Is there a definite article?
Reference annotation for a Torlak speech corpus

Max Wahlström
University of Helsinki
max.wahlstrom@helsinki.fi

Teodora Vuković
Slavisches Seminar/Language and Space Lab, University of Zurich
teodora.vukovic2@uzh.ch

This paper addresses the creation and implementation of a reference annotation fo r a South Slavic speech corpus. Among the Slavic languages the definite article is fo und only in Bulgarian and Macedonian. In an on-going corpus-based study, we examine the use of a clitic demonstrative pronoun, traditionally classified as a definite article, in the Timok variety of Torlak, which is a transitional group of dialects between Bulgarian and Macedonian, on the one hand, and Serbian, on the other. While the different stages of the grammaticalization of a definite article have been described in the languages of Europe (see, e.g., Pajusalu 2009, Heine & Kuteva 2006, 111–118, Laury 1997, Mladenova 2008), these processes have not been studied within a transitional variety of a dialectal continuum which displays both fully grammaticalized marking of definiteness and the complete lack of it.

In previous studies, even when the clitic is not treated as a definite article, these views are not based on a systematic examination (see, e.g., Ivić 1985, 115–116). We seek to demonstrate that the clitic has a distribution different from the full demonstrative pronoun regarding its anaphoric use. Yet we argue that it is not a marker of definiteness on par with the article in Bulgarian and Macedonian, since, fo r instance, its inferential use is not obligatory, as its omission from the word 'leash' in Example (1) demonstrates:

(1) Ćuštica, male, ~60 y.

oná	se	otkʲíne	znaš	blízu do	kaíšku=tu	a
it	REFL	tears.off	you.know	close to	collar.ACC.F.SG=CL.ACC.F.SG	CONJ
lánče	ostálo	a	oná	otíšla		
leash	left	CONJ	it	left		

'it [a fo x] tore itself off, you know, close to the collar, and the leash was left and it ran away'

The study is based on a recently collected speech corpus of the Timok varieties of Torlak spoken in Southeastern Serbia. In addition to semantic class, morphological features, and grammatical roles, the NPs in the corpus are being encoded for their reference. The reference annotation includes several discourse features, including coreference and number of mentions. For the type of reference, we employ a tagset created by Haug, Eckhoff, and Welo (2014, in their terminology, givenness annotation). However, designed for historical text corpora, its implementation for a speech corpus will be discussed in more detail.

For referential distance we propose a simple measure based on the number of intervening tokens. Unlike with a predefined threshold for discourse-old referents, our measure allows for a wider variety of statistical testing to assess the role of distance. In addition, we examine a way to extract the entire referential chains with the referential distance depicted visually. Finally, we discuss the caveats between the reference annotation and automated morpho-syntactic annotation, and propose a tagset for dislocated arguments, based on Juvonen (2000).

References

Haug, Dag, Hanne Eckhoff & Eirik Welo. 2014. The theoretical foundations of givenness annotation. In Patricia Cabredo Hofherr & Anne Zribi-Hertz (eds.), *Crosslinguistic studies on noun phrase structure and reference*, 17–52. Leiden: Brill.

Heine, Bernd & Tania Kuteva. 2006. *The changing languages of europe*. Oxford: Oxford University Press.

Ivić, Pavle. 1985. *Dijalektologija srpskohrvatskog jezika: uvod i štokavsko narečje*. Novi. Sad: Matica srpska.

Juvonen, Päivi. 2000. *Grammaticalizing the definite article: a study of adnominal determiners in a genre of spoken finnish*. Stockholm: Department of Linguistics, University of Stockholm.

Laury, Ritva. 1997. *Demonstratives in interaction: the emergence of a definite article in finnish*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Mladenova, Olga M. 2008. *Definiteness in bulgarian: modelling the processes of language change*. Berlin: Mouton de Gruyter.

Pajusalu, Renate. 2009. Pronouns and reference in estonian. *STUF -Language Typology and Universals* 62(1). 122–139.