Languages with More Second Language Learners Tend to Lose Nominal Case

Christian Bentz[*]
University of Cambridge, Department
of Theoretical and Applied Linguistics, UK
cb696@cam.ac.uk

Bodo Winter
University of California, Merced, Department
of Cognitive and Information Sciences, USA
bodo@bodowinter.com

Abstract

In this paper, we provide quantitative evidence showing that languages spoken by many second language speakers tend to have relatively small nominal case systems or no nominal case at all. In our sample, all languages with more than 50% second language speakers had no nominal case. The negative association between the number of second language speakers and nominal case complexity generalizes to different language areas and families. As there are many studies attesting to the difficulty of acquiring morphological case in second language acquisition, this result supports the idea that languages adapt to the cognitive constraints of their speakers, as well as to the sociolinguistic niches of their speaking communities. We discuss our results with respect to sociolinguistic typology and the Linguistic Niche Hypothesis, as well as with respect to qualitative data from historical linguistics. All in all, multiple lines of evidence converge on the idea that morphosyntactic complexity is reduced by a high degree of language contact involving adult learners.

Keywords

case – sociolinguistic typology – second language acquisition – language evolution

[*] Author names are in alphabetical order. Christian Bentz was responsible for project inception, data collection and writing. Bodo Winter was responsible for statistical analyses and writing.

© KONINKLIJKE BRILL NV, LEIDEN, 2013 | DOI: 10.1163/9789004281523_004

1 Introduction

Languages are diverse in the way participant roles and grammatical functions are encoded. Some languages rely on fixed word orders and adpositional phrases, others on case marking, and some on both. Within the set of those languages that use case, there are large differences with respect to how many cases are used. Hungarian, for example, has about twenty nominal cases, encoding information about who is doing what to whom, what belongs to whom, and what the spatial relationships between objects are. Other languages employ only few cases (e.g., German with four) or none at all (e.g., Chinese). What are the factors that drive this diversity?

When it comes to the complexity of case systems, there are three logical possibilities of what can happen diachronically: case paradigms can increase in complexity, decrease in complexity, or not change at all. Kulikov (2009: 456) calls these three "evolutionary types" of languages. What determines these evolutionary types could in principle be due to language-internal factors, language-external factors, or a combination of both. In this paper, we want to emphasize the contribution of language-external factors, in particular language contact, by providing a quantitative test of ideas from sociolinguistic typology (Trudgill, 2011) and the Linguistic Niche Hypothesis (Lupyan and Dale, 2010; Dale and Lupyan, 2012). We propose that languages that are widely acquired non-natively tend to lose nominal case, or have no case at all. Furthermore, we argue that this supports frameworks that see languages as changing entities adapting to the constraints of their users (e.g., Christiansen and Chater, 2008; Beckner et al., 2009).

It is not trivial to assume that case systems erode in language contact, as contact gives rise to a variety of different grammatical changes (Thomason and Kaufman, 1991: 115; Thomason, 2001: 75; Trudgill, 2002: 66), "simplification" being only one of them. In fact, there is considerable discussion about which contact scenarios lead to simplification and which lead to "complexification" (see, e.g., Trudgill, 2011: 15). For example, Nichols (1992) discusses how, in areas characterized by long-term contact between many different families, grammatical markers are often borrowed without replacing existing ones, leading to a net increase in morphosyntactic complexity (cf. discussion in Trudgill, 2011: 29). With respect to case, Aikhenvald (2003: 3) discusses how the language Tariana has developed entirely novel core cases via intensive contact. These examples show how case complexity can increase in contact situations.

On the other hand, researchers focusing on pidgins, creoles and koinés (Trudgill, 2002; 2004; 2011; McWhorter, 2007) have argued repeatedly that morphosyntax will be prone to regularization and reduction in many contact situations. Here, the simplifying processes that are at play in the development of pidgins, creoles and koinés are seen as equally effective in large-scale contact situations, such as in the case of English, Chinese or Persian, but perhaps to a lesser degree. On this account, pidgins and creoles are an extreme example of a more general process that applies to many contact situations. This view predicts that simplification due to contact should be *quantitatively dominant* compared to complexification, without neglecting that the latter sometimes happens.

Complexification is assumed to occur under long-term contact involving a high degree of child bilingualism (Trudgill, 2011: 40–41). The principle underlying simplification, on the other hand, is assumed to be imperfect learning by adult second language learners. McWhorter (2007: 14) states that "languages widely acquired non-natively are shorn of much of their natural elaboration." The work of McWhorter (2007) and Trudgill (2002, 2004, 2011) is replete with examples of qualitative studies of single languages that underwent some degree of morphosyntactic simplification due to L2 learning. To this growing body of evidence, we add a quantitative study of contact-induced reduction in nominal case complexity.

2 Multiple Mechanisms of Case Loss

Of particular relevance to our endeavor is Lupyan and Dale's (2010) finding that, across 28 typological features from the World Atlas of Language Structures (WALS; Dryer and Haspelmath, 2011), morphosyntactic complexity was inversely correlated with population size and the number of neighboring languages. Languages spoken by larger populations tended to have less morphosyntactic complexity, and the same applies to languages that are surrounded by more languages. However, population size and the neighboring languages only reflect contact in a very indirect fashion. As discussed by Lupyan and Dale (2010), their study requires the additional linking hypothesis that bigger language communities tend to have more contact with surrounding communities (as proposed by Wray and Grace, 2007).

By using information about the proportion of second language learners in a community, we have a more direct measure of language contact. Moreover, focusing on case allows us to make explicit links to the literature on the adult L2 acquisition of case. With reference to this literature, we would like to propose three potential mechanisms that can be relevant for case loss in contact situations in which adult second language speakers are involved:

a) There is abundant evidence for the idea that morphology in general and case in particular is difficult to acquire by adults. For example, Parodi, Schwartz, and Clahsen (2004) demonstrate that L2 speakers of German have problems learning morphological inflections, irrespective of their L1 (Korean, Turkish, Spanish, and Italian). Specifically with respect to case marking, Gürel (2000) shows that English L2 learners of Turkish experience serious problems with case, and Haznedar (2006) discusses evidence suggesting that these problems might persist even when the learner is very advanced. With some case forms in this particular study, case omission is observed to be very high (up to 90%). Papadopoulou and colleagues (Papadopoulou et al., 2011) demonstrate that Greek L2 learners of Turkish encounter persistent problems with the correct usage of case markers, despite the fact that Greek employs case as well. In this particular production study, the percentage of omission and substitution errors is higher overall than the percentage of correct uses, except for the highest proficiency level (ibid.: 186). In a similar vein, Jordens, De Bot, and Trapman (1989) test the acquisition of the correct usage of accusatives in two groups of Dutch L2 learners of German and find that learners tend to use the nominative as a default case, thus exhibiting a reduction in morphological differentiations.

Where this difficulty arises is somewhat less clear. The fact that case requires rote memorization of complex and sometimes irregular paradigms might play a role. Moreover, these memorized forms then have to be rapidly retrieved in the correct sentence context. Crucially, regardless of the exact cognitive mechanism, the abovementioned studies suggest that case substitution and omission are recurrent problems across L2 learners of varied languages. And the L2 evidence so far suggests that case is difficult, regardless of whether the learner's L1 has case marking or not. Following the evidence from the studies on L2 case acquisition, it is fair to assume that growing numbers of adult L2 speakers in a population will cause more omission and substitution errors in the overall spoken and written corpora.

b) Once the L2 speakers' difficulty with case is noticed by native speakers, the latter might in turn exhibit simplification as well ("foreigner talk" or FDS = foreigner directed speech). In Little (2011), two groups of participants had to learn an artificial language, and they reduced the morphosyntactic complexity of this language more when speaking to "foreigners" in the experiment than when speaking to non-foreigners, demonstrating FDS under controlled experimental conditions. Little (2011) argues that FDS is an underappreciated factor in sociolinguistically triggered language change; however, one has to recognize that FDS is closely linked to (and depends on) prior L2 learning difficulties.

c) Another potential mechanism of contact-induced case loss is proposed by Barðdal and Kulikov (2009): loan words tend to combine with more productive case inflections, biasing the distribution of case markers against less productive ones, which are then prone to disappear.

