First-mention Definites: More than Exceptional Cases

Marta Recasens, M. Antònia Martí, Mariona Taulé
CLiC, Centre de Llenguatge i Computació
University of Barcelona
{mrecasens,amarti,mtaule}@ub.edu

1 Introduction

Traditional linguistic theories of definiteness have characterized the definite article in terms of *uniqueness* or *familiarity*, *inclusiveness* or *identifiability*. From this perspective, anaphoric uses of definite noun phrases (NPs) are seen as the paradigm case, while non-anaphoric or first-mention uses are treated as exceptions deserving no special attention. The main weaknesses of such approach are its tendency to be based on constructed examples and its focus on one single language, English. When natural data is taken into account, classical treatments of definites collapse.

2 Theoretical background

Fraurud's (1990) empirical study, "Definiteness and the Processing of Noun Phrases in Natural Discourse", marked a turning point because of its stress on the high occurrence of non-anaphoric definite NPs in natural discourse. Based on a 10,455-word Swedish corpus of written texts, Fraurud reports that only one third of all definites are subsequent mentions (i.e. anaphoric), while the remaining two thirds are first mentions, contrary to what would be expected according to traditional theories. A second corpus-based study, Poesio and Vieira's (1998), corroborated that finding for English and pointed out the need for a new account of definites. In this respect, Löbner's (1985) distinction between sortal, relational and functional concepts seems to shed light on building a theory that does not treat all definite NPs uniformly. Depending on the head noun class, Löbner classifies definites into *semantic* (if the head noun is functional) and *pragmatic* (if the head is sortal or relational). His main claim is that the role of the definite article is precisely to signal that the noun is to be taken as a functional concept. Therefore, the article is redundant in semantic definites, as the unambiguous reference comes from the lexical meaning of the noun itself; whereas in pragmatic definites, the article is responsible for forcing a functional reading of the noun in the context of utterance.

From the perspective of the definite article, Lyons (1999) surveys how the expression of definiteness varies greatly cross-linguistically –not only in form, but also in the functions it serves–, which brings him to formulate a new theory that regards definiteness not as a lexical but as a grammatical category. Lyons takes a diachronic perspective to outline the evolution of definiteness from a category of meaning expressing *identifiability* to a grammatical category. At this point it is relevant to recall Bybee's (1998) functionalist approach to grammar evolution, based on the belief that the only linguistic universals we can talk of are those of change. The definite article does fit in the *grammaticization* process as characterized by Bybee: the article has abstracted its meaning and the range of contexts in which it appears has increased. The different position that each language occupies in this grammaticization continuum accounts for the variation across languages at present. While English uses the article only in simple definites, Spanish does so also in generics, and Catalan represents a further stage where the article appears not only in simple definites and generics, but also in possessives and personal names.

All in all, it seems indeed irrefutable that limiting the meaning of the definite article to anaphoricity fails short to account for the overwhelming number of non-anaphoric definites observed in real occurring data. Lyons' (1999) distinction between lexical definiteness and grammatical definiteness correlates with Löbner's (1985) distinction between pragmatic and semantic definites. On the one hand, Löbner emphasises the role of the noun semantics; on the other hand, Lyons subtracts lexical meaning from the article in favour of a more grammatical role. These two ideas meet in Fraurud's (1990) claim for a non-uniform treatment of definite NPs. Non-anaphoric uses of definites should not be ignored by any theory that aims at providing a full account of how NP interpretation takes place and, by extension, by any theory on natural language understanding. The aim of this paper is to merge the views by Löbner (1985), Fraurud (1990), Lyons (1999), and Bybee (1998) in order to cast light on the non-anaphoric uses of definites in real data from Spanish and Catalan.

3 Empirical evidence

We carry out two quantitative corpus studies based on the AnCora corpus (Annotated Corpora for Spanish and Catalan).[1] Firstly, evidence for the grammaticization of the article is found in a typological study that compares different languages –Spanish, Catalan, Swedish and English– with respect to what we call the *definiteness ratio*, that is, the number of definite NPs in relation with the total number of full NPs[2]. The definiteness ratio yields an insight into the extent to which languages differ in

[1] AnCora consists of two 500,000-word corpora, mainly newspaper articles, annotated from the morphological to the semantic level (PoS tags, constituents and functions, argument structures and thematic roles, strong and weak named entities, and WordNet synsets). http://clic.ub.edu/ancora

[2] By *full NPs* we mean NPs with a nominal head, thus omitting pronouns, NPs with an elliptical head as well as coordinated NPs.

relation to the position they occupy in the grammaticization process. In languages with a high definiteness ratio such as Spanish and Catalan, the article has generalized and lost its semantic load to become the unmarked form, the one used in most contexts by default. In contrast, results for English reveal that the presence of the definite article is still very informative from the perspective of NP interpretation.

The second study focuses on Spanish and Catalan. We design a quantitative study to measure to what extent certain nouns tend to cooccur with the definite article. Given that the existence of a modifier can imply a change in the (non-)appearance of the article, we split full NPs into those with a modifier and those without any. By means of confidence intervals, we obtain the list of nouns that do statistically cooccur with the article. Our expectations are confirmed: most of the nouns in this list are functional –semantic definites– in Löbner's terms, while the article is pleonastic, a grammatical category as claimed by Lyons. This list of nouns makes it possible to delimit the considerably large group of first-mention definites, a group that, despite not being contemplated by traditional accounts, turns out to be far from exceptional.

4 Conclusion

The research here presented can help in the development of a computational algorithm for coreference resolution. The first study lays the basis for defining a parameter that takes the language type into account so that, depending on the definiteness ratio, certain rules of the algorithm might be differently applied. As a direct result of the degree of grammaticization, the list of nouns likely to appear in first-mention definites will be longer –and so more relevant– for languages with a higher ratio. An algorithm able to detect those NPs that, despite being definite, rely on no previous referent in the text (i.e. NPs that are candidates to start a coreference chain) means a significant improvement in the state-of-the-art coreference resolution systems.

References

Bybee, J. (1998). A functionalist approach to grammar and its evolution. Evolution of Communication, 2: 249-278.

Fraurud, K. (1990). Definiteness and the processing of NPs in natural discourse. Journal of Semantics, 7: 395-433.

Löbner, S. (1985). Definites. Journal of Semantics, 4: 279-326.

Lyons, C. (1999). Definiteness. Cambridge University Press, Cambridge, UK.

Poesio, M. and R. Vieira (1998). A corpus-based investigation of definite description use. Computational Linguistics, 24(2): 183-216.