Regressing morphological differences on an empirical measure of learning difficulty

Job Schepens

Centre for Language Studies, Radboud University Nijmegen, the Netherlands
International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands.

Levinson & Gray (2012) argue that linguistic diversity takes up a key position in the cognitive sciences because the underlying evolutionary processes of diversity indicate what we are capable of. The concept of linguistic distance is inherently important because it quantifies linguistic diversity. Linguistic distance can be measured in degrees of evolutionary change over time (Gray & Atkinson, 2003), but this restricts quantification of linguistic distance within language families only. Newly available typological data also enable cross-family quantification of linguistic distance. It is unclear however how to weigh the structural features of typological data for a measure of linguistic distance. What is needed is a hierarchy of features according to their importance for a measure of linguistic distance.

Establishing a hierarchy of features may be possible by using an externally defined, empirical measure of linguistic distance, such as learning difficulty. Learning difficulty of Dutch can be used as a measure of linguistic distance by aggregating language testing scores on the basis of the learner's mother tongues (Van der Slik, 2010). In the Netherlands, a large database of language testing scores is available consisting of over 50,000 test scores on the state exam "Dutch as a Second Language". The scores on this exam are available for learners from all over world.

We aim to expose a feature hierarchy in one sub domain of typological configurations, namely morphological complexity. Encapsulating feature variance to the restricted domain of morphological complexity avoids normalization problems of number of features included and reduces artificial effects of missing data across multiple typological areas. Lupyan & Dale (2010) quantified morphological complexity using 28 structural features from the World Atlas of Language Structures (WALS; Dryer & Haspelmath, 2011). According to Lupyan & Dale's (2010) coding scheme, Dutch shows some degree of morphological complexity in at least 6 of the 28 features. Lupyan & Dale (2010) showed that the morphological complexity of a language is correlated with the number of speakers of a language. This correlation could have evolved through varying patterns of L2 acquisition across languages. This hypothesis provides us with a rationale to choose morphological complexity as sub domain for our study since we are interested in exposing feature dynamics in L2 learning difficulty. For other explanations of how this variable evolved see Sampson, Gil, & Trudgill (2009). Regression of morphological differences between languages (the mother tongues of the learners) and Dutch on differences in learning difficulty learners have may expose a feature hierarchy of cross-linguistic differences in the domain of morphological complexity.

In Wieling, Nerbonne, & Baayen (2011), pronunciation distance from standard Dutch to pronunciation variants was regressed on several social and geographical characteristics by sampling pronunciation distances from of a large number of words and locations in the Netherlands. Here, morphological distance from Dutch to other languages was regressed on learning difficulty by sampling morphological differences from 28 features and 72 different languages from 29 genera and 13 families. Forty of these languages were from the Indo-European language family tree. A random effect structure of features crossed with languages allowed an estimation of variance components between features and languages independently. We fitted a logistic mixed effects regression model and used its fitted values for a feature hierarchy of morphological differences that maps to learning difficulty. The feature hierarchy was exposed by correlating predicted morphological differences with observed morphological differences per feature. Correlating predicted and observed values per feature quantifies explained variance per unique feature, which, at the moment, is uncommon in (generalized) linear mixed effects regression modeling. We took three steps to estimate the logistic mixed effects regression model with a random effect structure of features crossed with languages.

The first step. We calculated a vector of aggregated proficiency measures for the 72 mother tongues with at least 20 test scores available, corrected for country characteristics (educational quality) and individual differences (level of education, length of residence, age of arrival, and gender). The interaction between quality and level of education was also taken into account, resulting in a total of six fixed effects. Calculation of these measures was based on aggregated random effects from a multilevel model based on Schepens, Van der Slik, & Van Hout (in press). Besides the six fixed effects, random effects were incorporated for the country of birth, the mother tongue, the best additional language, and the combinations of the best additional language with the mother tongue. This random effect structure was discussed in Schepens, Van der Slik, and Van Hout, "The L2 Impact on Acquiring Dutch as a L3: The L2 Distance Effect."

The second step. We extracted language vectors from WALS for the 28 features that were also used by Lupyan & Dale (2010) to quantify morphological complexity. Thanks to the availability of WALS, large-scale typological comparison becomes possible. For the six missing values for the Dutch language vector, we adopted feature values from the language neighbour German.

The third step. The logistic mixed effects regression model was fitted by the Laplace approximation with a random effect for feature crossed with nested random effects for language per genus per family. For a comparison of different methods of fitting a logistic mixed effects model see Zhang et al. (2011). Adding learning difficulty as a fixed predictor (covariate) resulted in a significant decrease of 12.329 points of the -2 log likelihood approximation ($p<.001$). Allowing learning difficulty to vary randomly across features (addition of a random slope) resulted in a further decrease of 19.162 points ($p<.0001$). Subsequently, a new model was fitted where the fixed effect for learning difficulty was removed entirely in order to model this effect using random slopes only. For the coefficients of the final fitted model see Table 1 and Table 2.