Crucially, our results do not hinge on any specific mechanism (learning difficulty, FDS, loan words). Our study is, in fact, agnostic to the exact mechanism of case erosion, especially because the mechanisms are mutually compatible with each other and are expected to pull languages in the same direction—towards less case. Thus all three mechanisms predict that the more L2 speakers exist in a population, the more case should be eroded. This is because with more L2 speakers, there are more erroneous and omitted forms in the joint L1 + L2 corpus (first mechanism), there will be more simplified foreigner-directed speech (second mechanism), and more loan words that disfavor less productive cases (third mechanism).

While our discussion so far has emphasized three language-external mechanisms, this is by no means intended to discount the importance of language-internal factors. In particular, it has been suggested that phonetic and phonological changes can lead to the loss of case distinctions. This has been argued for, among others, Classical Latin (Barðdal and Kulikov, 2009: 472), Old English (Allen, 1997: 75), Scandinavian languages (Norde, 2001) and Arabic (Barðdal and Kulikov, 2009: 472). An association between case loss and L2 speakers does not preclude that language-internal factors might also be at play, and language-internal and external factors might interact with each other, for example when an ongoing process of case erosion is facilitated or accelerated by L2 learners.

Having said this, in the following, we will quantitatively test the hypothesis that the three potential language-external mechanisms associated with adult L2 learning have an impact on case marking. We will outline our methodology, samples and statistics in Section 3, and report the results of our analyses in Section 4. The discussion in Section 5 will address potential concerns with our approach. Moreover, the quantitative data will then be linked to existing qualitative studies. We conclude by pointing out how our results only make sense in the light of a framework that views language as an adaptive system shaped by the linguistic niche of the speaker population.

3 Methodology

3.1 *Sample*

Our sample contains languages for which we could obtain reliable estimates or counts of the number of L2 speakers in the linguistic community. We define "L2 speakers" as adult L2 speakers as opposed to early bilinguals, following Trudgill (2011) in assuming that the reduction of case complexity is driven by adult L2 learners and not by child bilinguals. We were able to collect L2 speaker information for 226 languages using the SIL Ethnologue (Lewis, 2009), the Rosetta project website (www.rosettaproject.org), and the UCLA Language Materials Project (www.lmp.ucla.edu). Generally, these sources follow our L2 definition, although in some cases the exact "degree" of bilingualism might vary (see, e.g., "bilingualism remarks" in Ethnologue).

In this superset of 226 languages, we looked for overlap with the chapter on "Number of Cases" (Iggesen, 2011) in the World Atlas of Language Structures (WALS; Dryer and Haspelmath, 2011). This resulted in a sample of 66 languages (see Appendix for a list with detailed information). The sample comprises 26 language families from 16 different areas. The area and family information was taken from Balthasar Bickel and Johanna Nichol's AUTOTYP database (www.spw.uzh.ch/autotyp/).

Iggesen (2011) adopts 9 categories ranging from "No morphological case marking" to "10 or more cases." We excluded the category "Exclusively borderline case marking," since it was not clear how to rank this with respect to the other categories, and ended up with an 8-step continuum. In the context of the following study, case is operationally defined as in Iggesen (2011) to only include productive morphological inflections of nouns. Note that this definition is relatively loose, since it includes, for example, the possessive clitic *'s* in English (which is thus counted as having two cases, genitive and non-genitive). It has been argued that such clitics are not genuine case-markers since they can be attached to entire noun-phrases rather than inflected nouns only (see, e.g., Blevins, 2006, and Hudson, 1995). However, a look at the British National Corpus reveals that possessive markers are used with either proper nouns or common nouns in more than 90% of their occurrences. Hence, from the perspective of a second language learner, such possessive clitics behave very much like any other noun inflection they encounter. Moreover, while this is a potential concern for English and Swedish, it is not for most of the other languages in the dataset (e.g., Greek, Icelandic, Finnish, German, etc.), which only have genuine case affixes.

3.2 *Statistics*

As mentioned above, we do not intend to claim that the proportion of L2 speakers is the only factor affecting the number of case forms. Therefore, we expect exceptions to our hypothesis, e.g., a language with fairly few L2 speakers and a small number of nominal cases, or a language with a lot of L2 speakers and a large number of nominal cases. Individual languages might be exceptions because of particular sociohistorical developments, or because of particular structural features. We thus do not look for absolute universals but statistical ones (Bickel, 2010).

To assess the L2/case complexity association statistically, we used R (R Development Core Team, 2012) and the packages *lme4* (Bates, Maechler, and Bolker, 2012) and *glmmADMB* (Skaug et al., 2012) to construct generalized linear mixed effects models (for a discussion of linear models and mixed models in typology, see Cysouw, 2010, and Jaeger et al., 2011).

We constructed two models for our data, reflecting two ways of looking at the Iggesen (2011) variable. In one model, we ask the question: do languages with many L2 speakers tend to be those languages that have no case at all? For this model, we thus consider the presence and absence of case as a binary variable, which requires a logistic regression model. This analysis models the probability of a categorical dependent variable (here, no case vs. case) as a function of a predictor variable (in this case, the proportion of L2 speakers).

In the second model, we ask the question: do languages with many L2 speakers have *fewer* cases? For this analysis, we need a Poisson regression model. This analysis models case as a discrete count variable (1 case, 2 cases, 3 cases, etc.) as a function of a predictor variable (the L2 proportion). One important assumption of the Poisson distribution is that the sample mean and the sample variance are identical. In our case, this assumption was not met. The dispersion parameter was larger than 1.35, significantly above 1 ($\chi^2(66) = 89.3$, p = 0.029), indicating slight overdispersion (the sample variance exceeds the mean). In situations like this, the negative binomial distribution can be used as an alternative to the Poisson distribution. The negative binomial also models discrete count data, but it relaxes the overdispersion assumption (for a discussion, see Ismail and Jemain, 2007). One additional complication with regression of count data is the problem of too many zeros. In our case, 47% of all languages had no nominal case at all. We thus decided to use the function *glmmADMB*, which is able to account for zero inflation.

The predictor variable throughout both models was the proportion of L2 speakers in the overall L1 + L2 population. To control for areality and genealogy, we treated "Language Area" and "Language Stock" as crossed random effects (cf. Jaeger et al., 2011). We also included area-specific and family-specific random slopes for the effect of L2 proportion on case complexity.[1] A random slope model rather than an intercept-only model is necessary to account for the possibility that the effect of L2 proportion differs between language families and areas. This is to be expected, given that some areas such as South East Asia tend to have no case (see, e.g., Bickel and Nichols, 2009), therefore precluding the L2 proportion from affecting case. Having random slopes is also important because, as Barr et al. (2013) showed, models without random slopes for critical effects (in our case, the L2 speaker proportion) tend to be anticonservative (see also Schielzeth and Forstmeier, 2009). Our model also included a term to account for the correlation between random slopes and intercepts, e.g., languages with a high intercept (= many cases) might have a steeper slope (= more case loss).

Equations (1) and (2) show the general structure of the mixed logistic regression and the mixed negative binomial regression, respectively.

(1) $P(y_i = 1) = f^{-1}(\alpha_{j,\mathrm{k}[i]} + \beta_{j,k[i]}\, x_i)$
(2) $y_i = e^{\alpha_{j,k[i]} + \beta_{j,k[i]} \mathrm{x}_\mathrm{i}}$

In Equation (1), $P(y_i = 1)$ is the predicted probability of observing case (= 1) for each data point *i*. In both equations, the term $\alpha_{j,k[i]}$ represents the intercept for each i^{th} data point and $\beta_{j,k[i]}$ represents the slope for the effect of L2 speakers on case probability for each i^{th} data point. If this slope is negative, the probability of observing case decreases with higher values of L2 speakers; if it is positive, it increases. The subindices *j* and *k* represent adjustments of the intercept and slope for each language family and area respectively. Intercept and slope combined characterize the linear predictor. In Equation (1), estimated probabilities for observing case (as opposed to not observing case) can be derived by transforming this linear predictor by the inverse logit function f^{-1}. In Equation (2), the estimated count of nominal cases can be derived by transforming the linear predictor by the exponential function. In the context of the current discussion, it is crucial that the intercept and the slope are allowed to vary, that is, different language families and areas can have different "baseline" case occurrence levels, and within different families and areas, the L2 speaker proportion can have different effects on observed nominal case outcomes.

