Typological richness of the German gender system revealed by data mining

Author(s)

Affiliation(s)

1 Introduction

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.

2 A typological gem

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.

(1)	a.	ein	neu-er	Wagen
		a[NOM.M/N.SG]	new-NOM.M.SG	car(M)[NOM.SG]
		'a new car'		
	b.	ein-e	neu-e	Kutsche
		a-NOM.F.SG	new-NOM.F.SG	coach(F)[NOM.SG]
		'a new coach'		
	c.	ein	neu-es	Fahrrad
		a[NOM.M/N.SG]	new-NOM.N.SG	bicycle(N)[NOM.SG]
		'a new bicycle'		

Furthermore, the basic semantic assignment rules are relatively straightforward. Much more challenging are (i) the relation between gender and inflection class (see Augst 1975; Pavlov 1995; Bittner 1999; Kürschner & Nübling 2011) and (ii) the phonological assignment rules, investigated by Köpcke (1982) and Köpcke & Zubin (1983) among others.

3.1 Semantics

Sex-differentiable nouns, i.e. nouns which refer to male or female humans, or to male or female (higher) animals, are assigned gender on the basis of biological sex: e.g. *der Mann* 'man', *die Frau* 'woman', *der Eber* 'wild boar', *die Bache* 'wild sow'. In addition there are various non-core semantic assignment rules, some of which are highly specific and yet surprisingly robust.

3.2 Word formation

Those German nouns which are morphologically complex are governed by the Last Member Principle (*Letzt-Glied-Prinzip*, see Köpcke & Zubin 1984: 28-29, and references there): the gender of the whole word is determined by the gender of the last element. In compounds the last element is a word with its own gender value. For example *der Mutterschutz* 'maternity' consists of the feminine first member *die Mutter* 'mother' and the masculine last member *der Schutz* 'protection'; by the Last Member Principle it is masculine. Derivational affixes are similarly associated with a gender value, which is assigned to the derived word irrespective of the gender of the base (if this is a noun). For example, the suffix *-schaft* derives feminine nouns, e.g. *die Freundschaft* 'friendship' from the masculine noun *der Freund* 'friend', or *die Landschaft* 'landscape' from the neuter noun *das Land* 'land'.

3.3 Inflection

For many instances, gender can unambiguously (or nearly unambiguously) be predicted from inflection class, e.g. all nouns which inflect like *die Lampe* 'lamp', are feminine. Then, as a reduced prediction, there are several inflection classes whose nouns can be masculine or neuter but not feminine. For instance, we can predict that *Knopf* 'button' cannot be feminine based on its paradigm.

3.4 Phonology

Köpcke (1982) and Köpcke & Zubin (1983) establish a number of phonological rules to account for the gender of monosyllabic nouns. For example, almost all monosyllabic nouns starting with the cluster /kn/ are masculine (93%), e.g. *der Knopf*, 'button', *der Knick* 'crease', the only exception being the neuter noun *das Knie* 'knee'. The majority of nouns which end in the clusters /ft/, /xt/ or /çt/ are feminine (64%), e.g. *die Zunft* 'guild', *die Frucht* 'fruit', *die Sicht* 'visibility'. And in general, the more consonants a monosyllabic noun has in its onset or coda, the higher the probability that the noun is masculine.

This body of research has demonstrated clear regularities in the assignment of gender to German nouns. And yet, despite this progress in understanding parts of the system, and the great typological interest of German gender, no attempt has been made to analyse the system as a whole.

3 Pitfalls in the analysis of German gender

In analysing a system as complex as German there are at least three potential 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; without a clear baseline we do not know how successful a rule is compared to pure chance;
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.

4 Data mining

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. We have cleaned this database, and we have 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 Extreme Gradient Boosting trees (similar to Guzmán Naranjo 2020), including different combinations of predictors (morphology, semantics, phonology, inflection class). The baselines in this dataset are approximately 35% masculine, 45% feminine and 20% neuter.

Our choice of using a Boosting Tree model (Chen and Guestri 2016) is purely pragmatic, this type of model has been shown to work very well for this type of task (Bonami and Pellegrini, forth.). In our case, the models learn to predict the gender of a given noun based on a set of predictors.

