Data-driven models of referential choice: Antecedent distance and beyond

Nils Norman Schiborr
University of Bamberg
nils-norman.schiborr@uni-bamberg.de

A long line of research in the tradition of Chafe (1976) has examined the factors determining the choice of referential expressions in natural discourse, that is when speakers opt for a pronoun, zero anaphora, or a full lexical expression. Generally, explanations are framed in an information packaging approach (Prince 1981; Givón 1983; Gundel et al. 1993; Chafe 1994; Gundel 2003), appealing to some variant of "accessibility" (Ariel 1990; 2000) or "activation states" (Chafe 1976). The distance between a given referential expression and its antecedent has been recognized as a defining variable in work on both referential choice (Ariel 1990; Kibrik 2000; Botley & McEnery 2001; Gipper 2016) and co-reference resolution (Levinson 1987; Ariel 1996; see also Lappin & Leass 1994; Mitkov 2002). This paper presents a data-driven and cross-linguistic approach to referential choice, focusing on antecedent distance and drawing on natural spoken discourse from four languages.

Schiborr (2017) examines antecedent distance in spoken English, testing the claims of Ariel's accessibility hierarchy (1990; 2000; 2004). Accessibility theory postulates a monotonously graded scale of referring expressions whose specificity and informativeness is expected increase proportionally with the distance to their antecedent. Schiborr (2017) finds little support for most of the finer-grained distinctions on the accessibility hierarchy; instead, referring expressions display a dichotomy between lexical (full noun phrases) and non-lexical expressions (pronouns including demonstratives, and zero anaphora). Initial work on additional languages largely supports this conclusion, see Figure 1.

However, although antecedent distance is evidently an important predictor for the choice between lexical and non-lexical expressions, it cannot account for the choice between pronominal expressions and zero anaphora. This paper presents ongoing work on extending this approach to additional languages, particularly those exhibiting a higher tolerance for zero anaphora ("pro-drop"), as well as on the selection of additional factors for modelling other aspects of referential choice (e.g. the humanness of the referent in question, its overall prominence in the discourse, and competition with other referents).

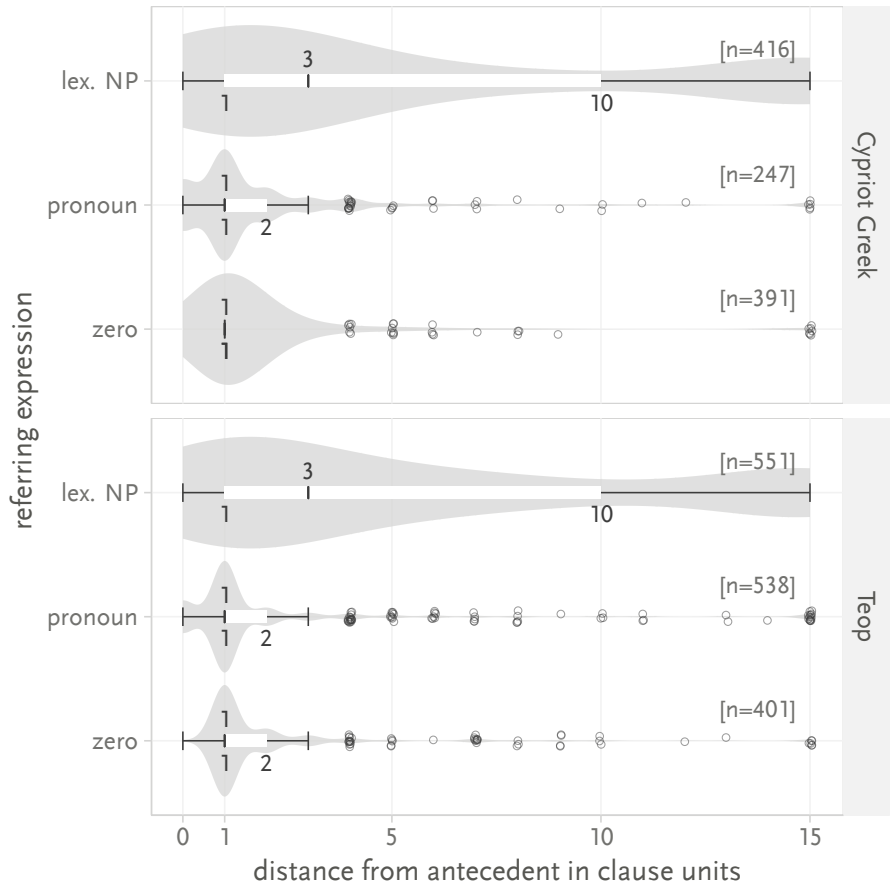

Figure 1 Distribution of referring expressions by the winsorized distance from their antecedent in clause units (same clause for $d = 0$). Data sources: Cypriot Greek (Vollmer & Hadjidas 2015) and Teop (Oceanic, Nehan-Bougainville; Mosel & Schnell 2015).