[1] For the negative binomial analyses, the model did not converge if a random slope term was introduced for the effect of L2 proportion depending on language areas. We thus proceeded with a model that only included random slopes for family and random intercepts for area. The model still accounts for areal tendencies to have more or less case because of the intercept component, but it does not account for differential effects of the L2 proportion predictor in different families. Because of these issues, we additionally ran a regular mixed Poisson regression with all necessary slopes (this model did converge). The results are the same.

When using mixed models, random effects should generally have 5 to 6 levels at a minimum. In our case, most languages or areas have less than that. This creates uncertainty in the estimation of some random intercepts and particularly random slopes. To show that the results reported below are not due to this problem, we additionally perform ordinary least squares regression (no nested random effect structure) on averages by family and by area (in analogy to a by-subjects or by-items analysis in psycholinguistics). For the mixed model analyses, we derive p-values using likelihood ratio tests (cf. Barr et al., 2013).

4 Results

We will first look at the distributions of the dependent and independent variables separately (see Fig. 1). In our sample, the speaker communities have on average 33 % second language speakers, with considerable spread around this value (SD = 27 %). The variable "case rank" shows a high proportion of 0's, indicating that many languages in our sample have no case at all (about 47 %). The distribution of the case rank variable looks somewhat bimodal (cf. Fig. 1b).

On average, the populations of languages with nominal case have about 16 % second language speakers, and the populations of languages without case have about 44 %. Mixed logistic regression indicates that languages with more second language speakers were more likely to have no case at all (logit estimates: -6.57 ± 2.03; p = 0.00014). Figure 2a displays the absence and presence of case as a function of the L2 speaker proportion. Each data point indicates a specific language and the curve indicates the fit of the logistic model. The height of this line indicates the probability of observing a language with case. As can be seen from looking at the plot, the curve drops to 0 around 50 %. In our sample, there were no communities of languages *with* case that had L2 speaker proportions up to about 50 %.

Looking at the number of nominal cases, Fig. 2b highlights the fact that languages with more L2 speakers tend to have fewer cases. A generalized linear mixed model with negative binomial error structure and a term for excess zeros (zero-inflation) indicates that this pattern is significant (log estimates: -3.6 ± 1.06; p = 0.00062).

Because of the random effects structure described above, these results generalize over language families and areas. A graphical way of depicting that these results are relatively independent from considerations of family and area is shown in Fig. 3, where the average case presence proportions and average case rank values are graphed for language families and areas, rather than individual languages. Here, the same pattern can be observed for all perspectives of looking at the data: languages and language areas that have many L2 speakers tend to have lower case proportion and case rank averages. We can also perform ordinary least squares regression (a general model) analysis on this averaged data. This analysis shows a significantly negative slope for all four plots in Fig. 3.

Despite the fact that mixed models with random effects for family and area account for areal and genealogical sources of non-independence, it is a potential concern that some families and areas are overrepresented. In particular,

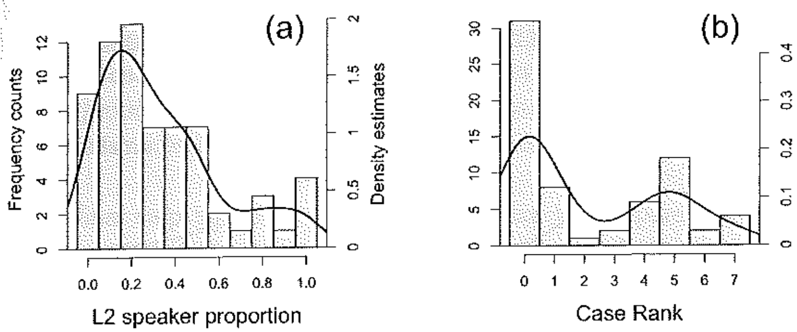

FIGURE 1 *(a) Frequency histogram and superimposed kernel density estimates of the L2 speaker proportion (independent variable). (b) Frequency histogram and superimposed kernel density estimates of the case rank variable (dependent variable).*

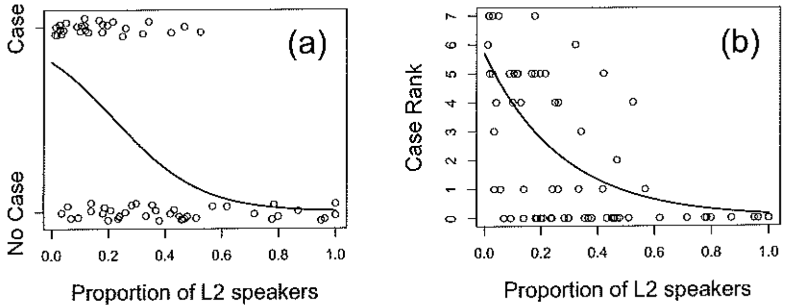

FIGURE 2 *(a) Presence and absence of case as a function of L2 speaker proportion. For better visibility, presence and absence points are shown with some random jitter along the y-axis. The curve indicates the fit of the logistic model, which represents the estimated probability of observing a language with case. (b) Case rank as a function of L2 speaker proportion. The curve indicates the fit of the negative binomial model, which represents the estimated number of nominal cases.*

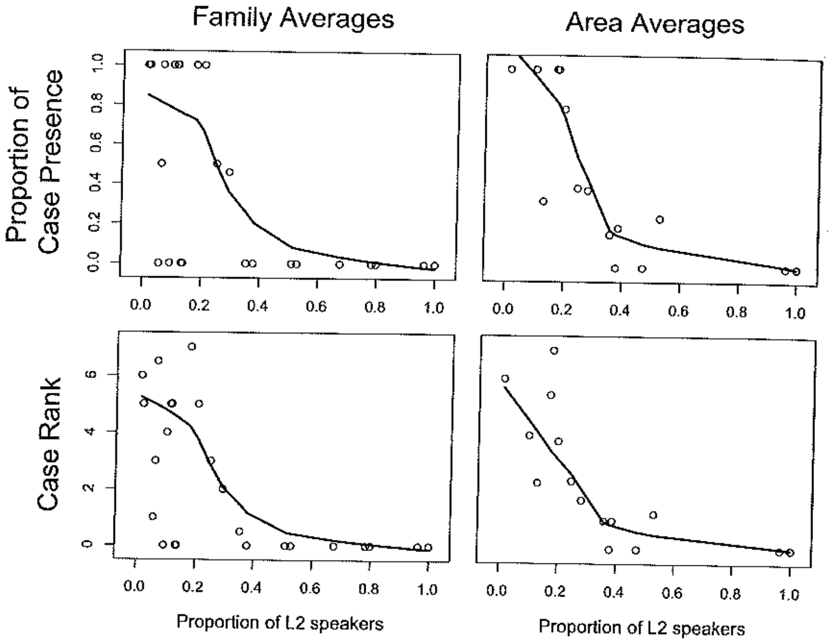

FIGURE 3 *Proportion of case presence (first row) and case rank (second row) as a function of* L2 *speaker proportion. Columns indicate averages by stock (left column) and by region (right column). Curves indicate lowess scatterplot smoothers.*

there are 24 Indo-European languages in our sample, constituting 36% of the total sample. Do the conclusions still hold if we exclude these languages? The answer is yes: the results for both the categorical measure presence vs. absence of case (logit estimates: -10.82 ± 4.55; p = 0.002) and case rank (-5.23 ± 1.45, p = 0.0003) hold if Indo-European languages are excluded.

