Reference production across human languages

Stefan Schnell & Guido Linders

In connected discourse, users of human languages express series of states-of-affairs (SOAs) that are intended to be interpreted as interconnected in various ways to yield a coherent whole, i.e., meaningful discourse. In addition to core conceptual aspects of events and states themselves, producers will verbalize entities that participate in each SOA and that are new in the current shared discourse model or identifiable with participants mentioned in previous SOAs. Entities that are kept stable across SOA verbalizations are discourse referents.

We here present corpus-based computational findings on reference production from diverse human languages, i.e., the introduction and tracking of discourse referents throughout discourse. Reference production involves a range of choices, foremost the choice of referential form, which can be a full description, a type of proform, or zero reference, but also a range of constructional choices (e.g. voice or certain non-canonical constructions like dislocations, inversions, etc.) as well as choices in the ordering of participant expressions (syntactic arguments) within a sentence. We focus here on referential choice and argument order.

Drawing on computational models of reference production in spoken-language corpora represented in the multilingual corpus Multi-CAST and (mostly) written-language corpora from Universal Dependencies, we find that referential choice abides by most of the functional factors known from the literature (e.g. Chafe's seminal work and work in that tradition). Yet, we also identify structural constellations as highly important. This points to the necessity for more balanced theories of reference production that take both functional-communicative as well as idiosyncratic-structural features into account. In relation to argument ordering, our current findings lend overall support to an agent-first bias, yet other features are likewise relevant, and their relevance depends in part on the position of the verb in a simple sentence.