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

Introduction

- First quantitative, statistical evaluation of the interaction between word order and definiteness in the article-less language Polish.
- No explicit markers of definiteness are found in (1) with the nouns *dzień* ‘day’ and *zegary* ‘clocks’ in the Polish translation in contrast to the English sentence.

(1) *Był jasny, zimny* ***dzień*** *kwietniowy i* ***zegary*** *biły trzynastą.*
was bright cold day April.ADJ and clocks struckt hirteen
‘It was **a bright day** in April, and **the clocks** were striking thirteen.
(Orwell 2008: 7)

- The position of an NP in relation to the position of the main verb is described to have an influence on the definiteness of the NP [Szwedek 1976, Błaszczak 2001].
- Investigation from a quantitative perspective to support previous qualitative studies.
- Far-reaching aim: to validate definiteness strategies which can be used for developing machine learning algorithms that determine automatically the definiteness of an NP in unannotated Polish corpora.

Theoretical Background

- **Definiteness**: We follow Löbner (1985, 2011) for whom uniqueness is the underlying concept of definiteness.
- If a noun is definite there is only one referent that fits the definite NP. Löbner uses two inherent properties: uniqueness [±U] and relationality [±R] resulting in the four noun types:

	[–U]	[+U] inherently unique
[–R]	Sortal nouns (SN): stone, table, chair	Individual nouns (IN): sun, Pope, Maria
[+R] inherently relational	Relational nouns (RN): brother, hand, uncle	Functional nouns (FN): head, mother, distance

Feature under investigation: Word order
Features for follow-up studies:
- Perfective and imperfective aspect [Wierzbicka 1967]
- Case marking [Sadziński 1995]
- Pronouns: possessive, demonstrative, indefinite pronouns
- Restrictive linguistic structures (relative clauses)
- NPs with ordinals and superlatives

Data and Annotation

- Annotation of definiteness of nouns in the first 479 sentences of a Polish translation of Orwell's novel *Nineteen Eighty-Four* [**Orwell**] (annotated with morpho-syntactical information according to TEI) using MMAX2 [**MMAX**]
- Annotation categories:
  - Main categories: (1) part of an idiom or proverb, (2) multiword lexeme, (3) (in-)definite noun
  - Categories of definiteness for (2) and (3): (i) generic, (ii) indefinite, (iii) definite, explicitly marked by a demonstrative, (iv) definite due to other reasons, (v) ambiguous
- 8664 word tokens, 2059 of 2447 nouns annotated with definiteness information.
- Standard annotation approach: Two annotators, one adjudicator, who merges the annotations; kappa = 0.985. Overestimates the real agreement.

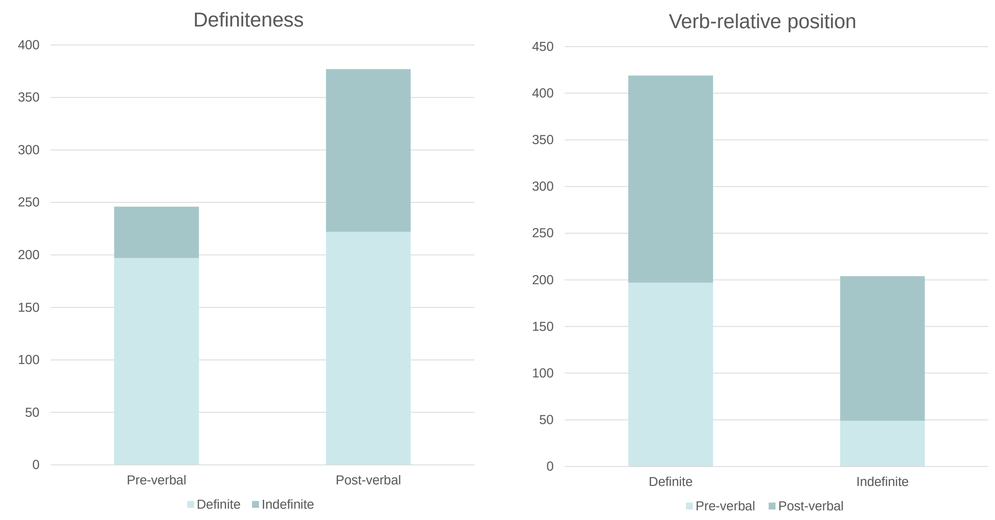
Statistical Evaluation

Heuristic chunking of the 479 sentences produces 304 chunks with exactly one main verb. Distribution of definiteness annotations depending on the pre-/post-verbal position in these chunks:

Type of definiteness	Postverbal	Preverbal
Definite (not explicit)	222	197
Indefinite	155	49

χ^2 –test:
- Null hypothesis: Deniteness of NPs is not related to the verb-relative of a noun.
- Significance level: 10%
- $\chi^2_{crit} = 2.71 < 30.367$: highly significant

Result of the test supported by a graphical representation:



Evidence for the claim: “in a postverbal position […] a nominal phrase not accompanied by any determiner […] is in principle ambiguous (definite or indefinite)” whereas “[i]n a preverbal position a nominal is normally interpreted as definite [Błaszczak 2001: 11, 15].

Conclusions

- Our results substantiate Błaszczak’s (2001) claim.
- Preverbal position is strongly associated with definiteness, postverbal position is basically ambiguous with respect to definiteness.
- The syntactic position of definite NPs cannot be predicted, whereas indefinite NPs are predominantly found in the postverbal position.
- Unexpected result: Comparatively high number of 49 indefinite preverbal NPs.

Next step:
It can be observed that inherently unique nouns (IN and FN) are definite regardless of their syntactic position. This could explain why definite NPs do not show clear positional preferences. To obtain a more detailed picture of the connection between syntactic position and definiteness, we plan to annotate the concept types of the nouns in our corpus.

Bibliography

[**Błaszczak**] Błaszczak, J.: Investigation into the interaction between the Indefinites and Negation. Akademie Verlag, Berlin (2001)
[**Löbner**] Löbner, S.: Definites. Journal of Semantics 4(4), 279-326 (1985)
[**Löbner**] Löbner, S.: Concept Types and Determination , vol. 28. Oxford University Press (2011)
[**MMAX**] Müller, C., Strube, M. In: Multi-Level Annotation of Linguistic Data with MMAX2. Peter Lang, Frankfurt (2006) 197214
[**Orwell**] Orwell, G.: Rok 1984. Warszawskie Wydawnictwo Literackie MUZA SA. (2008)
[**Sadziński**] Sadziński, R.: Die Kategorie der Determiniertheit und Indeterminiertheit im Deutschen und Polnischen. WSP, Częstochowa (1995)
[**Szwedek**] Szwedek, A.: Word Order, Sentence Stress and Reference in English and Polish. Linguistic Research, Edmonton (1976)
[**Wierzbicka**] 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)