With a small sample such as the current one, particular languages might potentially have large effects on the result. However, this can be informative, as it might direct us to investigate the reasons why a given language might be different from the general pattern. To assess leverage statistically, we performed influence diagnostics. This is done by successively excluding each language, rerunning the same analysis as reported above,[2] and observing how much the model estimates change if the dataset does not include a particular language (this is called DFBeta). It turns out that the estimates do not change much at all: all estimated coefficients have the same sign as the ones from the original model, indicating that, when particular languages were excluded, the direction of the L2 proportion effect on case did not change. More generally, this suggests that there is no particular language that influenced our results disproportionately.

As a final step in our analysis, we would like to compare our approach (looking at L2 speaker figures directly) to the one taken by Lupyan and Dale (2010), where population size was considered a shorthand for the degree of language contact. How well does population size (L1 + L2) predict the case complexity in our sample? For both categorical and count analyses, there was no effect of population size (presence vs. absence logit estimate: 0.73 ± 0.61, p = 0.23; case rank log estimate: -0.012 ± 0.03, p= 0.69), suggesting that L2 speaker information is crucial when looking at case complexity. In a small sample such as ours, L2 speaker proportion is statistically associated with case complexity, but it is impossible to detect the (presumably weaker) effect of population size, which only reflects language contact indirectly.

5 Discussion

To sum up, we demonstrated a statistical association between the proportion of L2 speakers and the presence and absence of case. This result is independent of biases coming from individual language families and language areas. Moreover, once the proportion of L2 speakers passes the threshold of 50%, case seems to be disfavored.

The apparent nonlinear pattern evidenced by the lowess lines in Fig. 3 is intriguing and reminiscent of nonlinear patterns in language evolution (e.g., Blythe and Croft, 2012) or nonlinear phase transitions in other domains of cognition (e.g., Spivey, Anderson, and Dale, 2009). In conjunction with the observation that for both count and categorical analyses (Fig. 2), there were no nominal cases with L2 speaker proportions above 50%, this might suggest some degree of discontinuity or bifurcation in the relationship between L2 speaker proportion and nominal case. It could be that case becomes vastly disfavored once the proportion of L2 speakers reaches a certain threshold. However, a kink in the graph can also be generated by a continuous nonlinear function such as the exponential function (Lamberson and Page, 2012). To be able to assess whether this nonlinearity is in fact a tipping point, one would need to have historical data to observe the relation between case and L2 speakers in time.

[2] In fact, we had to use a regular Poisson model for the rank data because the negative binomial model with zero-inflation did not converge often enough. The failure of models to converge is a common problem with small datasets.

5.1 *Addressing Potential Concerns*

There are several potential concerns that are inherent in our approach (and similar approaches such as Lupyan and Dale, 2010). In this section, we address the issue of ancestral case, the issue of time depth, a potential alternative hypothesis for our finding ("the reverse hypothesis"), the indetermination of mechanism, and growing case paradigms.

5.1.1 Ancestral Case

In our data collection procedure, we did not distinguish between languages whose ancestral languages had case and languages whose ancestors did not have case. However, if, for example, a protolanguage of a family is reconstructed without case, then any association between the absence of case and the proportion of L2 speakers might be entirely accidental. And, more generally, a given case system of, say, 5 nominal cases means something different depending on whether the ancestral language had 3 cases, 9 cases, or 0 cases.

We do control for this possibility via the inclusion of random slopes for the effect of L2 speaker proportion with respect to language family. Our model weighs the evidence for a relationship between L2 speakers and case with respect to the possibility that a language might not be able to exhibit such a relationship because it comes from a language family that has never had case. In our model, a language family is allowed to have a specific baseline value (e.g., 0 case for the Sino-Tibetan family) and a specific slope for the effect of L2 proportion (e.g., no relationship between L2 and case). Since our model also includes a term for the correlation between slopes and intercepts, languages from caseless language families are assumed to have shallower slopes. In other words, the model accounts for the fact that the absence of case within a family precludes the family from showing any effect of L2 proportion. However, the random effects structure does not include any complex phylogenetic structure beyond this. In particular, it would be desirable to have historic L2 data and link this with specific complexity-related morphosyntactic changes within a phylogenetic tree to make a closer causal connection between morphosyntactic complexity and L2 speaker proportion.

5.1.2 Time Depth

Can synchronic data from L2 speakers be used to make inferences about the past? Given that case loss is expected to be a process acting on larger historical time scales, what should actually be of importance is the number of second language speakers *in the past*. Sometimes, the number of adult learners can vary abruptly in the history of languages, e.g., when populations migrate, when new trade routes become accessible, or when new trade relations are made. While this certainly happens, the question is how frequently such abrupt changes occur. We know of no quantitative data that could be used to assess this frequency. However, we believe that across the board, L2 data from the present reflects the degree of language contact in the past, barring some noise. The noisy nature of the inferences based on these L2 figures is precisely why it is crucial to use a statistical approach that reflects broad-ranging trends. While short-term fluctuations might make *specific* data points less reliable, our approach is able to generalize across particularities. In any case, this is the best we can do given the lack of historical data on L2 figures. We only need to assume that the L2 figures of today reflect past language contact to some degree.

5.1.3 The Reverse Hypothesis

The reverse hypothesis states that, for the statistical association we found, causality runs the other way around: languages which are easier to learn attract more L2 learners. The ease or difficulty of learning a language may guide students' learning preferences in our present-day schooling system, where students have to learn a specific language and can choose to study the easiest one available to them. However, we think that the reverse hypothesis is unlikely to be a valid explanation for L2 learning in the past. Presumably, people did not have much of a choice when it came to learning languages, as socioeconomic factors must have been very important: if there was a population of speakers with whom someone wanted to (or needed to) trade, the language of those speakers had to be learned. It seems unlikely that learning preferences determine socioeconomic choices in these circumstances.

5.1.4 The Indetermination of Mechanism

In our introduction, we pointed towards several mechanisms that might have played a role with regards to case loss: first, the presence of a high proportion of incorrect and omitted forms in the joint L1 + L2 corpus due to imperfect learning, second, the accommodation of native speakers' speech to second language learners (= FDS), and third, loan words. Crucially, our data do not allow us to conclude with certainty that any of these mechanisms actually are the cause of the observed statistical relationship. According to the statistical mantra "correlation is not causation," there might be hidden variables which are somehow connected both to case complexity and L2 figures. We currently cannot think of such a lurking variable, but this cannot conclusively be ruled out.

Consideration of mechanisms is important, however. Here it is crucial to point out that our study is not characterized by post-hoc reasoning about a correlation we happened to find by chance, but that we *predicted* the association between case and L2 speakers based on prior empirical data (from second language acquisition studies, FDS) and a specific framework of linguistic theorizing related to sociolinguistic typology: the hypothesis of language as an adaptive system (Beckner et al., 2009; cf. Bentz and Christiansen, 2010; Christiansen and Chater, 2008) and the Linguistic Niche Hypothesis (Lupyan and Dale, 2010). On these grounds, we believe that our data support the involvement of at least some of the discussed mechanisms, and indirectly support the view of language as an adaptive system. Moreover, so far, no other contact-related mechanisms seem to be readily available to explain the patterns we found.

Linguists have often argued that case loss is due to language-internal mechanisms, such as the phonological erosion of case markers. While phonological erosion of case markers certainly does happen, sound change often does not explain the full pattern of case loss (see, e.g., Weerman and de Wit, 1999). Additionally, the work of Blevins and Wedel (2009) on inhibited sound change suggests that there should be pressures against the sound change happening with respect to case markers because of functional pressure to maintain the marking of important grammatical roles.

However, even though we prefer contact-induced case erosion as one of the quantitatively dominant mechanisms (in line with our analysis above), it should be pointed out that our results do not inherently stand in opposition against phonological or any other language-internal accounts of case loss. In fact, in our results, not all languages fall exactly onto the curves in Fig. 2 and Fig. 3, indicating there is a lot of variance that is left unexplained, some of which could be due to language-internal factors. Moreover, as pointed out above, language-internal factors and language-external factors can interact with each other (cf., e.g., discussion in Norde, 2001).

