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

SLE Athens/SLE 2021 Platform
1 September 2021

Thanks to the ESRC, ANR, Labex and the Centre for Advanced Study (CAS), Oslo

Outline

1. **Typology**: why is German gender special?
2. German gender assignment: the essentials
3. **Data-mining**: new data and new methods
4. Results
5. Nested generalizations
6. Conclusions

1. Typology: why is German gender special?

Fedden, Guzmán Naranjo & Corbett:
Typology meets data mining

Typology

Gender assignment systems:

- semantic (Tamil, Bininj Gun-wok)
- semantic and formal
- morphological (Russian)
- phonological (Qafar)

German

Arguably the most complex assignment system

- 378 assignment principles in the literature, to date (Corteen 2019)
  - 159 based on semantic properties
  - 219 based on the formal properties
  - and they conflict
- Yet, of course, children learn the system (Mills 1986)

BUT only three German gender values

(1) a. der Film
DEF.NOM.M.SG film(M)[NOM.SG]
'the film'

b. die Symphonie
DEF.NOM.F.SG symphony(F)[NOM.SG]
'the symphony'

c. das Buch
DEF.NOM.N.SG book(N)[NOM.SG]
'the book'

Previous research

- Notable work by Klaus-Michael Köpcke and David Zubin on semantics and phonology
  - Köpcke 1982; Köpcke & Zubin 1983, 1984, 1996; Zubin & Köpcke 1984

- Specific interest in the relation between gender and inflection
  - Augst 1975; Pavlov 1995; Bittner 1999; Kürschner & Nübling 2011

2. German gender assignment: the essentials

Fedden, Guzmán Naranjo & Corbett:
Typology meets data mining

Basic semantic assignment principles

- Sex-differentiable nouns, i.e. nouns which refer to female or male humans or to female or male (higher) animals, are assigned gender on the basis of sex:

- *die Frau* ‘woman’, *die Kuh* ‘cow’ → F

- *der Mann* ‘man’, *der Bulle* ‘bull’ → M

Formal assignment principles

- Formal assignment for most German nouns
  - morphology (compounding, derivation, inflection class)
  - phonology (stem shape)

Word formation

- Compounding, derivation
- Last Member Principle (*Letzt-Glied-Prinzip)* :

The gender of the whole word is determined by the gender of the last element.

(Köpcke & Zubin 1984: 28-29,
and references there)

Compounding

- Last member determines gender:

  die Mutter ‘mother’ + *der Schutz* ‘protection’ → *der Mutterschutz* ‘maternity’

  [[Mutter][schutz]]
  F *M* → M

Derivation

- Derivational affixes are associated with a gender value
- This value is assigned to the derived word (irrespective of the gender of the base if this is a noun)
- for instance, suffix *-schaft* → F, e.g.
  - *der Freund* 'friend' → *die Freundschaft* 'friendship'
  - *das Land* 'land' → *die Landschaft* 'landscape'

Inflection

- Inflection class is important, e.g. nouns which inflect like *Lampe* 'lamp' (genitive SG *-∅*, nominative PL *-(ə)n*) are all F

'lamp'	SG	PL
NOM	Lampe	Lampen
ACC	Lampe	Lampen
GEN	Lampe	Lampen
DAT	Lampe	Lampen

- For at least four inflection classes, gender can be predicted unambiguously (or nearly unambiguously)

Inflection

- Several Inflection classes allow partial prediction of gender, e.g. *der Knopf* 'button' from IC *-(ə)s*/Umlaut+ *-e*

'button'	SG	PL
NOM	Knopf	Knöpfe
ACC	Knopf	Knöpfe
GEN	Knopf(e)s	Knöpfe
DAT	Knopf	Knöpfen