Methodologically, the approach is bottom-up, in that the relevant analytical categories are derived secondarily from the annotated corpus data, rather than being pre-imposed. Thus, although most researchers agree that some notion of topicality is relevant for referential choice, the present approach does not operate within a pre-conceived theory of topicality. Distance calculations are performed with the aid of the RefIND scheme (Referent Indexing in Natural-language Discourse, Schiborr et al. 2018), which tracks co-referential mentions throughout a discourse. Other information is gleaned from annotations with the GRAID scheme (Grammatical Relations and Animacy in Discourse, Haig & Schnell 2014), which marks all referential expressions, including covert expressions (i.e. zero

anaphora). This paper examines natural language data from four languages – Cypriot Greek, English, Teop, Vera'a – all taken from the freely accessible Multi-CAST collection (Haig & Schnell 2015),[1] and may include more depending on progress until December.

References

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

Ariel, Mira. 1996. Referring expressions and the +/– coreference distinction. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and referent accessibility*, 13–36. Amsterdam: John Benjamins.

Ariel, Mira. 2000. The development of person agreement markers: From pronouns to higher accessibility markers. In Barlow, Michael & Kemmer, Suzanne (eds.), *Usage-based models of language*, 197–220. Stanford: Center for the Study of Language and Information.

Ariel, Mira. 2004. Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse Processes* 37(2). 91–116.

Botley, Simon P. & McEnery, Tony. 2001. Proximal and distal demonstratives: A corpus-based study. *Journal of English Linguistics* 29(3). 214–233.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, Charles N. (ed.), *Subject and topic*, 25–55. New York: Academic Press.

Chafe, Wallace. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: The University of Chicago Press.

Gipper, Sonja. 2016. Constraints on choice of referring expression in Yurakaré. In Holler, Anke & Suckow, Katja (eds.), *Empirical perspectives on anaphora resolution*, 143–168. Berlin: de Gruyter.

Givón, Talmy (ed.). 1983. *Topic continuity in discourse* (Typological Studies in Language 3). Amsterdam: John Benjamins.

Gundel, Jeanette K. 2003. Information structure and referential givenness/newness: How much belongs in the grammar? In Müller, Stefan (ed.), *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar, Michigan State University*, 122–142. Stanford: CSLI Publications.

Gundel, Jeanette K. & Hedberg, Nancy & Zacharski, Ron. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2). 274–307.

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (version 7.0)*. (`https://lac.uni-koeln.de/en/multicast/`) (Accessed 2015-12-30).

Haig, Geoffrey & Schnell, Stefan (eds.). 2015. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (`https://lac.uni-koeln.de/multicast/`) (Accessed 2016-02-08).

Kibrik, Andrej A. 2000. A cognitive calculative approach towards discourse anaphora. In Baker, Paul & Hardie, Andrew & McEnery, Tony & Siewierska, Anna (eds.), *Proceedings from the 3rd Discourse Anaphora and Reference Resolution Colloquium (DAARC 2000)*, 72–82. Lancaster: Lancaster University Centre for Computer Corpus Research on Language.

1 Online at `https://lac.uni-koeln.de/en/multicast/`.

Lappin, Shalom & Leass, Herbert J. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4). 535–561.
Levinson, Stephen C. 1987. Pragmatics and the grammar of anaphora: A partial pragmatic reduction of binding and control phenomena. *Journal of Linguistics* 23(2). 379–434.
Mitkov, Ruslan. 2002. *Anaphora resolution*. London: Longman.
Mosel, Ulrike & Schnell, Stefan. 2015. Multi-CAST Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://lac.uni-koeln.de/multicast-teop/) (Accessed 2016-02-22).
Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Cole, Peter (ed.), *Radical pragmatics*, 223–255. New York: Academic Press.
Schiborr, Nils N. 2015. Multi-CAST English. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://lac.uni-koeln.de/multicast-english/) (Accessed 2016-02-28).
Schiborr, Nils N. 2017. *Antecedent distance and the accessibility hierarchy: A quantitative approach*. Master's thesis. Bamberg: University of Bamberg.
Schiborr, Nils N. & Schnell, Stefan & Thiele, Hanna. 2018. *RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines (v1.1)*. Bamberg / Melbourne: University of Bamberg / University of Melbourne. (https://www.uni-bamberg.de/fileadmin/aspra/misc/RefIND-guidelines-v1.1.pdf) (Accessed 2018-04-09).
Vollmer, Maria C. & Hadjidas, Harris. 2015. Multi-CAST Cypriot Greek. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (https://lac.uni-koeln.de/multicast-cypriot-greek/) (Accessed 2016-02-22).