A final mechanism that needs to be discussed is "selective copying," where the grammatical idea behind a morphological form or a word order construction is copied into the L1 from a surrounding L2 (see, e.g., Johanson, 2009: 495). This seems to explain specifically why a lot of varieties of immigrant languages in English-speaking countries tend to lose case (e.g., Clyne, 2003: 124–130), but it does not necessarily predict any association between the number of L2 speakers in a language and the degree of case erosion. While selective copying certainly happens, it is unlikely to explain the full patterns of our results, which include contacts between many different languages—sometimes between languages that both have case and nevertheless tend to lose it. Thus, to sum up, we believe that the current data fits neatly within a relatively broad set of theoretical frameworks (sociolinguistic typology, language as an adaptive system) and is neatly predicted *a priori* based on the experimental data discussed above. This makes it very likely that the pattern we found is, in fact, connected to L2 speakers.

5.1.5 Growing Case Paradigms

At first sight, our approach seems to suggest that languages quite generally rather lose case than enhance their nominal case paradigms. This raises the question of how case marking could come into existence in the first place, and how mechanisms of enhancing case complexity are associated with the proportion of native speakers and non-native speakers in a language community. According to Wray and Grace (2007), *esoteric* linguistic communities, i.e. close-knit, culturally coherent groups of L1 learners with few or no language contact, will be prone to develop more opaque morphological marking strategies than *exoteric* societies, i.e. culturally rather heterogenic groups of 'strangers' that are associated with language contact and a high proportion of adult L2 learning. With regards to the varying degrees of learnability of languages, Wray and Grace (2007: 557) conclude that "a language that is customarily learned and used by adult non-native speakers will come under pressure to become more learnable by the adult mind, as contrasted with the child mind."

Trudgill (2011: 185) refines this argument by naming the exact factors that are potential predictors of linguistic complexity: 1) small population size; 2) dense social networks; 3) large amounts of shared information; 4) high stability; and 5) low contact. Crucially, Trudgill (ibid.) notes that these factors "*permit* linguistic complexity development; but they do not compel it to occur." This suggests that in a globalized world, the conditions for enhancement of morphological complexity in general, and case marking in particular, may be more 'rigorous' and harder to meet than the ones for loss of inflectional marking. This could explain why we see so many languages losing case, as well as other morphosyntactic features (Lupyan and Dale, 2010).

This is not to neglect that there are interesting examples of growing case marking paradigms in recent history. For instance, the indigenous language Wappo of the Yukian language family was spoken until the 1990s in a small territory near San Francisco Bay. This language is reported to have had an 8-case system (see Li, Thompson, and Sawyer, 1977: 90). The subject marking *-i* inflection is analyzed as a generalized form of an ergative marker and a recent development in the language's history (ibid.: 100). Note that from 1910 onwards, there was only a small, strongly interrelated group of 73 native speakers of Wappo (Cook, 1976: 239). Other examples of developing nominal case markers involve the Estonian *-ga/-ka* comitative/instrumental marker, which was derived from the Balto-Finnic noun **kansa* 'people,' 'society,' 'comrade' (Heine and Kuteva, 2007: 66), the derivation of a Basque comitative case suffix *-ekin* from the noun *kide* 'companion' (ibid.), and the Hungarian inessive and elative markers *-ben/-ban* and *-ból/-ből*, which both derived from the locative noun *bél* 'interior' (ibid.). Interestingly, Estonian and Hungarian are in our sample and have very low L2 ratios of 0.05 and 0.015, respectively.

While such examples are suggestive, a quantitative approach would also be important to gather further evidence for the hypothesis that small, close-knit societies are more likely to develop case markers in their languages than societies with recurrent L2 influence.

5.2 *Convergence with Qualitative Studies*

Our statistical approach dovetails nicely with many individual accounts of the histories of specific languages or language families. For example, Herman (2000) argues that L2 speakers that have been "recruited" into the Latin speech community when the Roman Empire spread throughout Europe were one important factor contributing to case erosion (cf. Bentz and Christiansen, 2010; Clackson and Horrocks, 2007: 276). Swedish and Danish also underwent considerable case erosion, for which Norde (2001: 243) states that "internal factors alone are not a sufficient explanation for the disappearance of inflectional case." Interestingly, these "contact-varieties" of the Germanic branch can be shown to be significantly more impoverished in terms of case marking than the relatively isolated Icelandic and Faroese (Trudgill, 2011: 72), which tend to conserve morphological complexity much more—and which, due to their isolation, also tend to have much less contact than other Germanic languages.

English, too, has been suggested to have been subject to contact-induced case erosion: while Old English and Old High German displayed four to five distinct cases (Admoni, 1990: 30; Hutterer, 2002: 313), these were lost to disproportionate extents in English and less so in German (Dal, 1962: 4). Many believe that this case loss must be connected to the influence of speakers of Scandinavian populations (McWhorter, 2007: 91 pp.), to assimilation of Late British speakers into the Old English population (Trudgill, 2011: 55) and to the invasion of the French-speaking Normans (Baugh and Cable, 2006: 108; Milroy, 1984). However, this position is still relatively controversial, with considerable counterarguments and a long-lasting debate surrounding this topic (see, e.g., Görlach, 1986; Allen, 1997; Dalton-Puffer, 1995; Thomason and Kaufman, 1991: 265).

Thomason and Kaufman argue against the assumption of large-scale simplification of English through French second language speakers, drawing on the following factors: 1) the comparatively low numbers of adult learners in the relevant areas (a maximum of 50,000 compared to 1.5–2 million English native speakers); 2) the fact that the degrees of simplification in the relevant dialects do not correlate with the degrees of borrowing of lexical material from French; 3) the fact that at least some of the changes resulting in simpler morphology in Middle English occurred earlier than the Norman Conquest, that is, before 1066. Similar objections can be raised with regards to a potential impact of Norse speakers on Northern English dialects of the Old English and Middle English period.

However, note that Thomason and Kaufman (1991) refuse general claims of creolization and morphological simplification in Middle English, not claims about nominal case marking in particular. In fact, even in Thomason and Kaufman's analyses there is some evidence that case markers were particularly prone to disappear between approximately 1200 and 1350. For example, they state that the Southern dialects of Middle English—which were in contact with the Normans—were still rather conservative with regards to verbal inflection. However, even in these varieties the following morphological simplifications can be observed (see Thomason and Kaufman, 1991: 310–311): 1) dative affixes on nouns are lost; 2) genitive plural affixes on nouns are lost; 3) gender and case agreement markers are reduced on pronominal modifiers; 4) subclasses of Old English nouns with less than 10 members are eliminated. Also, as pointed out earlier, we do not want to claim that language-internal factors, i.e. 'normal changes' (Thomason and Kaufman, 1991: 264) are irrelevant for case loss. Especially in complex contact scenarios like the Middle English one, it seems reasonable to consider both internal and external factors, rather than defining them as mutually exclusive.

With regards to pidgin and creole languages, the picture is somewhat less controversial. Pidgins are associated with incomplete adult language learning and interrupted transmission. Trudgill (2011: 182–183) asserts that this is exactly the reason why pidgin languages quite generally lack morphological marking strategies. Cases, numbers, tenses, moods, voices, aspects, persons and genders are encoded in periphrastic constructions, if at all. Although natively learned creole languages often employ "repair" mechanisms to overcome this inflectional scarcity, these mechanisms are mostly limited to optional aspect and tense markers as well as optional plural markers. Moreover, in most creole languages case relations are marked by word order rather than affixes. This pattern has led McWhorter (2011) to argue that one of the most salient features of a *Creole Prototype* is the extreme rarity or, in fact, non-existence of inflectional marking. Some interesting counterexamples to this general claim are given in Plag (2005). Sri Lanka Creole Portuguese, for instance, employs the same set of case markers as the substrate languages Tamil and Singhalese. In this creole variety case affixes are derived from lexical material of the lexifier Portuguese, e.g., the dative marker *-pa* as an eroded form of the preposition *para*. Another potential counterexample is the Arabic noun for 'property,' which was reduced to *ta* and became a genitive case marker in the Arabic-based creole Nubi (Heine and Kuteva, 2007: 66). Such morphological complexification in contact situations is unexpected from our overall point of view. However, McWhorter (2011) encounters Plag's (2005) criticism by pointing out that rare examples to the opposite do not refute the overall claim that adult learner varieties are among the morphologically simplest languages. This is exactly the point we are trying to make with our quantitative and statistical approach.