- → M *or* N

Phonology

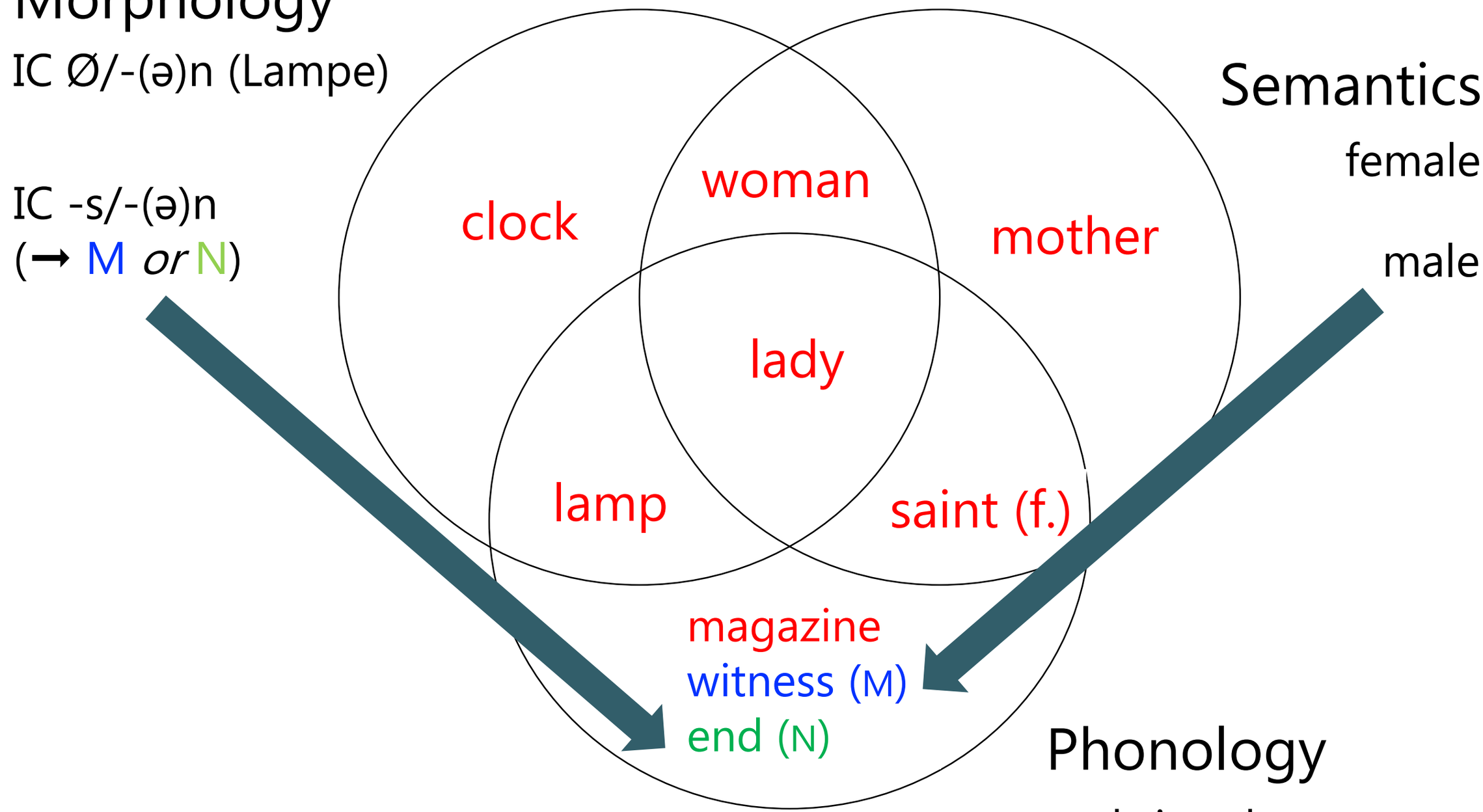
- Phonological cues can help:
  - 64% of monosyllabic nouns are masculine
    - 93% of monosyllabic nouns with initial /kn/ are masculine (exception *das Knie* ‘knee’)
- Phonological assignment principles studied by Köpcke (1982); Köpcke & Zubin (1983); Köpcke & Zubin (1984)
- Phonological assignment rules are not exceptionless

Overlap of predictors

Morphology

IC Ø/-(ə)n (Lampe)

IC -s/-(ə)n
(➞ M *or* N)



Semantics

female

male

Phonology

ends in schwa

3. Data-mining: new data and new methods

Fedden, Guzmán Naranjo & Corbett:
Typology meets data mining

Three pitfalls

1. “Cherry picking”: suggested regularities are sometimes based on positive data only

