

Supplementary Materials: Typology meets statistical modeling: the German gender system

Sebastian Fedden^{1,2,4} Matías Guzmán Naranjo³ Greville G. Corbett⁴

¹*Université Sorbonne Nouvelle/LACITO*, ²*Ludwig-Maximilians-Universität München*, ³*Universität Freiburg*, ⁴*University of Surrey*

In these supplementary materials we provide (i) model comparisons, (ii) a list of all predictors and their values, and (iii) a list of changes to the CELEX database (none of which led to significant differences in the results).

SECTION 1. MODEL COMPARISONS. As we mention in the paper, there is no deep reason for choosing boosting trees over other machine learning algorithms. The main reason was implementation efficiency and overall accuracy. In this section we provide some comparisons to other algorithms and show that we can obtain very similar results as long as we keep the data representation in terms of distances to neighbors.

While there are many different alternatives, we choose to focus on two here: multinomial (i.e. categorical) logit regression, and simple conditional inference trees. We present comparisons of three predictor groups which are the most crucial: all predictors, only phonology, only semantic vectors.

ACCURACY COMPARISON FOR MULTINOMIAL AND CONDITIONAL INFERENCE TREES. In all three cases the boosting trees perform better than the multinomial model and than the conditional inference tree model.

Phonology:

model	accuracy
boosting tree	0.91
multinomial	0.89
conditional inference tree	0.89

Semantics (vectors):

model	accuracy
boosting tree	0.66
multinomial	0.65
conditional inference tree	0.62

Phonology+Semantics+Derivation+Inflection:

model	accuracy
boosting tree	0.96
multinomial	0.95
conditional inference tree	0.92

AIC-BASED FEATURE SELECTION. A reviewer suggested using a more traditional approach to model selection, and comparing that to the results we obtain using cross-validation. Here we present the AIC comparison for models fitted using multinomial logit regression and compare these to the results presented in the paper.

Table 1. shows the AIC vs accuracy for the binomial model, and comparison with the accuracy values of the boosting tree model. We observe that using AIC produces a very similar ordering of variable importance to using accuracy and cross-validation, with either multinomial or boosting trees. Boosting trees are between 1 and 2 percentage points more accurate than multinomial models, and this shows especially in the phonology. While the multinomial model does infer that morphology plays an additional role to phonology, the boosting tree model does not.

Table 1. AIC vs accuracy for the binomial model, and comparison with the accuracy values of the boosting tree model

N	predictors	AIC	Acc	Acc (XGBoost)
1	phon ic sem	7869	0.95	0.96
2	phon morph ic sem	9380	0.95	0.96
3	phon ic	9921	0.94	0.95
4	phon ic morph	11146	0.94	0.95
5	phon sem	12484	0.92	0.93
6	phon sem morph	13779	0.93	0.93
7	sem ic morph	15004	0.9	0.91
8	sem ic	15435	0.88	0.89
9	phon	17133	0.89	0.91
10	phon morph	17891	0.9	0.91
11	morph ic	18681	0.85	0.85

12	ic	20985	0.83		0.83
13	singular	26885	0.79		0.79
14	sem morph	37374	0.73		0.74
15	plural	37527	0.73		0.74
16	sem	43983	0.68		0.69
17	morph	48080	0.59		0.59
18	sem (distances)	48700	0.65		0.66
19	sem (lexical)	52900	0.56		0.58

SECTION 2. PREDICTOR LEVELS. Here we provide a full listing of all predictors and their values.