Table 1. Parameter inferences are displayed for the standard deviations of the four random intercepts, the standard deviation of the random slope, and a correlation coefficient. The final fitted model included a random intercept for feature, a random slope for learning difficulty per feature, and one covariance parameter. The final fitted model also included random intercepts for 1) language nested within genus per family, 2) genus nested within family, and 3) family. The log-likelihood was estimated with the Laplace approximation.

Groups	Name	Std. Dev.	Corr
Feature	(Intercept)	1.409087	
	Learning Difficulty	0.049532	0.644
Language : (Genus : Family)	(Intercept)	0.000000	
Genus : Family	(Intercept)	0.521079	
Family	(Intercept)	0.377069	

Table 2. Estimation of the only fixed effect included in the final fitted model: the fixed intercept.

Fixed effects:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3995	0.2758	-1.448	0.148

Figure 1 shows the predicted probabilities as based on predicted log-odds (the log of the odds of a morphologically different feature value) for each feature (blue lines) as a function of learning difficulty. A learning difficulty of +30 indicates that speakers of that language on average score 30 points higher than the language sample's average, corrected for the confounding variables as described in step 1. The predicted log-odds were calculated by adding the estimated coefficients of the random intercepts to a multiplication of the estimated random slope coefficients with learning difficulty. Due to the transformation of odds into log-odds space, a linear model could be fitted and predicted log-odds could be replicated. Figure 1 shows a mixture of positive and negative relations between the probability of observing a morphological difference and learning difficulty per feature.

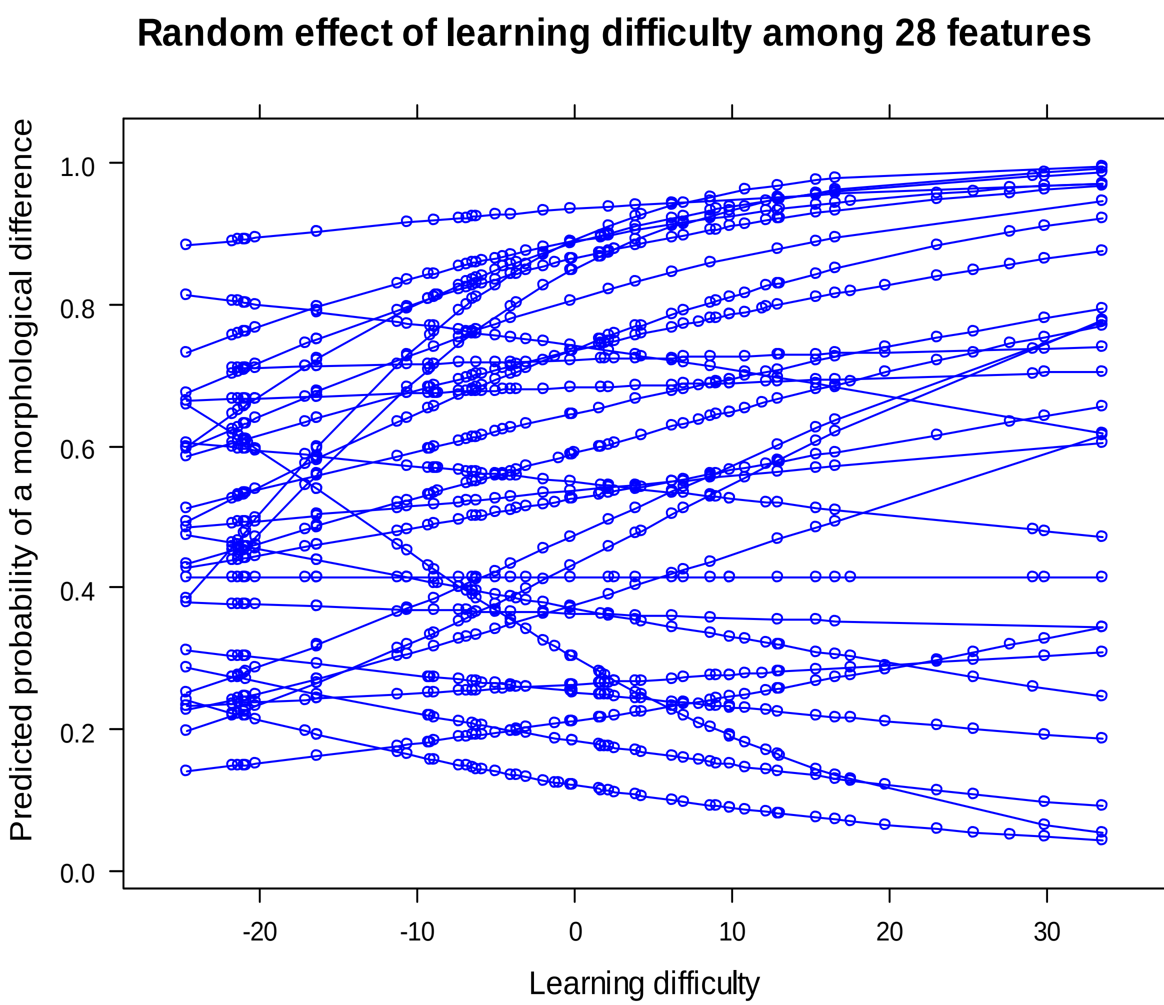

Figure 1. Predicted probabilities (data points) of a model that maps morphological differences (Y-axis) to learning difficulty (X-axis). Learning difficulty was measured by aggregating corrected language testing scores by mother tongue. Morphological difference was measured in terms of identical (1) or different (0) feature values. The slopes in blue represent the regressions on learning difficulty for each feature.

