Guying Zhou / Jing Li

die Frau – sie – ø: A corpus-based multifactorial analysis of the use of anaphoric forms in German

This article delves deep into the intricacies of the German language, specifically focusing on the use of anaphoric forms. In linguistic terms, *anaphora* denotes the reference to entities which have already been mentioned in the text and serves as the continuation of a topic. This study aims to shed light on the underlying mechanisms that dictate the choice of anaphoric forms in German, with a specific focus on lexical NP anaphora (NA), pronominal anaphora (PA), and zero anaphora (ZA).

Different languages utilize various anaphoric forms to varying degrees. For instance, some East Asian languages, such as Chinese and Japanese, are more characterized by zero anaphors compared to others. Consequently, foreign language learners and translators often find it challenging to choose the appropriate anaphoric form as they must overcome the influence of their native language. One of the primary observations in the study is the intuitive ability of native German speakers to choose the correct anaphoric form, while non-native speakers often grapple with this choice. This observation underscores the complexity and depth of linguistic nuances that native speakers often take for granted. This research focuses on the production-side use of different anaphoric devices in German. It aims to describe the anaphoric choice mechanism (NA vs. PA vs. ZA) in a transparent way based on authentic corpus data.

In linguistic studies, anaphora is considered one of the most extensively researched topics, having garnered significant attention from linguists across various domains. Broadly, one can differentiate between formal and functional approaches. The current study aligns with the functional approaches and draws heavily on Ariel's Accessibility Theory. This theory posits that anaphora ultimately depends on the degree of accessibility of the antecedent in mental representation. For anaphora at both discourse- and sentence-level, the very same factors play a role. Additionally, this study is also influenced by Construction Grammar, which adopts a usage-based perspective and views anaphora (or the alternation of PA/NA/ZA) as a construction in a broader sense. Consequently, the three types of anaphora, like competing construction elements, differ in form and function (including semantics, pragmatics, etc.) and these differences are reflected in statistical probability distributions.

The present study aims to elucidate the distributional differences of PA, NA, and ZA in relation to contextual factors. By utilizing authentic corpus data and employing *multinomial logistic regression*, it seeks to identify and weigh up potential contextual influencing factors. This is done within the framework of Accessibility Theory to extract broader principles governing the choice of anaphoric forms in German. This leads to the following research questions:

1) Which contextual influencing factors emerge empirically as key determinants in the choice between PA, NA, and ZA in German out of several potential factors?
2) How do the key factors identified influence the choice between PA, NA, and ZA?
3) What are the typical contexts (i. e., combination of contextual factors) for PA, NA, and ZA?

To address these research questions, a multifactorial analysis was conducted. Initially, a balanced corpus, modeled after the Brown Corpus (1961), was created, from which a total of 3,000 anaphor instances were identified for direct examination. For each instance, all influencing factors within its immediate context were annotated manually, the specific influencing factors and their annotation methodology being derived from existing research. The annotated data subsequently underwent statistical analysis using R. By applying *multinomial logistic regression*, this study systematically explores and evaluates the conglomerate of factors influencing anaphoric choices. It revealed that nine contextual factors play a pivotal role in the selection of anaphoric forms in German: referent animacy (Belebtheit), the syntactic function of the antecedent (AnteSynF), the syntactic function of the anaphor (AnapSynF), syntactic parallelism between the antecedent and the anaphor (Parallelismus), the position of the anaphor (AnapPos), the clause type of the anaphor (AnapTeilTyp), the referential distance between the antecedent and the anaphor (RefDist), potential competitors of the antecedent (Mitbewerber), and text type (Textsorte). They can be used to model the choice made by native German speakers with a high degree of accuracy (84.24%). From this, the typical contexts for each anaphoric form (i. e., NA, PA, and ZA) can be deduced.

To a certain degree, the results obtained align with the four dimensions of Ariel's (1990) Accessibility Theory and can further be categorized into two groups (intrinsic vs. extrinsic). Represented by factors such as Belebtheit, AnteSynF, AnapSynF, AnapPos, and Parallelismus, the intrinsic category pertains directly to the ontological salience of the reference object itself. An animate reference object, acting syntactically as a subject, with the anaphor also positioned before the verb in a subject-like position, tends to have a heightened activation level, thus facilitating increased accessibility. In contrast, the extrinsic category encompasses the factors AnapTeilTyp, RefDist, Mitbewerber, and Textsorte. It focuses on how discourse-immanent conditions can influence the activation level of the reference object. If the reference object appears in an embedded subordinate clause, there is a significant distance between the anaphor and antecedent, potential competing referents are present, and the text is academic, the activation of the reference object is likely to be hindered, thus reducing its accessibility. Against this backdrop, the choice of an anaphoric form can be understood as a balancing process between these factors. With heightened activation, there is a tendency to opt for semantically and morphologically less complex anaphoric forms like ZA or PA. Conversely, with reduced activation, the preference leans towards morphologically complex and semantically richer forms like NA.

Lastly, it is essential to address potential limitations of the current study. As Ariel (1990) herself articulated, accessibility is quite often the only condition for appropriate use of lower Accessibility Markers (primarily NA). However, languages typically impose additional syntactic constraints on expressions with higher accessibility (mainly ZA). In other words, obeying the principle of accessibility is considered more as a necessary rather than a sufficient condition. Thus, this study does not intend to challenge traditional syntactic theories concerning anaphora. Instead, it aims to provide an interpretative perspective from a functional standpoint. Methodologically, it is worth noting that the modeling technique employed in this study is designed to identify key factors. In reality, the simplest model was chosen based on the principle of Occam's razor, which states that entities should not be multiplied beyond necessity. This model offers the most explanatory power within the data. Despite the aforementioned limitations, this research is significant for

both linguistic research and foreign language didactics. By identifying key factors and their mechanisms and describing typical contexts, the decision-making process that native speakers often follow intuitively can be elucidated more transparently. Moreover, this study could provide methodological insights for future contrastive language analyses by attempting to holistically calculate distributional differences concerning influencing factors (i. e., predictor variables).

In conclusion, the article emphasizes that the choice of anaphoric form in German is far from arbitrary. Instead, it is influenced by a plethora of contextual factors. This study not only demystifies the decision-making process that native speakers subconsciously undertake but also offers invaluable insights for linguistic research and foreign language pedagogy. The research also paves the way for future contrastive language analyses, providing a robust framework and methodology.

References

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