2. Generalizations without a baseline

3. Missing the importance of overlapping factors

- We take a more holistic view

New data

- WebCelex (Baayen, Piepenbrock & Gulikers 1995), available at: http://celex.mpi.nl
  - Based on the Mannheim corpus (~6 million words, ~31.000 noun types)
- Annotated for gender, phonology, inflection class, and derived/compounded status
- We added semantic information (human, animal, object, abstract, mass) and frequency (based on the much larger COW corpus; Schäfer & Bildhauer 2012, Schäfer 2015)
- We fixed inconsistencies in the original annotation

Overall picture

- Our aim: To put numbers to predictions
- Overall gender distribution based on ~31,000 nouns

M: 36% **F: 45%** **N: 19%**

- Baseline against which we will measure predictor accuracy

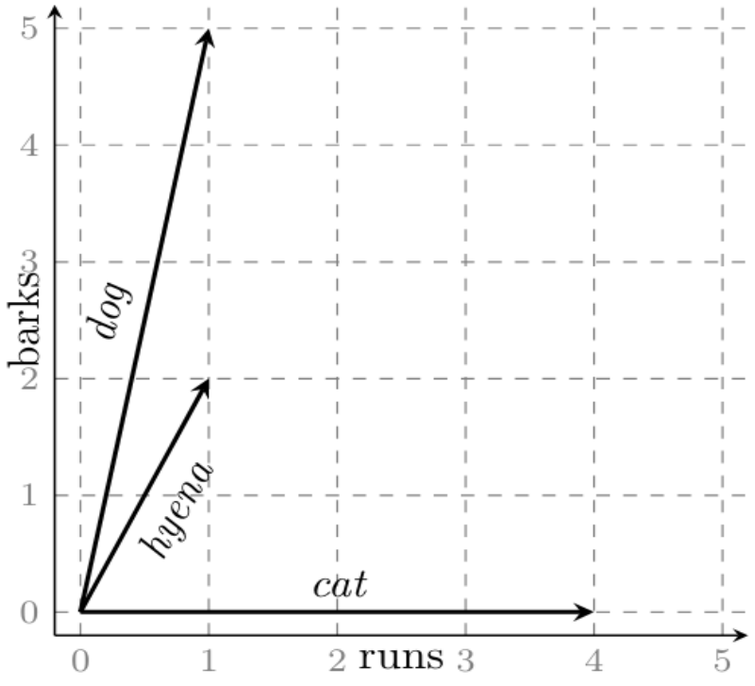
New methods

- Semantic and phonological representation
- Analogical principle: Words which are semantically or phonologically similar will belong to the same class
  - (Bybee & Slobin 1982; Skousen 1989; Daelemans et. al 1998; Albright & Hayes 2002; Matthews 2005; Guzmán Naranjo 2019, 2020)
- Therefore, we represent the semantics and phonology of nouns as their similarity to other masculine, feminine and neuter nouns

Semantic distances

- Distributional view of semantics (Firth 1957): the meaning of a word is given by its distribution

	runs	barks
dog	1	5
hyena	1	2
cat	4	0



- This implies that words which have a similar distribution in a corpus have similar meaning
- We can use semantic vectors to represent and calculate the semantic distance between words (using the COW corpus)