We see another connection between our study and existing work on language enclaves. Here, a common finding is that inflectional paradigms are maintained in the first generations after immigration, but in the following generations morphological systems are quickly simplified (see, e.g., Boas, 2009; Salmons, 1994; Franke, 2008; Trudgill, 2004). For example, in Texas German, use of the dative went down from 64% to 28.5% (Salmons, 1994: 61) within only one generation. This dramatic change happened when, after World War I, the German language suffered a heavy loss of reputation, and a considerable number of parents (Boas, 2009: 349) decided not to speak Texas German with their children. Thus, the children of this variety successively became L1 speakers of English and L2 learners of Texas German (Franke, 2008, shows a similar pattern for Springbok German in South Africa). This opens up the possibility that case loss is at least partly due to imperfect L2 learning.

Similar tendencies of case loss are reported for Haysville East Franconian in Indiana (Nützel, 1993), for younger speakers of Michigan German (Born, 2003), for Volga German spoken in Kansas (Keel, 1994), for German varieties in the Transcarpathia region in Ukraine (Keel, 1994), for Mennonite communities in the Altai region (Jedig, 1981), for a Low German variety spoken in Kyrgyzstan (Hooge, 1992), and for a German variety spoken in Hungary (Knipf-Komlósi, 2006) (for a review, see Franke, 2008). It appears that case loss in German language islands is a common pattern, regardless of whether the contact language has a rich case system (e.g., Hungarian) or hardly any case at all (e.g., English). Therefore, this case loss cannot be primarily due to selective copying from surrounding non-case languages. This suggests that, when younger speakers learn their own minority language as a second language, learning constraints come into play and may affect subsequent language change, a view that is very much compatible with the frameworks outlined above.

It should be pointed out that our results do not hinge on whether any of the particular historical cases discussed in the preceding section is actually due to L2 learning-induced simplification or not. Ultimately, our approach speaks for itself, but we see the historical cases as a nice convergence of qualitative studies with our present quantitative one as well as broader, more large-scale studies such as Lupyan and Dale (2010).

6 Conclusions

Second language acquisition studies suggest that nominal cases are particularly hard to learn for adult learners. This micro-scale learning difficulty in individual people might have macro-scale effects on the development of languages, as long as there are: a) enough second language learners to affect the whole system abruptly, or b) a permanent influx of new L2 speakers over several generations, or both. As would be expected based on adult case learning difficulties, we found an inverse association between the proportion of L2 learners and the presence of nominal case, as well as an inverse association between the proportion of L2 learners and the number of nominal case markers.

Taken together, the historical cases discussed by many linguists, the evidence from second language acquisition, and the evidence from our analyses reported above dovetail nicely with the idea that languages adapt to the sociocultural niches of their speaking communities and the cognitive constraints of their speakers. This is expected based on such proposals as the Language as Shaped by the Brain Hypothesis (Christiansen and Chater, 2008), the Linguistic Niche Hypothesis (Lupyan and Dale, 2010; Dale and Lupyan, 2012; cf. Wray and Grace, 2007), Trudgill's Sociolinguistic Typology (2011) and McWhorter's framework outlined in "Language Interrupted" (2007). To these proposals, we add another piece of quantitative evidence.

In addition, we would like to point out that our study makes an important methodological point. Typologists know it is crucial to control for the non-independences in a dataset that stem from language areas and language families (e.g., Dryer, 1989, 1992). The best remedy for an areally and genealogically biased typological analysis is to balance the sample with respect to families and areas. However, this was not possible in our case, as L2 speaker information is very limited. We thus had to resort to a non-balanced sample. Mixed models make it possible to work with such a sample even if the sample size is small, because they allow us to account for areal and genealogical effects in a single model (Jaeger et al., 2011). We also show that influence diagnostics are of importance in typological analyses: in our case, we were able to show that excluding individual languages does not greatly affect the results and thus, the processes that underlie the discussed pattern seem to be uniform to the point that no particular languages play a dominant role. Given that some languages have quite extreme linguistic histories, this is a fairly unexpected result. It suggests that the language contact processes leading to case erosion are present in many different languages, despite idiosyncratic historical trajectories.

Finally, our approach makes reference to independent evidence from experiments and second language studies. This shows that linguists and typologists can gain a lot from looking outwards to other fields to find converging evidence for existing hypotheses and, ultimately, more confidence in these ideas. Future research will show whether a similar integration can be made across other linguistic domains, such as phonology, syntax, and other aspects of morphology. For all of these different domains, we expect the same principle to hold: those aspects of grammar that adults find most difficult to learn will be most disfavored in language change when a language is spoken by many second language speakers. Languages are thus seen as having to fit the specific niche of their speakers and communities—if the fit is less than optimal, specific linguistic features will not be passed on to future generations. Thus, languages themselves are seen as adapting entities (Beckner et al., 2009; cf. Bentz and Christiansen, 2010), constantly changing as a function of a multitude of different environmental factors, including the cognitive systems that need to be able to sustain them.

Acknowledgments

We thank Roger Mundry for statistical advice and Søren Wichmann as well as three anonymous reviewers for helpful comments and suggestions. We thank the Max Planck Institute for Evolutionary Anthropology in Leipzig for supporting this project.

References

Admoni, Vladimir G. 1990. *Historische Syntax des Deutschen*. Tübingen: Niemeyer.

Aikhenvald, Alexandra. 2003. *A Grammar of Tariana, from Northwest Amazonia*. Cambridge: Cambridge University Press.

Allen, Cynthia. 1997. Middle English case loss and the 'creolization' hypothesis. *English Language and Linguistics* 1: 63–89.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68: 255–278.

Barðdal, Jóhanna and Leonid Kulikov. 2009. Case in decline. In Andrej Malchukov and Andrew Spencer (eds.), *The Oxford Handbook of Case*, 471–478. Oxford: Oxford University Press.

Bates, Douglas, Martin Maechler, and Ben Bolker. 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999–0.

Baugh, Albert C. and Thomas Cable. 2006. *A History of the English Language*. London: Routledge.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. 2009. Language is a complex adaptive system. *Language Learning* 59: 1–26.

Bentz, Christian and Morten H. Christiansen. 2010. Linguistic adaptation at work? The change of word order and case system from Latin to the Romance languages. In Andrew D.M. Smith, Marieke Schouwstra, Bart de Boer, and Kenny Smith (eds.), *Proceedings of the 8th International Conference on the Evolution of Language*, 26–33. Singapore: World Scientific.

Bickel, Balthasar. 2010. Absolute and statistical universals. In Patrick C. Hogan (ed.), *The Cambridge Encyclopedia of the Language Sciences*, 77–79. Cambridge: Cambridge University Press.

Bickel, Balthasar and Johanna Nichols. 2009. The geography of case. In Andrej Malchukov and Andrew Spencer (eds.), *The Oxford Handbook of Case*, 479–493. Oxford: Oxford University Press.

Blevins, James P. 2006. English inflection and derivation. In Bas Aarts and April M. McMahon (eds.), *The Handbook of English Linguistics*, 507–536. Oxford: Blackwell.

Blevins, Juliette and Andrew Wedel. 2009. Inhibited sound change. An evolutionary approach to lexical competition. *Diachronica* 26: 143–183.

Blythe, Richard and William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88: 269–304.

Boas, Hans C. 2009. Case loss in Texas German: The influence of semantic and pragmatic factors. In Jóhanna Barðdal and Shobhana L. Chelliah (eds.), *The Role of Semantic, Pragmatic, and Discourse Factors in the Development of Case*. Amsterdam: John Benjamins.

