Definiteness in languages with and without articles: a parallel corpus study

Laura Becker
(University of Cologne)

Overview This talk presents an empirical approach to comparing the expressions of referentiality in languages with and without articles using parallel movie subtitles. I show that based on the types of expressions used to refer, all languages point towards a three-way distinction of referential values. While the article is the most important factor to determine the referential value of an expression, languages with and without articles generally rely on very similar cues. In languages with both a definite and an indefinite article, we find a clear difference between the importance of the two articles: overall, the indefinite article is a much more importance marker for determining the referential function of an expression.

The present approach In order to compare coding strategies for (in)definiteness in 5 European languages, parallel subtitles from 5 movies (Pirates of the Caribbean 1, Harry Potter 1, Inception, The Lion King, Lord of the Rings 1) have been used. Parallel texts ensure that the semantics/pragmatics of the contexts are directly comparable in the different languages, which is especially important for languages without articles. From those subtitles, 500 parallel referring expressions have been extracted for German, Spanish, Hungarian (def. and indef. articles), Macedonian (def. article), Russian (no articles).

Annotation The following parameters were annotated: (i) referential value (deictic, anaphoric, establishing, situationally unique, bridging, specific indefinite, nonspecific indefinite referents); (ii) syntactic position (subject, object, oblique object, non-argument); (iii) clause type (main / subordinate); (iv) semantic features of the expression (human, concrete, abstract, place); (v) other elements in the noun phrase (possessives, demonstratives, adjectives, other attributes); (vi) type of the expression (noun phrase, pronoun, pro drop); (vii) type of the article (in/definite, none), (viii) number value of the expression (singular, plural). Using various parameters that are independent of each other differs from previous "hierarchy" approaches to referentiality (e.g. Gundel, Hedberg, and Zacharski 1993; Dryer 2013; Dryer 2014): it allows us to analyse the combination of different values and the impact of their interaction on the referential interpretation of the expression.

Results Based on the types of expressions that are used in the various referential contexts, the 7 levels of referentiality can be clustered according to their similarity, as is shown in Figure 1.

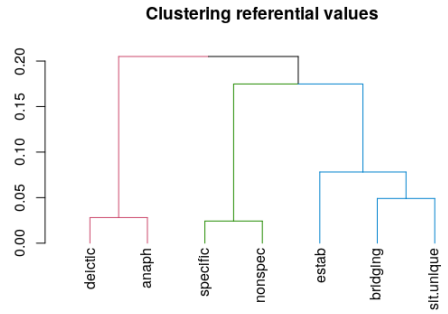

Figure 1: Similarity of the referential values

We see that instead of the traditional two-way split into the definite and the indefinite domains, the data points towards three main types of referential values: a first definite cluster including deictic and anaphoric referents (most of pronominal forms), a second definite cluster of establishing, bridging, and situationally unique referents, and the indefinite cluster consisting of specific and nonspecific referents.

In order to address the importance of the different formal properties annotated for the expression of definiteness in the different languages, random forest models have been used to examine the influence of each annotated factor for determining the referential value (I distinguish between the three main values determined by clustering) of the expression. For languages with articles, the latter is clearly the most important factor. However, the performance of the model fitted for Russian is similar to the ones of the other languages; also, the factors that are important to determine the referential value of the expression do not crucially differ across languages. While their ranking varies, it is the presence/absence of possessives, demonstratives, and the type of expression (noun vs. pro noun vs. drop) that contribute most information on the referential value of the expression. The syntactic function of the expression, the clause type, semantics of the expression show less influence; however, word order effects (such as left and right peripheral positions of the expressions) are expected to play a role and will be tested. As for the importance of the article, a comparison of the performances of models fitted with and without the articles as predictors for Spanish, German, and Hungarian show that it is mostly the indefinite article that contributes to the models overall accuracy, and to a lesser extent the definite article.

Examples

(1) **Bridging referent**
I'm told it's the latest fashion in London. Well, **women in London** must've learned not to breathe!

de	**Die Frauen** müssen gelernt haben, nicht zu atmen.	(art:def + noun)
sp	¡**Las londinenses** deben haber aprendido a no respirar!	(art:def + noun)
hu	Akkor **a londoni nők** újabban nem lélegeznek.	(art:def + adj + noun)
mk	**Жените во Лондон** веројатно научиле да не дишат.	(art:def + noun + other)
ru	Наверное, **лондонские модницы** научились обходиться без воздуха.	(adj + noun)

(2) **Specific (indefinite) referent**
If you have a few moments, Mr. Cobb has **a job offer** he'd like to discuss with you.

de	...würde Mr. Cobb gerne **ein Jobangebot** mit Ihnen besprechen.	(art:indef + noun)
sp	...el Sr. Cobb quiere ofrecerte **un trabajo**.	(art:indef + n)
hu	...Mr. Cobb ajánlana önnek **egy állást**.	(art:indef + noun)
mk	...г-дин Коб има **бизнис понуда** за тебе.	(adj + noun)
ru	...у мистера Кобба к тебе **деловое предложение**.	(adj + noun)

Selected references

Dryer, Matthew S. (2013). "Definite Articles". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, Matthew S. (2014). "Competing Methods for Uncovering Linguistic Diversity: The Case of Definite and Indefinite Articles (Commentary on Davis, Gillon, and Matthewson)". In: *Language Language* 90.4, pp. 232–249.

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski (1993). "Cognitive Status and the Form of Referring Expressions in Discourse". In: *Language* 69.2, pp. 274–307.