Information structure and cognitive states in the structure of German sentences

The word order of German and English is usually described as quite rigid, however with substantial differences. Thanks to case marking, German can easily flip its unmarked SVO order and bring non-subject elements to the topic position, while English has to use specific constructions (Durrell 2017: 935-936). By contrast, the structure of German sentences is organized around the predicate into three topological fields: the prefield, the midfield and the postfield, with specific restrictions for certain constituents; for instance, adverbial, predicate and arguments are forbidden in the postfield. Accordingly, German can easily modify the order of verbal arguments for information structure. But what about the order of elements in the sentence? And how does this interact with the cognitive accessibility of these elements, which Gundel, Hedberg and Zacharski (1993) has described as the Givenness hierarchy?

We conduct a quantitative analysis on 2,500 German sentences from miniCIEP+ (Verkerk and Talamo 2024), a parallel corpus parsed according to the Universal Dependency (UD) framework and annotated for information structure using the schema described in Anonymous (2024). The annotation provides referents with labels describing the information status (given vs. new) as well as mention type (anaphor, cataphor, predicate, apposition, discourse deixis and lexical coreference). The annotation is not currently available for German, but we plan to project it from English to German sentences using AWESOME, a word-by-word aligner (Dou & Neubig 2021). For each annotated mention of a referent, we extract features building up the German forms of the six types of cognitive states described by the givenness hierarchy (Gundel, Hedberg and Zacharski 1993): uniquely identifiable (indefinite article + N), referential (indefinite pronoun/referential expression + N), uniquely identifiable (definite article + N), familiar (distal demonstrative + N), activated (distal/proximate demonstrative, proximate demonstrative + N) and in focus (personal pronoun). We then extract the information structure (status and coreference) and compute the relative position of the referent in the sentence, operationalized as a scale ranging from 0 (at the very beginning of the sentence) to 1 (at the very end of the sentence). For instance, take the German sentence *Im letzen Jahr war er zu dem ersten Mal dort gewesen* ‘‘Last year he has been there for the first time’, whose annotation is shown in Figure 1; the three annotated mentions are: *Im letzen Jahr*, which is in the referential state, *er*, which is the focus state, and *dort*, which is in the activated state.

We fit a non-linear regression models with the information status as the response variable and relative position, coreference and type of cognitive state as the independent variables. We expect that given referents show lower values for the relative position i.e., when they appear at the beginning of the sentence and that different degrees of coreference play a meaningful role in the sentence structure and processing, similarly to what is discussed for English in Ye, Tu, and Pustejovsky, 2023.

```
1-2  Im    _    _    _    _    _    _    _    _
1    In    in   ADP  APPR _    4    case _    _
2    dem   der  DET  ART  Case=Dat|Definite=Def|Gender=Neut|Number=Sing|PronType=Art  4  det  _  Entity=(e20-Time-1-CorefType:coref,InfStat:new
3    letzten letzt ADJ ADJA Case=Dat|Degree=Pos|Gender=Neut|Number=Sing  4  amod _  _
4    Jahr  Jahr NOUN NN   Case=Dat|Gender=Neut|Number=Sing  12 obl  _  Entity=e20)
5    war   sein AUX  VAFIN Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin  12  aux  _  _
6    er    er   PRON PPER Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs  12  nsubj _  Entity=(e1-Person-1-CorefType:ana,InfStat:given)
7-8  zum   _    _    _    _    _    _    _    _
7    zu    zu   ADP  APPR _    10   case _    _
8    dem   der  DET  ART  Case=Dat|Definite=Def|Gender=Neut|Number=Sing|PronType=Art  10  det  _  _
9    ersten erst ADJ  ADJA Case=Dat|Degree=Pos|Gender=Neut|Number=Sing|NumType=Ord 10  amod _  _
10   Mal   Mal  NOUN NN   Case=Dat|Gender=Neut|Number=Sing  12 obl  _  _
11   dort  dort ADV  ADV  _    12   advmod _  Entity=(e18-Place-1-CorefType:ana,InfStat:given)
12   gewesen sein AUX VAPP VerbForm=Part  0  root _  SpaceAfter=No
13   .     .    PUNCT $.   _    12   punct _  _
```

Figure 1. The sentence *Im letzen Jahr war er zu dem ersten Mal dort gewesen* ‘Last year he has been there for the first time’ annotated for Universal Dependencies (column 1-9) and for information structure (column 10).

References

Anonymous (2024). A Multilingual Parallel Corpus for Coreference Resolution and Information Status in the Literary Domain.

Dou, Zi-Yi and Graham Neubig. (2021). Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2112–2128, Online. Association for Computational Linguistics.

Durrell, Martin (2017). Hammer's German Grammar and Usage (Routledge Reference Grammars). Sixth Edition. Abingdon, Oxon ; New York, NY : Routledge.

Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski (1993). “Cognitive Status and the Form of Referring Expressions in Discourse”. In: Language 69.2, pp. 274–307.

Verkerk, Annemarie and Luigi Talamo (2024). “mini-CIEP+ : A Shareable Parallel Corpus of Prose”. In: Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024. Ed. by Pierre Zweigenbaum, Reinhard Rapp, and Serge Sharoff. Torino, Italia: ELRA and ICCL, pp. 135–143.

Ye, Bingyang, Jingxuan Tu, and James Pustejovsky (2023). “Scalar Anaphora: Annotating Degrees of Coreference in Text”. In: Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023). Ed. by Maciej Ogrodniczuk et al. Singapore: Association for Computational Linguistics, pp. 28–38. doi: 10.18653/v1/ 2023.crac-main.4.