To include the morphological predictors, inflection class predictors and hand-annotated semantic predictors is a simple matter of adding factors to the model. For phonology and semantics we use a technique based on similarity neighbourhoods. For phonology we calculate a right-hand-side weighted Levenshtein distance between all nouns (we use the phonological transcription). For semantics we induce gender-neutralized semantic vectors using Word2Vec from the COW corpus and calculate a cosine distance matrix between all nouns. With these distance matrices we extract for each noun the nearest five neuter neighbours, nearest five feminine neighbours and nearest five masculine neighbours (once for semantics and once for phonology). We then use these distance values as our phonological and semantic predictors. The intuition behind this technique is that the gender of a noun depends (in part) on how similar phonologically or semantically the noun is to other neuter, masculine and feminine nouns. While using a simple *K*-nearest neighbours algorithm also works, we found that our implementation performed considerably better.

For all models we report the 10-fold cross-validation accuracy. In cross-validation we divide our dataset in 10 groups; and fit a model leaving out one of the 10 groups, we then try to predict the gender of the nouns in the omitted group. We repeat this process for all groups.

5 Results

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).

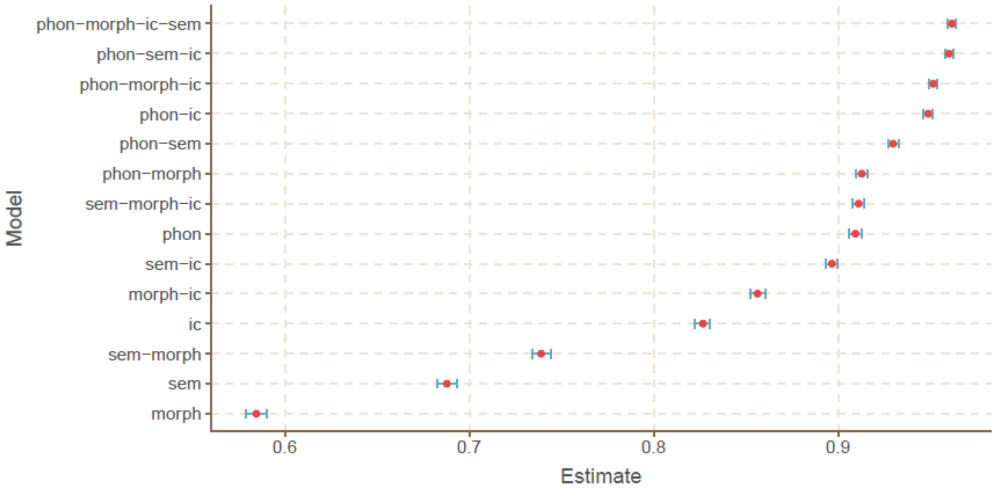

Figure 1. Accuracy and uncertainty intervals by model
Abbreviations: sem - semantics, phon - phonological shape of the stem,
morph - morphologically complex (derived) nouns, ic - inflection class

6 Conclusions

Our conclusions relate first to German gender, where we see increasingly clearly the interlocking regularities of the system. We hope to reduce the ill-informed comments still made about German gender, sometimes even by linguists. Second, we make a larger point by showing how typologists can benefit from data mining. From both sides of the collaboration, it is important to keep asking what the generalizations which are established actually mean. In earlier work on gender assignment there was an emphasis on distinguishing regularities from each other, to establish which was responsible for a particular assignment. The current work takes a broader view of inter-connected and mutually reinforcing regularities. And for this, German is indeed a typological gem.

References

Augst, Gerhard. 1975. *Untersuchungen zum Morpheminventar der deutschen Gegenwartssprache*, Forschungsberichte des Instituts für deutsche Sprache Mannheim 25, Tübingen: Gunter Narr.

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.), *Gender in Grammar and Gognition*, Part 1: *Approaches to Gender*, 1–23. Berlin: Mouton de Gruyter.

Bonami, Olivier & Matteo Pellegrini. *Derivation predicting inflection. The role of families, series and morphotactics*. Paper presented at the *19th International Morphology Meeting*, Vienna, February 2020.

Chen, Tianqi & Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

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. 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.

Köpcke, Klaus-Michael & David A. Zubin. 1984. Sechs Prinzipien für die Genuszuweisung im Deutschen. Ein Beitrag zur natürlichen Klassifikation. *Linguistische Berichte* 93. 26–50.

Kürschner, Sebastian & Damaris Nübling. 2011. The interaction of gender and declension in Germanic languages. *Folia Linguistica* 45. 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.