Because syntactic substructures are not marked in [1], we split input sentences into syntactic chunks that contain exactly one main (non-auxiliary) verb. For this sake, we used a heuristic function that describes typical sentence structures in Polish in terms of regular expressions. Subsequent statistical analysis was restrained to these one-verb chunks. It should be noted that each of these chunks may consist either of a main clause or of a subordinate clause. We were able to extract 304 chunks with exactly one main verb from 46.6% of all 479 sentences in this way, while the remaining sentences had unclear chunkings. As this study focusses on bare NPs, we further excluded 101 nouns and NPs that were used with a determiner such as a demonstrative, indefinite, or possessive pronoun, because these determiners influence the definiteness of the noun at the NP level. For each resulting chunk, we recorded the number of nouns occurring before and after the main verb, and the respective definiteness annotations of the nouns. Raw counts of this procedure are given in Table 1.

To test the research hypothesis that definiteness of NPs is related to their verb-relative positions, we constructed a 2 × 2 contingency table, using both columns and the two rows “definite (not explicit)” and “indefinite” from table 1. The content of this table is displayed as a bar plot in figure 1, grouped by definiteness (left) and the verb-relative position (right). Because the expected frequencies for all cells of the 2 × 2 contingency table were higher than 5, we applied a χ^2 test as a statistical test for count data to this table. The null hypothesis of the test claims that definiteness of NPs is not related to the verb-relative

Type	postverbal	preverbal
ambiguous	7	0
definite (demonstr. pron.)	4	9
definite (not explicit)	222	197
generic	3	1
indefinite	155	49

Table 1. The positions of nouns with definiteness annotations relative to the main verb

position, while the alternative hypothesis postulates such a relationship. Because this paper is an exploratory study, we chose a comparatively high significance level of $\alpha = 10\%$, which produces decisions that are in favour of the alternative hypothesis. The χ^2 test yields a value of 30.367 for the 2×2 contingency table constructed from Table 1, showing highly significant differences between the factors at the given significance level α.

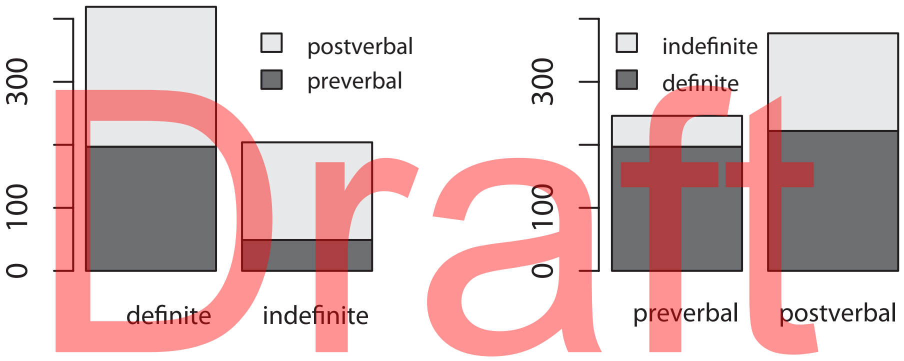

Fig. 1. Absolute frequencies of (in-)definite nouns in the pre- and postverbal position

The right plot in figure 1 clearly supports what Błaszczak [5, pp. 11–15] and other authors state about the influence of the verb-relative position on definiteness (refer to Section 2.2): Bare nouns in the preverbal position are mostly interpreted as definite, whereas the postverbal position is ambiguous in our corpus. In addition to these ideas formulated in previous research, the left subplot in figure 1 demonstrates that indefinite nouns show the tendency to occur in the postverbal position, while no positional preference is found with definite nouns.

5 Conclusion

The results of our study show a strong interaction between the definiteness of an NP and its position in relation to the main verb. This is in accordance with the observations made in the literature ([20, p. 235], [21, pp. 232–233], [6, p. 217]). The quantitative evaluation in Section 4 showed that the postverbal position is basically ambiguous in terms of definiteness, while the preverbal position is strongly associated with definite NPs. Analyzing our data in the opposite direction, the syntactic position of definite NPs cannot be predicted, whereas indefinite NPs are prominently found in the postverbal position, as can be observed in Figure 1. The comparatively high number of 49 indefinite preverbal NPs (refer to Table 1) is unexpected and should be submitted to a closer examination.

