Grammatical properties that favor the development of indefinite articles

Background Definite articles are crosslinguistically more frequent than indefinite ones (Dryer 1989). However, in Western Asia, the trend towards indefinite marking is higher than in other areas of the world (Becker 2018, Dryer 2013), which could be accounted for by reference-based differential object marking (DOM), found in a number of languages in the area (cf. Haig & Khan 2018, Johanson & Csató 1998, Windfuhr 2013): Case marking takes over the expression of definiteness/specificity in the object position, leaving mainly the subject position unmarked for referentiality. Arguing that this favors the development of indefinite articles over definite articles relies on the hypothesis that lexical subjects mostly contain given, topical referents. Then, the referential status of such subjects should be definite by default, making definite marking functionally redundant while motivating explicit indefinite marking.

Aim of the talk This talk presents a crosslinguistic corpus study that verifies this hypothesis by analysing the distribution of referent types, expressed lexically, pronominally, or left unexpressed, across different syntactic functions.

Corpus & methodology The study is based on the multicast corpus (Haig & Schnell 2019), containing morphosyntactically annotated monologic natural narrative text from 12 languages (1000-5700 clauses per language), with an additional referential annotation for 9 of those (used for this study): Cypriot Greek, English, Northern Kurdish, Mandarin, Sanzhi Dargwa, Nafsan, Teop, Tulil, and Vera'a. The annotation tracks all discourse-new, bridging (identifiable but first mention), and discourse-old referents, including pro and zero forms, and distinguishes intransitive, transitive subject; transitive, oblique object; locative, goal argument. A Bayesian logistic regression model was fitted to predict referents (new, old, bridging) from the interaction between linguistic form (lexical, pronominal, zero) and syntactic positions.

Results The model confirms that across languages, the association between the subject position and old referents is stronger than for other positions (also because it has the highest proportions of pro and zero forms). Furthermore, if the referent is expressed lexically, the proportion of given referents is higher in the transitive (and intransitive) subjects as opposed to objects, obliques, locatives, and goals: the model predicts around 80% of old transitive subject referents for lexical NPs and around 10% of new and 10% of bridging referents, as opposed to a proportions of 60% vs. %20 and %20 in the other non-subject positions.

This means that lexical NPs in subject positions are comparatively rare to begin with and if they occur, they are very likely to have a discourse-old referent.

Implications for the development of indefinite articles The findings suggest that (i) the (transitive) subject position is functionally more relevant for the development of indefinite articles than object and oblique argument positions, while (ii) definite marking of lexical subject NPs is functionally less motivated. On the one hand, this accounts for the hypothesis that indefinite articles are mainly used with topical referents in their initial stages (Givón 1984, Heine 1997, Hopper & Martin 1987). On the other hand, this study provides crosslinguistic evidence in favor of the association between DOM and a preference for indefinite over definite articles, because definiteness being marking on (transitive) objects, definite marking in subjects is functionally redundant.

Becker, Laura. 2018. *Articles in the world's languages.* University of Leipzig dissertation.

Dryer, Matthew S. 1989. Article-noun order. *Chicago Linguistic Society* 25. 83–97.

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

Givón, Talmy. 1984. *Syntax: a functional-typological introduction. Volume 1.* Amsterdam: Benjamins.

Haig, Geoffrey & Geoffrey Khan. 2018. *The languages and linguistics of Western Asia: An areal perspective.* Berlin: De Gruyter Mouton.

Haig, Geoffrey & Stefan Schnell. 2019. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts.* https://multicast.aspra.uni-bamberg.de.

Heine, Bernd. 1997. *Cognitive foundations of grammar.* Oxford: Oxford University Press.

Hopper, Paul J. & Janice Martin. 1987. Structuralims and diachrony: The development of the indefinite article in English. In Anna Giacalone Ramat, Giuliano Bernini & Onofrio Carruba (eds.), *Papers from the 7th International Conference on Historical Linguistics*, 295–304. Amsterdam: John Benjamins.

Johanson, Lars & Éva Csató (eds.). 1998. *The Turkic languages.* London and New York: Routledge.

Windfuhr, Gernot. 2013. *The Iranian languages.* London: Routledge.