Phonological distances

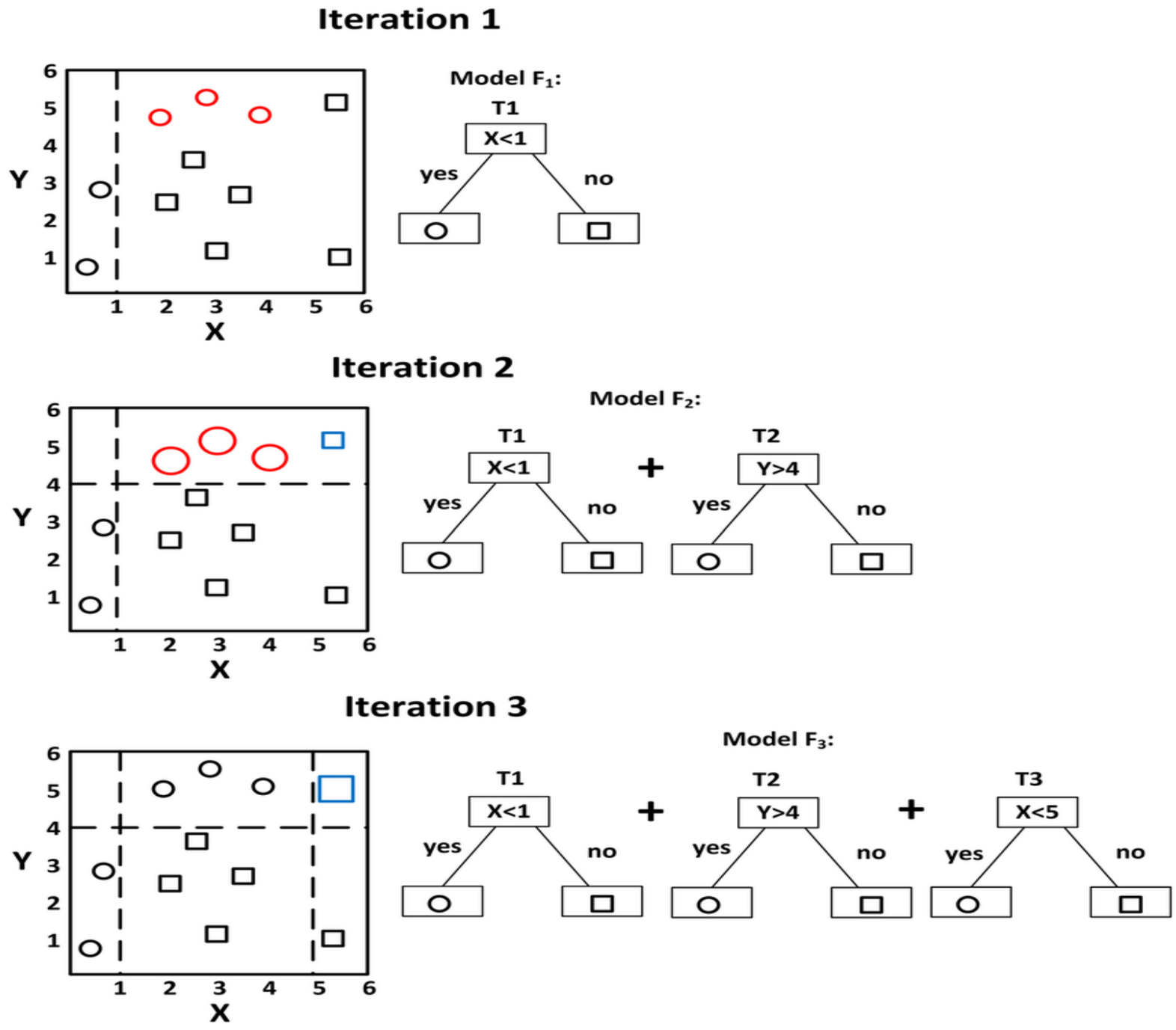
- Right-edge weighted Levenshtein distances
- We calculate the phonological distance between all nouns. We assume that (for German) similarity on the right hand side is more important than similarity on the left hand side:
  - Distance between /baʊ**m**/ ‘tree’ and /baʊ**x**/ ‘belly’: 1
  - Distance between /**b**aʊm/ ‘tree’ and /**z**aʊm/ ‘seam’: 0.25

New methods

- Instead of finding rules manually, we use a **Boosting Tree** model:
  - build many small decision trees
  - join them together into a stronger system
  - system learns automatically all rules and their interactions

- Our predictors are:
  - Semantics (lexical: sex, animacy, concrete, mass)
  - Semantics (distance to the 5 nearest M, F and N nouns)
  - Phonology (distance to the 5 nearest M, F and N nouns)
  - Derivation
  - Inflection class