Born, Renate. 2003. Regression, convergence, internal development: The loss of the dative case in German-American dialects. In William D. Keel and Klaus J. Mattheier (eds.), *German Language Varieties Worldwide: Internal and External Perspectives*, 151–164. Frankfurt: Lang.

Christiansen, Morten H. and Nick Chater. 2008. Language as shaped by the brain. *Behavioral and Brain Sciences* 31: 489–509.

Clyne, Michael. 2003. *Dynamics of Language Contact*. Cambridge: Cambridge University Press.

Clackson, James and Geoffrey Horrocks. 2007. *The Blackwell History of the Latin Language*. Oxford: Blackwell.

Cook, Sherburne F. 1976. *The Conflict between the California Indian and White Civilization*. Berkeley: University of California Press.

Cysouw, Michael. 2010. Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14: 253–287.

Dal, Ingerid. 1962. *Kurze deutsche Syntax: Auf historischer Grundlage*. Tübingen: Niemeyer.

Dale, Rick and Gary Lupyan. 2012. Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems* 15: 1150017 (doi: 10.1142/S0219525911500172).

Dalton-Puffer, Christiane. 1995. Middle English is a creole and its opposite: On the value of plausible speculation. In Jacek Fisiak (ed.), *Linguistic Change under Contact Conditions*, 35–50. Berlin: Mouton de Gruyter.

Dryer, Matthew S. 1989. Large linguistic areas and linguistic sampling. *Studies in Language* 13: 257–292.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68: 81–138.

Dryer, Matthew S. and Martin Haspelmath (eds.). 2011. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. Online: http://wals.info/.

Franke, Katharina. 2008. "We call it Springbok-German!" Language contact in the German communities in South Africa. Unpublished ms., Monash University.

Görlach, Manfred. 1986. Middle English—a creole? In Dieter Kastovsky and Aleksander Szwedek (eds.), *Linguistics across Historical and Geographical Boundaries*, 329–345. Berlin: Mouton de Gruyter.

Gürel, Ayse. 2000. Missing case inflection: Implications for second language acquisition. In Catherine Howell, Sarah A. Fish and Thea Keith-Lucas (eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development* 45, 379–390. Somerville, MA: Cascadilla Press.

Haznedar, Belma. 2006. Persistent problems with case morphology in L2 acquisition. In Conxita Lleó (ed.), *Interfaces in Multilingualism*, 179–206. Amsterdam: John Benjamins.

Heine, Bernd and Tania Kuteva. 2007. *The Genesis of Grammar*. Oxford: Oxford University Press.

Herman, József. 2000. *Vulgar Latin*. University Park, PA: The Pennsylvania State University Press.

Hooge, David. 1992. Deutsche Sprachinseln: Ein Forschungsobjekt der Mehrsprachigkeit. *Germanistische Mitteilungen* 35: 107–114.

Hudson, Richard. 1995. Does English really have case? *Journal of Linguistics* 31: 375–392.

Hutterer, Claus J. 2002. *Die germanischen Sprachen. Ihre Geschichte in Grundzügen*. Wiesbaden: VMA-Verlag (Albus).

Iggesen, Oliver A. 2011. Number of cases. In Matthew S. Dryer and Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, ch. 49. Munich: Max Planck Digital Library.

Ismail, Noriszura and Abdul A. Jemain. 2007. Handling overdispersion with negative binomial and generalized Poisson regression models. *Casualty Actuarial Society Forum*, Winter 2007: 103–158.

Jaeger, T. Florian, Peter Graff, William Croft, and Daniel Pontillo. 2011. Mixed effects models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15: 281–320.

Jedig, Hugo. 1981. Zur Entwicklung des Synkretismus im Kasussystem des Niederdeutschen in der Sowjetunion. In Wilhelm-Pieck-Universität Rostock, Wissenschaftsbereich Germanistik (eds.), *Das Niederdeutsche in Geschichte und Gegenwart* 75, 164–172. Berlin: Zentralinstitut für Sprachwissenschaft der Akademie der Wissenschaften der DDR.

Johanson, Lars. 2009. Case and contact linguistics. In Andrej Malchukov and Andrew Spencer (eds.), *The Oxford Handbook of Case*, 494–501. Oxford: Oxford University Press.

Jordens, Peter, Kees de Bot, and Henk Trapman. 1989. Linguistic aspects of regression in German case marking. *Studies in Second Language Acquisition* 11: 179–204.

Keel, William D. 1994. Reduction and loss of case marking in the noun phrase in German-American speech islands: Internal development or external interference? In Nina Berend and Klaus J. Mattheier (eds.), *Sprachinselforschung: Eine Gedenkschrift für Hugo Jedig*, 93–104. Frankfurt: Peter Lang.

Knipf-Komlósi, Elisabeth. 2006. Sprachliche Muster bei Sprachinselsprechern am Beispiel der Ungarndeutschen. In Nina Berend and Elisabeth Knipf-Komlósi (eds.), *Sprachinselwelten—The World of Language Islands*, 39–56. Frankfurt: Peter Lang.

Kulikov, Leonid. 2009. Evolution of case. In Andrej Malchukov and Andrew Spencer (eds.), *The Oxford Handbook of Case*, 439–457. Oxford: Oxford University Press.

Lamberson, Peter J. and Scott E. Page. 2012. Tipping points. *Quarterly Journal of Political Science* 7: 175–208.

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the World*. 16th ed. Dallas, Texas: SIL International. Online: http://www.ethnologue.com/.

Li, Charles N., Sandra A. Thompson, and Jesse O. Sawyer. 1977. Subject and word order in Wappo. *Journal of American Linguistics* 43: 85–100.

Little, Hannah. 2011. *The Role of Foreigner Directed Speech in the Cultural Transmission of Language and the Resulting Effects on Language Typology*. PhD dissertation, University of Edinburgh.

Lupyan, Gary and Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS ONE* 5: e8559–e8559.

McWhorter, John. 2007. *Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars*. New York: Oxford University Press.

McWhorter, John. 2011. *Linguistic Simplicity and Complexity: Why Do Languages Undress?* Boston: Mouton de Gruyter.

Milroy, James. 1984. The history of English in the British Isles. In Peter Trudgill (ed.), *Language in the British Isles*, 5–32. Cambridge: Cambridge University Press.

Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Norde, Muriel. 2001. The loss of lexical case in Swedish. In Jan T. Faarlund (ed.), *Grammatical Relations in Change*, 241–272. Amsterdam: John Benjamins.

Nützel, Daniel. 1993. Case loss and morphosyntactic change in Haysville East Franconian. In Joseph Salmons (ed.), *The German Language in America: 1683–1991*, 307–321. Madison: Max Kade Institute for German-American Studies, University of Wisconsin-Madison.

Papadopoulou, Despina, Spyridoula Varlokosta, Vassilios Spyropoulous, Hasan Kaili, Sophia Prokou, and Anthi Revithiadou. 2011. Case morphology and word order in second language Turkish: Evidence from Greek learners. *Second Language Research* 27: 173–205.

Parodi, Teresa, Bonnie D. Schwartz, and Harald Clahsen. 2004. On the L2 acquisition of the morphosyntax of German nominals. *Linguistics* 42: 669–705.

Plag, Ingo. 2005. Morphology in pidgins and creoles. In Keith Brown (ed.), *Encyclopedia of Language and Linguistics*, vol. 8, 304–308. Oxford: Elsevier.

R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Downloadable at http://www.r-project.org/.

Salmons, Joseph. 1994. Naturalness and morphological change in Texas German. In Nina Berend and Klaus J. Mattheier (eds.), *Sprachinselforschung: Eine Gedenkschrift für Hugo Jedig*, 59–72. Frankfurt: Peter Lang.

Schielzeth, Holger and Wolfang Forstmeier. 2009. Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology* 20: 416–420.

Skaug, Hans J., David A. Fournier, Anders Nielsen, Arni Magnusson, and Benjamin M. Bolker. 2012. Generalized linear mixed models using AD Model Builder. R package version 0.7.2.12.

Spivey, Michael J., Sarah Anderson, and Rick Dale. 2009. The phase transition in human cognition. *Journal of New Mathematics and Natural Computing* 5: 197–220.