As announced, we correlated the fitted values with observed morphological difference in order to obtain a feature hierarchy. The highest correlations were observed for features *100* (r=.79, p<.05), *34* (r=.74, p<.05), *102* (r=.59, p<.05), and *28* (r=.55, p<.05). The lowest correlations were observed for features *70* (r=.14, ns), *59* (r=.01, ns), *38* (r=.00, ns). The highest negative correlation was found for feature number *101* (r=-.45, p<.05). The set of correlations implies a hierarchy of linguistic distance to Dutch with on top accusative alignment of verbal person marking (a feature of verb morphology considered medium complex in Dutch). Inspection of the data reveals that speakers of languages where there exists no alignment of verbal person marking experience more difficulty in learning the more morphologically complex alignment pattern of Dutch. On second place is the coding or occurrence of plurality, which is always obligatory for all nouns in Dutch. Although this is a feature of plurality considered morphologically simple in Dutch, speakers of languages in which plurality coding is not obligatory experience generally more difficulty in learning Dutch. On the third place is person marking on verbs, which Dutch employs for the agentive argument only. This seems especially hard for learners who speak a language that either does not apply verbal person marking at all (simpler) or applies it for both the agentive and the patient at the same time. On fourth place is case syncretism, which exists in Dutch (adopted from German) for core and non-core cases. Speakers of languages that have no inflectional case marking (considered simpler) experience relatively more difficulty in learning Dutch.

In this paper, we showed how to expose a feature hierarchy of learning difficulty of Dutch. The variety across features resulted in both high and low correlations between feature values and learning difficulty (ranging from .79 to -.45). Interestingly, it seems that learners of Dutch experience most difficulty for the few features of Dutch that may be regarded as morphologically complex. Besides these top 4 features discussed above, mixed positive and negative correlations were found for the other features that Lupyan & Dale identified as relatively complex in Dutch: *28* (also in top 4), *112* (r=.36), *26* (r=.32), *70* (r=.14), versus: *22* (-.19) and *77* (-.24). For feature *77* (if grammatical distinctions of evidentiality exist or not), Japanese and Korean, like Dutch, code evidentiality grammatically, while closely related languages such as Spanish and English lack this grammatical property. Judging from the difficulty that native speakers of Korean and Japanese have in learning Dutch, one morphologically complex feature alone does not determine learning difficulty, although it may contribute to the measure of learning difficulty. Despite these two exceptions, we confirm a trend between increasing morphological complexity and increasing learning difficulty. In relation to Lupyan & Dale (2010), this finding is compatible with the expectation that L2 learning of morphologically complex structures is more difficult and that complexity therefore decreases as more people have to learn the language as an L2.

Dryer, M. S., & Haspelmath, M. (Eds.). (2011). *The World Atlas of Language Structures Online*. Max Planck Digital Library. Retrieved from http://wals.info/

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*(6965), 435–439. doi:10.1038/nature02029

Levinson, S. C., & Gray, R. D. (2012). Tools from evolutionary biology shed new light on the diversification of languages. *Trends in Cognitive Sciences*, *16*(3), 167–173. doi:10.1016/j.tics.2012.01.007

Lupyan, G., & Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. (D. O'Rourke, Ed.). *PLoS ONE*, *5*(1), e8559. doi:10.1371/journal.pone.0008559

Sampson, G., Gil, D., & Trudgill, P. (2009). *Language complexity as an evolving variable*. Oxford University Press.

Schepens, J, Frans Van der Slik, and Roeland Van Hout. "The L2 Impact on Acquiring Dutch as a L3: The L2 Distance Effect". Paper presented at the Leuven Statistics Days: Mixed models and modern multivariate methods in linguistics, Leuven, Belgium, June 10, 2012. [http://jobschepens.ruhosting.nl/JSchepensLeuvenStatisticsDays2012.pdf]

Schepens, J., Van der Slik, F., & Van Hout, R. (in press). The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. *Comparing Approaches to Measuring Linguistic Differences*. Manuscript available at [http://jobschepens.ruhosting.nl/SchepensVanDerSlikVanHoutChapter.pdf]

Van der Slik, F. W. P. (2010). Acquisition of Dutch as a Second Language. *Studies in Second Language Acquisition*, *32*(03), 401–432. doi:10.1017/S0272263110000021

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLoS ONE*, *6*(9), e23613. doi:10.1371/journal.pone.0023613

Zhang, H., Lu, N., Feng, C., Thurston, S. W., Xia, Y., Zhu, L., & Tu, X. M. (2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine*, *30*(20), 2562–2572. doi:10.1002/sim.4265