Illustrating Boosting Trees

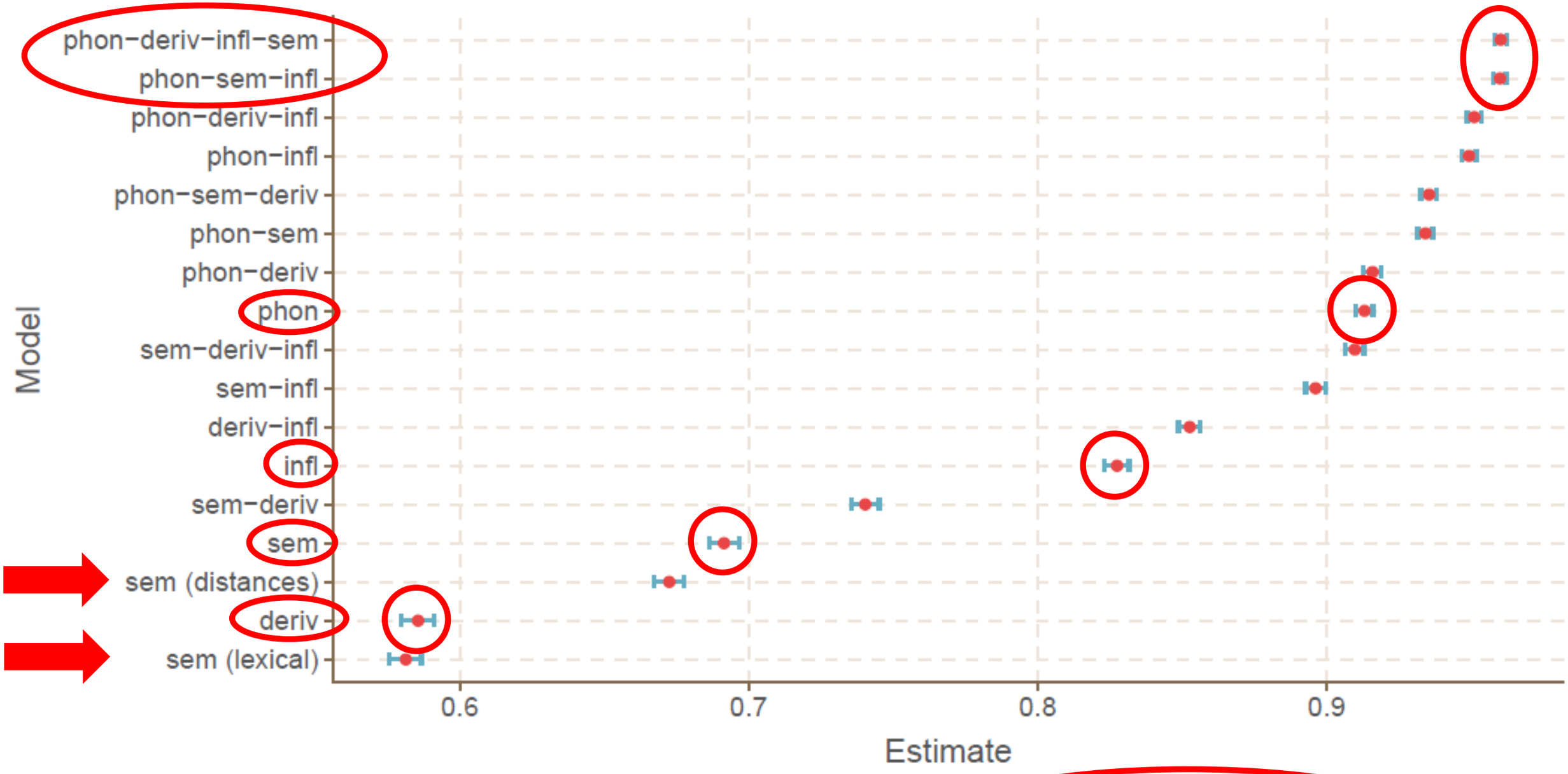


(Zhang et al. 2018)

4. Results

Fedden, Guzmán Naranjo & Corbett:
Typology meets data mining

Overall results



Accuracy and uncertainty intervals by model (majority case baseline: .45)

Variable importance (models)

model to be evaluated	overall accuracy	minus	overall accuracy w/out the model to be evaluated	equals	variable importance
PHON:	0.96	–	0.91	=	0.05
INFL:			0.935		0.025
SEM:			0.951		0.009
DERIV:			0.96		0

Additive vs. subtractive variable importance

- Additive variable importance
  - not worried about overlap, multiple predictors can point in the same direction
  - assignment principles **reinforce** each other
- Subtractive variable importance
  - eliminate all redundancy
  - **parsimonious** assignment principles