Thomason, Sarah G. 2001. *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press.

Thomason, Sarah G. and Terrence Kaufman. 1991. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Trudgill, Peter. 2002. *Sociolinguistic Variation and Change*. Washington, DC: Georgetown University Press.

Trudgill, Peter. 2004. *New-dialect Formation*. Oxford: Oxford University Press.

Trudgill, Peter. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Weerman, Fred and Petra De Wit. 1999. The decline of the genitive in Dutch. *Linguistics* 37: 1155–1192.

Wray, Alison and George W. Grace. 2007. The consequences of talking to strangers. Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117: 543–578.

Appendix. Sample Languages in Alphabetical Order

Language	Case Category	Stock	Region	L1 est.	L2 est.	L2/sum
Abkhaz	2 cases	North Caucasian	Greater Mesopotamia	117350	7041	0.057
Amharic	2 cases	Semitic	Greater Abyssinia	17464250	5500000	0.240
Araona	5 cases	Pano-Tacanan	NE South America	81	9	0.100
Bambara	No case	Mande	African Savannah	2772340	10000000	0.783
Bench (Gimira)	6–7 cases	Omotic	Greater Abyssinia	174000	22640	0.115
Bulgarian	No case	Indo-European	Europe	7897189	2067262	0.207
Burmese	8–9 cases	Sino-Tibetan	Southeast Asia	32309850	15250000	0.321
Catalan	No case	Indo-European	Europe	9115080	5000000	0.354
Chamorro	No case	Austronesian	Oceania	76350	30000	0.282
Chinese	No case	Sino-Tibetan	Southeast Asia	875744667	192383000	0.180
Chuvash	6–7 cases	Turkic	Inner Asia	1720000	200000	0.104
Croatian	5 cases	Indo-European	Europe	5546590	1953410	0.260
Dutch	No case	Indo-European	Europe	21865145	5000000	0.186
English	2 cases	Indo-European	Europe	339004069	169000000	0.333
Estonian	10 or more cases	Uralic	Inner Asia	1274330	67000	0.050
Ewe	No case	Kwa	African Savannah	3112000	500000	0.138
Fijian	No case	Austronesian	Oceania	392500	320000	0.449
Finnish	10 or more cases	Uralic	Inner Asia	5504695	161243	0.028
French	No case	Indo-European	Europe	88612817	80700000	0.477

(cont.)

Language	Case Category	Stock	Region	L1 est.	L2 est.	L2/sum
Frisian (Western)	No case	Indo-European	Europe	410500	174000	0.298
Georgian	6–7 cases	Kartvelian	Greater Mesopotamia	4255270	80000	0.018
German	4 cases	Indo-European	Europe	97100000	50250000	0.341
Greek	3 cases	Indo-European	Europe	12542245	11000000	0.467
Hausa	No case	Chadic	African	24594000	15000000	0.379
Hebrew	No case	Semitic	Greater Mesopotamia	5316700	4683300	0.468
Hungarian	10 or more cases	Uralic	Europe	13500635	203966	0.015
Icelandic	4 cases	Indo-European	Europe	269025	9553	0.034
Ilokano	No case	Austronesian	Oceania	7348300	2300000	0.238
Indonesian	No case	Austronesian	Southeast Asia	21033000	140000000	0.869
Irish (Gaelic)	2 cases	Indo-European	Europe	1656790	581574	0.260
Italian	No case	Indo-European	Europe	65848339	50000000	0.432
Japanese	8–9 cases	Japanese	N Coast Asia	126026700	1500000	0.012
Kannada	6–7 cases	Dravidian	Indic	36863800	9000000	0.196
Kanuri	6–7 cases	Saharan	African Savannah	3820250	500000	0.116
Khmer (Central)	No case	Austroasiatic	Southeast Asia	13603400	1000000	0.068
Kongo	No case	Benue-Congo	S Africa	5955908	5000000	0.456
Kutenai	No case	Kutenai	Basin and Plains	12	310	0.963
Latvian	5 cases	Indo-European	Inner Asia	1504880	500000	0.249
Lithuanian	6–7 cases	Indo-European	Inner Asia	3327090	720000	0.178
Makah	No case	Wakashan	Alaska-Oregon	0	2224	1.000
Malayalam	6–7 cases	Dravidian	Indic	36531330	10000000	0.215
Maori	No case	Austronesian	Oceania	35130	121980	0.776
Marathi	5 cases	Indo-European	Indic	70012927	3000000	0.041
Nenets	6–7 cases	Uralic	Inner Asia	41302	8260	0.167
Oromo (All)	6–7 cases	Cushitic	S Africa	20600000	2000000	0.088
Persian	2 cases	Indo-European	Greater Mesopotamia	47430600	62000000	0.567
Pitjantjatjara	10 or more cases	Pama-Nyungan	S Australia	2310	500	0.178
Polish	6–7 cases	Indo-European	Europe	39990670	1145760	0.028
RapaNui	No case	Austronesian	Oceania	3320	1925	0.367
Romanian	2 cases	Indo-European	Europe	24808318	4000000	0.139
Russian	6–7 cases	Indo-European	Inner Asia	151776975	110000000	0.420
Sango	No case	Adamawa-Ubangi	African Savannah	404000	1600000	0.798
Serbian	5 cases	Indo-European	Europe	10010275	11000000	0.524
Sinhalese	5 cases	Indo-European	Indic	15584375	2300000	0.129
Somali	2 cases	Cushitic	Greater Abyssinia	13871700	500000	0.035
Spanish	No case	Indo-European	Europe	343259405	105000000	0.234
Swahili (Tanzania)	No case	Benue-Congo	S Africa	2893815	55000000	0.950
Tagalog	No case	Austronesian	Oceania	20426600	51000000	0.714
Thai	No case	Tai-Kadai	Southeast Asia	33231195	34500000	0.509
Tiwi	No case	Tiwi	N Australia	0	1830	1.000

(cont.)

Language	Case Category	Stock	Region	L1 est.	L2 est.	L2/sum
Turkish	6–7 cases	Turkic	Greater Mesopotamia	62375060	8390200	0.119
Urdu	2 cases	Indo-European	Indic	60586800	43413200	0.417
Vietnamese	No case	Austroasiatic	Southeast Asia	64317000	16000000	0.199
Welsh	No case	Indo-European	Europe	317000	271000	0.461
Yoruba	No case	Niger-Congo	African Savannah	19380800	2000000	0.094
Zulu	No case	Benue-Congo	s Africa	9790000	15850000	0.618

Using Phylogenetic Networks to Model Chinese Dialect History

Johann-Mattis List
Forschungszentrum Deutscher Sprachatlas,
Philipps University Marburg, Marburg, Germany
mattis.list@uni-marburg.de

Shijulal Nelson-Sathi
Institute of Molecular Evolution,
Heinrich Heine University Düsseldorf, Düsseldorf, Germany
shijulalns@uni-duesseldorf.de

William Martin
Institute of Molecular Evolution,
Heinrich Heine University Düsseldorf, Düsseldorf, Germany
w.martin@uni-duesseldorf.de

Hans Geisler
Institute of Romance Languages and Literature,
Heinrich Heine University Düsseldorf, Düsseldorf, Germany
geisler@uni-duesseldorf.de

Abstract

The idea that language history is best visualized by a branching tree has been controversially discussed in the linguistic world and many alternative theories have been proposed. The reluctance of many scholars to accept the tree as the natural metaphor for language history was due to conflicting signals in linguistic data: many resemblances would simply not point to a unique tree. Despite these observations, the majority of automatic approaches applied to language data has been based on the tree model, while network approaches have rarely been applied. Due to the specific sociolinguistic situation in China, where very divergent varieties have been developing under the roof of a common culture and writing system, the history of the Chinese dialects is complex and intertwined. They are therefore a good test case for methods which no longer take the family tree as their primary model. Here we use a network approach to study the lexical history of 40 Chinese dialects. In contrast to previous approaches, our method is character-based and captures both vertical and horizontal aspects of language history. According to our results, the majority of characters in our data (about 54 %) cannot be