Typology meets data-mining: the German gender system

Sebastian Fedden, Matías Guzmán Naranjo & Greville G. Corbett
(Université Sorbonne Nouvelle and University of Surrey; Universität Tübingen; & University of Surrey)

Keywords: assignment rules, data mining, gender, German, morphosyntax, typology

In recent years linguistic typology has increasingly profited from computational methods; the hope is to discover patterns in large data sets more quickly and more accurately than would be possible for a human researcher. This is commonly known as 'data mining'. A linguistic system which could benefit from such an approach is German gender.

The German gender system is a gem among the assignment systems found in the world, for the complexity of its interacting semantic, morphological and phonological assignment principles. As fast as it offers partial results it raises new questions. This is the more remarkable since there are just three gender values: masculine, feminine, and neuter. Furthermore, the basic semantic assignment rules are relatively straightforward. Much more challenging are (i) phonological assignment (investigated by Köpcke 1982, Köpcke & Zubin 1983, among others), and (ii) the relation between gender and inflection class (see Pavlov 1995, Bittner 1999, and Kürschner & Nübling 2011). And yet, despite the progress which has been made, and the great typological interest of German gender, no attempt has been made to analyse the system as a whole.

In a system as complex as German there are at least three pitfalls:

1. cherry picking: observations of alleged regularity are sometimes based on few examples and the overall applicability of these regularities is left unexplored;
2. generalizations without a baseline: thus a prediction of a particular gender value for, say, 35% of the nouns is hardly remarkable if 35% of the nouns overall are of that gender;
3. not allowing for overlapping factors: given that phonological, morphological and semantic properties may make the same gender value more probable, making a claim for a particular generalization (e.g. phonological) requires us also to eliminate the possible effects of morphology and semantics.

To avoid these pitfalls and make progress towards a holistic analysis of the German gender system, we mine a database of more than 30,000 German nouns from WebCELEX (Baayen et al. 1995), coded for gender, frequency, phonological shape, inflection class, and derived/compounded status, which we have cleaned and to which we added semantic information (human, animal, object, abstract, mass) and frequency (based on the COW corpus, Schäfer 2015). We then built a series of analogical models using machine learning algorithms (similar to Guzmán Naranjo 2020), including different combinations of predictors (morphology, semantics, phonology, inflection class).

The overall accuracy results (Figure 1) show clearly that the system is anything but arbitrary. The combined factors reach a predictive success of over 96% (top line of Figure 1). Individual

factors are also strong predictors, most notably phonological shape and inflection class. The German gender assignment system – while complex and unusual – represents a typologically well-known type: a combination of semantic and formal (morphological/phonological) assignment principles (Corbett 1991). Our conclusions relate to German gender, but we also make a larger point by showing how typologists can benefit from data mining. And we hope to reduce the ill-informed comments still made about German gender, sometimes even by linguists.

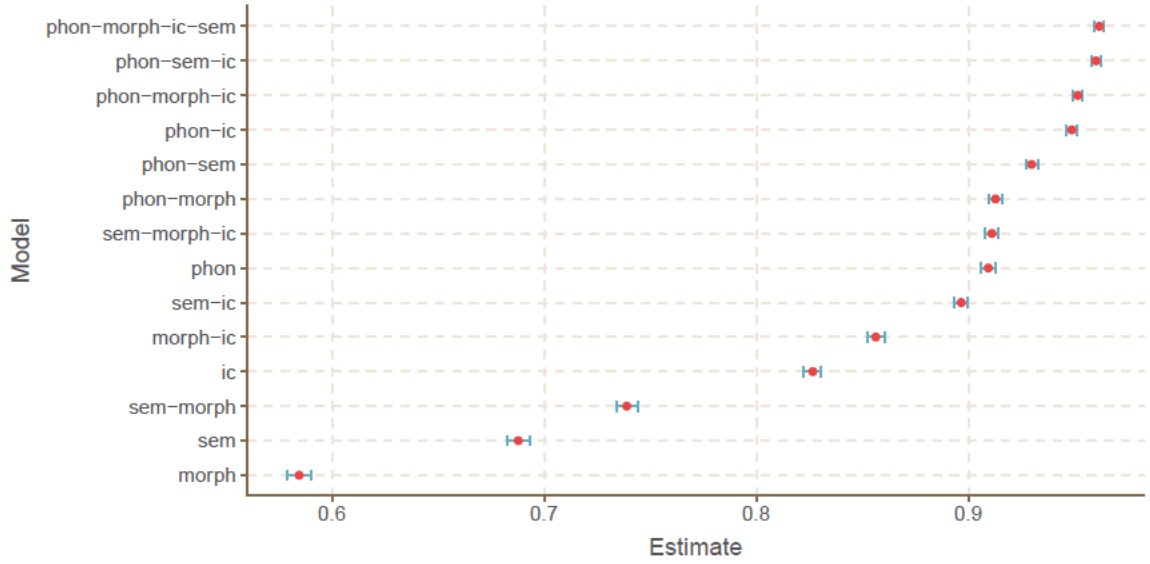

Figure 1. Accuracy and uncertainty intervals by model (ic = inflection class)

Acknowledgments: This work was partly funded by the grant "Optimal categorisation: the origin and nature of gender from a psycholinguistic perspective" (ESRC UK Grant RN0362A) and the public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR10LABX 0083).

References

Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers (1995), The CELEX Lexical Database (CD-ROM), Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Bittner, Dagmar (1999), Gender classification and the inflectional system of German nouns, in Barbara Unterbeck (ed), (1999), *Gender in Grammar and Cognition*, Part 1: *Approaches to Gender*, Berlin: Mouton de Gruyter, 1–23.

Corbett, Greville G. (1991), *Gender*, Cambridge: Cambridge University Press.

Guzmán Naranjo, Matías (2020), Analogy, complexity and predictability in the Russian nominal inflection system, *Morphology* 30(3), 219–262.

Köpcke, Klaus-Michael (1982), *Untersuchungen zum Genussystem der deutschen Gegenwartssprache*, Tübingen: Niemeyer.

Köpcke, Klaus-Michael & David A. Zubin (1983), Die kognitive Organisation der Genuszuweisung zu den einsilbigen Nomen der deutschen Gegenwartssprache, *Zeitschrift für germanistische Linguistik* 11, 166–182.

Kürschner, Sebastian & Damaris Nübling (2011), The interaction of gender and declension in Germanic languages, *Folia Linguistica* 45(2), 355–388.

Pavlov, Vladimir (1995), *Die Deklination der deutschen Substantive. Synchronie und Diachronie*, Frankfurt a. M.: Peter Lang.

Schäfer, Roland (2015), Processing and Querying Large Web Corpora with the COW14 Architecture, in *Proceedings of Challenges in the Management of Large Corpora (CMLC-3)*, 28–34.