Subtractive variable importance

predictors	total remaining	number of nouns predicted	accuracy
phonology (distances)	1584	1201	0.76
semantics (distances)	597	218	0.37
inflection class	564	185	0.33
derivation	459	80	0.17
semantics (lexical)	436	57	0.13
Köpckian rules	1584	3	0.002

5. Nested generalizations

Fedden, Guzmán Naranjo & Corbett:
Typology meets data mining

Example - Nested generalizations: two-way

Morphology

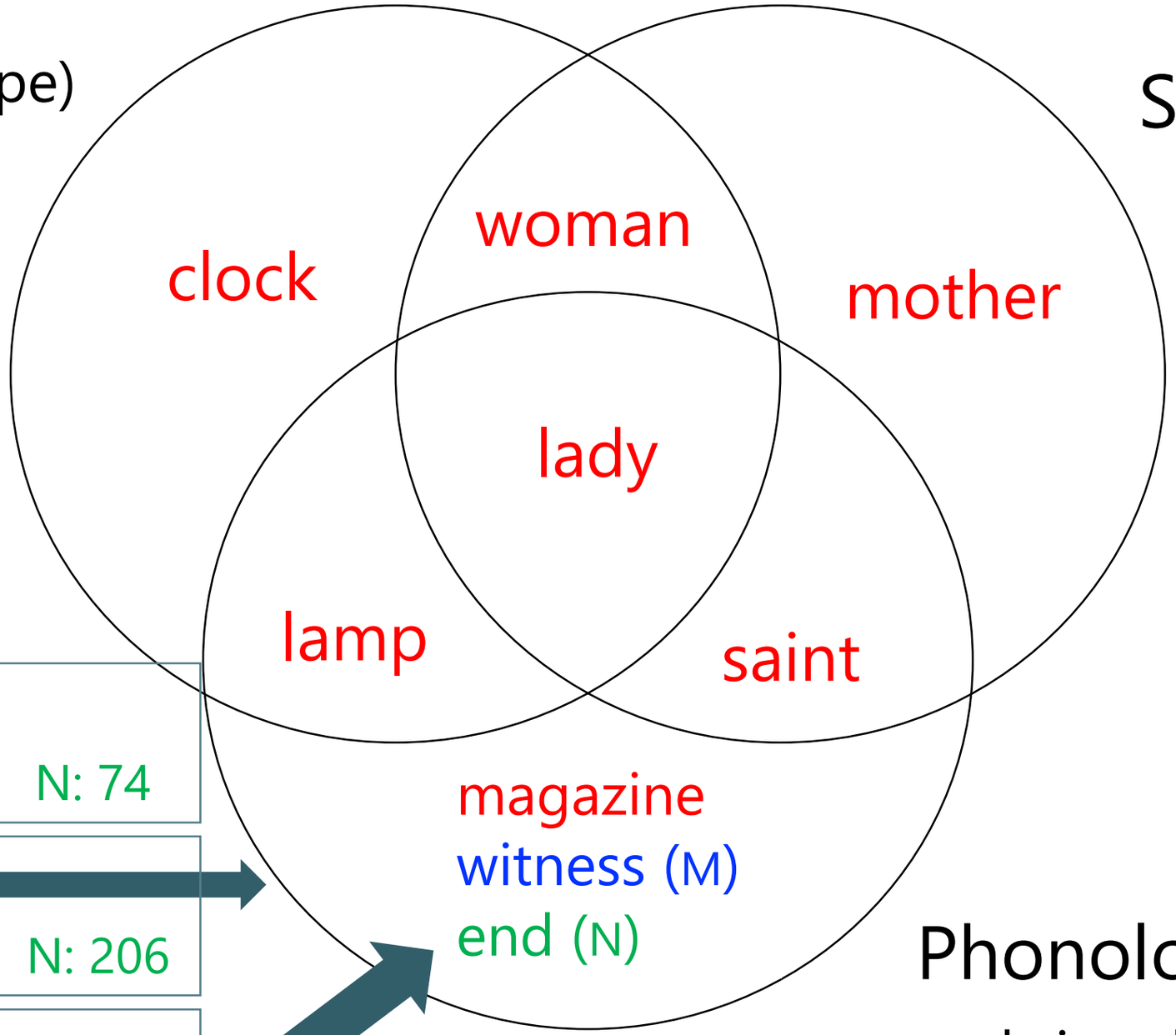
IC Ø/-(ə)n (Lampe)