Animacy: Animate, Inanimate, Human, Indeterminate

Sex: Epicene, None, Male, Female, Indeterminate

Concrete: Yes, No, Indeterminate

Mass: Yes, No, Indeterminate

Inflection-class: A, S0/P10, S0/P1u, S0/P2, S0/P3, S0/P4u, S1/P0, S1/P1, S1/P1/P10b, S1/P1/P10c, S1/P1/P10j, S1/P1/P10k, S1/P1/P1u, S1/P1/P1u/P5, S1/P1/P2, S1/P1/P3, S1/P1/P3/P5, S1/P1/P4, S1/P1/P4u, S1/P1/P5, S1/P1/P5/P2, S1/P10a, S1/P10a/P10k, S1/P10a/P5/P10i, S1/P10b, S1/P10b/P1, S1/P10c, S1/P10c/P1, S1/P10c/P10ai, S1/P10d, S1/P10d/P10ai, S1/P10e, S1/P10f/P4u, S1/P10h, S1/P10i, S1/P10i/P5, S1/P10j/P10ah, S1/P10k, S1/P10l, S1/P10m, S1/P1u, S1/P1u/P1, S1/P1u/P3, S1/P1u/P4u, S1/P1u/P5, S1/P2, S1/P2/P1/P3, S1/P2/P10l, S1/P2/P10m, S1/P2/P2u, S1/P2/P3, S1/P2/P4, S1/P2/P5, S1/P2u, S1/P2u/P2, S1/P3, S1/P3/P1, S1/P3/P5, S1/P3/P1, S1/P4, S1/P4/P1, S1/P4/P10g, S1/P4/P5, S1/P4u, S1/P4u/P1, S1/P4u/P1/P1u, S1/P4u/P3, S1/P5, S1/P5/P1, S1/P5/P10a, S1/P5/P10ak, S1/P5/P10c, S1/P5/P10e, S1/P5/P10i, S1/P5/P10k, S1/P5/P10k/P10a, S1/P5/P1u, S1/P5/P2, S1/P5/P3, S1/P5/P8, S1/P7, S1/P8, S1/P8/P9, S1/P9, S1/P9/P8, S1/S2/P1/P2, S1/S2/P1/P3, S1/S2/P3, S1/S3/P0, S1/S3/P1/P10m, S1/S3/P1/P3, S1/S3/P10i, S1/S3/P10s/P1, S1/S3/P2, S1/S3/P2/P5, S1/S3/P3/P1, S1/S3/P5, S1/S3/P5/P1, S1/S3/P5/P1/P2, S1/S3/P5/P2, S2/P0, S2/P1/P3, S2/P3, S2/P3/P10q, S2/S1/P0, S2/S1/P1/P3, S2/S1/P3, S2/S1/P3/P1, S3/P0, S3/P0/P3, S3/P1, S3/P1/P10q, S3/P1/P2, S3/P1/P3/P5, S3/P10_french, S3/P10_unclear, S3/P10a, S3/P10a/P10ab, S3/P10a/P5, S3/P10a/P5/P10ab, S3/P10aa, S3/P10ab, S3/P10ad, S3/P10ae/P2, S3/P10af, S3/P10ag, S3/P10c, S3/P10c/P10ai, S3/P10j, S3/P10k, S3/P10n, S3/P10o, S3/P10p, S3/P10p/P1, S3/P10p/P10a/P1, S3/P10q, S3/P10r, S3/P10t, S3/P10u, S3/P10v, S3/P10w, S3/P10x, S3/P10y, S3/P10z, S3/P1u, S3/P1u/P1, S3/P2, S3/P2/P10a, S3/P2/P10p, S3/P2/P10v, S3/P2/P3, S3/P2/P5, S3/P2/P7, S3/P2u, S3/P3, S3/P3/P1u, S3/P5, S3/P5/P1, S3/P5/P10a, S3/P5/P10ab, S3/P5/P10aj, S3/P5/P10i, S3/P5/P2, S3/P5/P3, S3/P7, S3/P7/P10ag, S3/P7/P2, S3/P9, S3/P9/P1, S3/P9/P10p, S3/S1/P0, S3/S1/P1, S3/S1/P10v/P7, S3/S1/P2, S3/S1/P2/P1, S3/S1/P2/P5, S3/S1/P2/P5/P3, S3/S1/P2/P7, S3/S1/P3/P5/P10i,

S3/S1/P5, S3/S1/P5/P2, S3/S1/P7, S3/S1/P7/P10a, S3/S1/P7/P2/P10p, S4/P0, S4/P7, S5/P0, S5/P3, S5/S2/P3

Suffix: -a, -ade, -aer, -aet, -age, -al, -an, -and, -aner, -ant, -anz, -ar, -arier, -asmus, -ast, -at, -atur, -bold, -chen, -dar, -de, -e, -ei, -el, -elei, -ell, -elle, -en, -end, -ens, -ent, -enz, -er, -erei, -erich, -erie, -erin, -eske, -ess, -esse, -et, -ett, -ette, -eur, -euse, -eut, -heit, -i:n, -iakum, -ial, -ian, -iast, -iat, -icht, -id, -ide, -ie, -ien, -ient, -ier, -iere, -ik, -ikum, -ille, -in, -ina, -ine, -ing, -ion, -ise, -ismus, -issimus, -ist, -it, -itis, -itum, -ium, -iv, -keit, -lein, -ler, -ling, -ment, -mus, -ner, -nis, -o, -oer, -oid, -on, -or, -ose, -s, -sal, -schaft, -sel, -st, -ste, -stel, -t, -taet, -tel, -ter, -thek, -tik, -tion, -tum, -um, -ung, -ur, -us

Prefix: ab-, an-, an- (2), anti-, auf-, aus-, be-, bei-, des-, dis-, durch-, ein-, erz-, ge-, il-, im-, im3-, in-, in- (2), in- (3), inter-, ko-, kon-, miss-, mit-, mono-, nach-, prae-, pro-, re-, rueck-, trans-, ueber-, ultra-, um-, un-, unter-, ur-, ver-, vor-, wider-, zu-

We distinguish between two different *an-* prefixes, and three different *in-* prefixes.

SECTION 3. CHANGELOG. This section compiles all changes made to the original CELEX database. The changes are mainly additive, but also include corrections and removal of data. We have run all our calculations on vanilla CELEX, as well as on the revised CELEX, and the results are not significantly different.

1. Changes to cells

a. Inflection class: We changed a total of 595 cells, of these

266 were errors, 241 of which were reassignments of motion nouns in *-in* (*Lehrerin* ‘female teacher’, etc.) to the same inflection class as *Biene* ‘bee’, CELEX had a separate (in our view unwarranted) inflection class for them.

330 nouns (almost exclusively loans), which CELEX lumped these together as ‘irregular’ in the plural, have been assigned to one of a set of small inflection classes.

b. Gender: We changed a total of 3 cells (all errors).

c. Proper name: We changed a total of 3 cells (all errors).

2. Rows eliminated

1 pronoun
1 duplicate entry
29 letters of the alphabet
24 orthographic variants
313 hybrids

3. Rows added

We added 241 rows: (i) for nouns that were coded as MF, MN, FN, or MFN in CELEX, so *Selerie* ‘celery’ (MF in CELEX) has 2 rows one M and one F in our database, or (ii) for nouns which belong to two different inflectional paradigms

4. Columns added

We added five columns: animacy (human, animal, inanimate), sex (male, female, epicene, none), concreteness (yes, no), hybrid (yes, no) and count (yes, no). These give us lexical semantic information which we compare with distributional semantics.

Morphological status

Original CELEX does not differentiate between compounds and derivations, and codes all of these as C 'complex'. We have automatically divided all complex words into compounds and derivations, and then manually checked all of them for consistency. This means changes in at least 20,000 cells for morphological status.