The results of this study indicate several directions for future research. First, we will focus on sentences with more than one NP placed postverbally, and investigate whether there is a tendency of placing indefinite NPs rather in sentence-final position in contrast to the postverbal, but not the sentence-final position. This approach is motivated by Szwedek's ([22, p. 80]) observation that the postverbal, but not sentence-final unstressed NP is always interpreted as definite. For this task, we need to annotate syntactic chunks either manually or by using a shallow syntactic parser (chunker). Second, it can be observed that inherently unique nouns such as individual and functional nouns are interpreted as definite regardless of their placement within the sentence. Löbner's theory of concept types and determination could explain our observation that definite nouns do not show clear positional preferences, as stated above and shown in Figure 1. Therefore, we are planning to annotate the concept type of the nouns in our corpus in a second step. This additional layer of information will make it possible to obtain a much more detailed picture of the connection between the syntactic position, definiteness, and the concept types. A working hypothesis for such a follow-up study would claim that sortal nouns have a tendency to be definite if placed preverbally, whereas they tend to be indefinite in the postverbal position, which is not the case with functional and individual nouns.

References

1. Orwell, G.: Rok 1984. Warszawskie Wydawnictwo Literackie MUZA SA. (2008)
2. Szwedek, A.: Some aspects of definiteness and indefiniteness of nouns in Polish. Papers and Studies in Contrastive Linguistics (2) (1974) 203–211
3. Szwedek, A.: Word Order, Sentence Stress and Reference in English and Polish. Linguistic Research, Edmonton (1976)
4. Grzegorek, M.: Thematization in English and Polish. A study in Word Order. PhD thesis, Drukarnia Uniwersytetu IM. A. Mickiewicza, Poznań (1984)
5. Błaszczak, J.: Investigation into the Interaction between the Indefinites and Negation. Akademie Verlag, Berlin (2001)
6. Mendoza, I.: Nominaldetermination im Polnischen. Die primären Ausdrucksmittel [Nominal determination in Polish. The primary means of expression]. PhD thesis, Ludwig-Maximilians-Universität, München (2004) Habilitationsschrift.
7. Löbner, S.: Definites. Journal of Semantics **4**(4) (1985) 279–326
8. Löbner, S.: Concept Types and Determination. Volume 28. Oxford University Press (2011)
9. Wierzbicka, A.: On the Semantics of the Verbal Aspect in Polish. In Jakobson, R., ed.: To Honor Roman Jakobson. Essays on the Occasion of his Seventieth Birthday. Mouton, The Haque (1967)

10. Witwicka-Iwanowska, M.: Artikelgebrauch im Deutschen. Eine Analyse aus der Perspektive des Polnischen [The article use in German. An analysis from the Polish perspective]. Narr Verlag, Tübingen (2012)
11. Topolińska, Z.: Remarks on the Slavic Noun Phrase. Wydawnictwo Polskiej Akademii Nauk, Wrocław (1981)
12. Topolińska, Z.: Składnia grupy imiennej [The syntax of the noun phrase]. In Topolińska, Zuzanna, ed.: Gramatyka współczesnego języka polskiego. Składnia. Państwowe Wydawnictwo Naukowe, Warszawa (1984) 301–386
13. Tokarski, J.: Fleksja polska [Polish inflection]. Wydawnictwo Naukowe, Warszawa (2001 [1973])
14. Sadziński, R.: Die Kategorie der Determiniertheit und Indeterminiertheit im Deutschen und im Polnischen [The category of determinedness and indeterminedness in German and Polish]. WSP, Częstochowa (1995)
15. Kotsyba, N., Radziszewski, A., Derzhanski, I., Erjavec, T.: Multext- East cesAna: Nineteen Eighty-Four, Polish, Warszawa (2010)
16. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B., eds.: Narodowy Korpus Języka Polskiego [The National Corpus of Polish]. Wydawnictwo Naukowe PWN, Warszawa (2012)
17. Müller, C., Strube, M. In: Multi-Level Annotation of Linguistic Data with MMAX2. Peter Lang, Frankfurt (2006) 197–214
18. Dubisz, S., Sobol, E.: Uniwersalny Słownik Języka Polskiego [Universal dictionary of the Polish language]. Wydawnictwo Naukowe PWN, Warszawa (2006)
19. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5) (1971) 378–382
20. Weiss, D.: Indefinite, definite und generische Referenz in artikellosen slavischen Sprachen [Indefinite, definite, and generic reference in article-less Slavic languages]. In Mehlig, H., ed.: Referate des VIII. Konstanzer Slavistischen Arbeitstreffens. Kiel 28.9. - 1.10.1982. Verlag Otto Sagner, München (1983) 229–261
21. Lyons, C.: Definiteness. University Press, Cambridge (1999)
22. Szwedek, A.: A Linguistic Analysis of Sentence Stress. Gunter Narr Verlag, Tübingen (1986)