IC -s/-(ə)n
(→ M *or* N)

IC -s/-(ə)n
F: 0 M: 260 N: 74

Ends in schwa
F: 4195 M: 557 N: 206

Both
F: 0 M: 0 N: 19



Semantics

female

Phonology

ends in schwa

Nested generalizations: two-way

predictors	total remaining	number of nouns predicted	accuracy
sem-infl	2668	1011	0.38
phon-infl	944	148	0.16
phon-sem	1128	154	0.14
sem-deriv	4924	334	0.07
phon-deriv	1710	89	0.05
deriv-infl	3278	68	0.02

Nested generalizations: three-way

predictors	total remaining	number of nouns predicted	accuracy
sem-deriv-infl	1050	27	0.03
phon-infl-sem	265	6	0.02
phon-deriv-sem	656	9	0.01
phon-deriv-infl	596	6	0.01

6. Conclusions

Fedden, Guzmán Naranjo & Corbett:
Typology meets data mining

Conclusions

- **Typology** meets data-mining
  - German fits the typology of assignment well
  - semantic plus formal principles:
    - the latter primarily phonological and inflection class
- Typology meets **data-mining**
  - accurate, discover patterns in large data sets quickly and accurately
  - broader view of inter-connected and mutually reinforcing regularities
- Typology **meets** data-mining
  - exciting results
  - only possible together

thank you very much

Bibliography available
in the slides

Bibliography

Albright, A. & B. Hayes. 2002. Modeling English past tense intuitions with minimal generalization. *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* 6. 58-69.

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

Baayen, R. Harald, R. Piepenbrock & L. 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 cognition*, Part 1: *Approaches to gender*, 1-23. Berlin: Mouton de Gruyter.

Bybee, J. L. & D. I. Slobin. 1982. Rules and Schemas in the Development and Use of the English past Tense. *Language* 58(2). 265-289.

Corteen, Emma C. 2019. The Assignment of Grammatical Gender in German: Testing Optimal Gender Assignment Theory (Doctoral thesis). https://doi.org/10.17863/CAM.37638

Bibliography

Daelemans, W., J. Zavrel, K. van der Sloot & A. van den Bosch. 1998. *TiMBL: Tilburg Memory-Based Learner*. Universiteit van Tilburg. https://research.tilburguniversity.edu/en/publications/timbl-tilburg-memory-based-learner-version-10-reference-guide

Firth, J. R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, 1-32.

Guzmán Naranjo, M. 2019. *Analogical classification in formal grammar*. Language Science Press. https://doi.org/10.5281/zenodo.3191825

Guzmán Naranjo, M. 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.

Bibliography

Köpcke, Klaus-Michael & David A. Zubin. 1996. Prinzipien für Genuszuweisung im Deutschen. In Ewald Lang & Gisela Zifonun (eds.), *Deutsch-typologisch*. Jahrbuch des Instituts für Deutsche Sprache 1995, 473-491. Berlin: De Gruyter Mouton.

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

Matthews, C. A. 2005. French Gender Attribution on the Basis of Similarity: A Comparison Between AM and Connectionist Models. *Journal of Quantitative Linguistics* 12. 262-296.

Mills, Anne E. 1986. *The acquisition of gender: A study of English and German*. Berlin: Springer.

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. *Proceedings of Challenges in the Management of Large Corpora (CMLC-3)* (IDS publication server), 28-34.

Schäfer, Roland & Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 486-493.

Bibliography

Skousen, R. 1989. *Analogical modeling of language*. Kluwer Academic Publishers.

Zhang, Z., G. Mayer, Y. Dauvilliers, G. Plazzi, F. Pizza, R. Fronczek, ... & R. Khatami. 2018. Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from European Narcolepsy Network database with machine learning. *Scientific reports*, 8(1). 1-11.

Zubin, David & Klaus-Michael Köpcke. 1984. Affect classification in the German gender system. *Lingua* 63. 41-96.