



A dynamical-systems approach to the evolution of morphotactic and lexical consonant clusters in English and Polish

Andreas Baumann¹ and Kamil Kaźmierski²

¹ *University of Vienna, andreas.baumann@univie.ac.at*

² *Adam Mickiewicz University, Poznań, kamil.kazmierski@wa.amu.edu.pl*

Abstract

Consonant clusters appear either lexically within morphemes or morphotactically across morpheme boundaries. According to extant theories, their diachronic dynamics are suggested to be determined by analogical effects on the one hand as well as by their morphological signaling function on the other hand. This paper presents a mathematical model which allows for an investigation of the interaction of these two forces and the resulting diachronic dynamics. The model is tested against synchronic and diachronic language data. It is shown that the evolutionary dynamics of the cluster inventory crucially depend on how the signaling function of morphotactic clusters is compromised by the presence of lexical items containing their morpheme internal counterparts.

Keywords: morphotactics; mathematical modeling; evolutionary linguistics; morphological complexity; analogy.

1. Introduction

There is strong evidence that consonant clusters in general, and final consonant clusters in particular, are dispreferred. First, they are much rarer typologically than languages allowing at most one coda consonant. For example, in the World Phonotactics Database (Donohue et al. 2013), which contains data on 2,378 languages, as many as 79.3% of the languages allow no more than one coda consonant. Second, consonant clusters are reduced in casual speech (Dziubalska-Kończak and Zydorowicz 2014; Madelska 2005; Shockey 2003). Third, they are acquired late in first language acquisition (Jarosz et al., submitted). Despite this evidence of bias against them, however, consonant clusters are well-established in a number of languages, including Polish and English.

Language specific phonotactics often restrict the set of clusters that are allowed in morphologically simple words. For example, English does not allow its words to end in /-md/ and Polish does not allow its simplex words to end in /-gw/. However, such phonotactic restrictions are considerably relaxed when it comes to morphologically complex forms. And so /md/ can surface in English *seemed* /si:md/ and /gw/ in Polish *pomógł* /'pɔmugw/ 'he helped' when a past tense suffix is attached. The difference in the range of sound sequences allowed to surface lexically (i.e. phonotactically) on the one hand, and those allowed to surface in morphologically complex words (i.e. "morphonotactically"), is accounted for in the theory of "morphonotactics", as proposed by Dressler and Dziubalska-Kořaczyk (2006). One of the claims they make is that markedness restrictions are less tight on morphonotactic clusters than on lexical ones, because the markedness of morphonotactic clusters might help to signal morphological operations. In fact, it is the very ill-formedness of morphonotactic clusters that makes them semiotically conspicuous, and therefore well-designed for signaling morphological complexity.¹

Some clusters are purely morphonotactic, i.e. in a given language they can be found only in morphologically complex words. This happens to be the case with /md/ in English (e.g. *seemed*, *doomed*) which never occurs in any morphologically simple form. Other clusters, such as /mp/ in English (e.g. *lamp*, *damp*) are purely lexical, since they can be found only in morphologically simple forms, and never arise from morphological operations. These two cases can be seen as located at the opposite ends of a scale indicating the degree to which a cluster is morphonotactic (Dressler and Dziubalska-Kořaczyk 2006). Between the two endpoints, intermediate cases can be found. Moving one step away from purely morphonotactic clusters, clusters which are morphonotactic by "strong default" can be encountered. A good example is English /ts/, which in the vast majority of cases shows up across word boundaries, as in *cats*, *sits* but occasionally inhabits morphologically simple words, as in *waltz*. The next step on the scale is occupied by clusters that are morphonotactic by "weak default", such as /ks/ in English *lacks*, *sacks* but also *tax*, *box* along with a number of other simple forms. These are more often morphonotactic than lexical, just like the strong default clusters are, but they show up as lexical more often than the

¹ On a related note, as one of the Reviewers points out, statistical information about phoneme sequences can also be used to help segment speech into words. For example, English phoneme pairs have been found to follow a strongly bimodal distribution in that each phoneme pair tends to appear either within words or across word boundaries (Hockema 2006).

strong default clusters do. And finally, there are clusters which are roughly as likely to appear in morphologically complex and morphologically simple forms, and these can be termed ambiguous. Such is the status of /nd/ in English, frequent in past tense and past participle forms such as *banned*, *sinned* but also in numerous simple forms such as *hand* or *kind*. In sum, the cluster types that occur in any language can always be arranged on a scale reflecting the different proportions to which they are lexical or morphonotactic. Now, as Dressler and Dziubalska-Kořaczyk (2006: 83) very plausibly suggest, the position of a given cluster on this scale influences the adequacy of that cluster to signal morphological complexity. The more lexical it is, the less well does it work as a signal of morphological complexity.

Crucially, this claim has interesting implications regarding the historical evolution of languages (cf. Dressler et al. 2010): although the factors it refers to are grounded in cognition and physiology, their reflexes should be observable in diachronic data.² Thus, if speakers seek to enhance the signaling function of clusters they should prefer (and select for) clusters to be more frequent near the two ends of the lexicality scale than near the middle (see also 2.1 below).

At the same time, however, there are also arguments for predicting the opposite. In particular, analogy effects have been posited (e.g. in Hogg and McCully 1987) by which morphonotactic cluster models support the emergence and stabilizes lexical counterparts – and *vice versa*. This clearly predicts that clusters should be more frequent near the middle of the lexicality scale than near its ends (see 2.2. below)

As for how such preferences shape the cluster inventory of a language, there is a number of mechanisms which can alter the status of a cluster as morphonotactic/lexical, as well as its prevalence diachronically. These include the reinterpretation of a morphonotactic cluster as lexical through loss of transparency (/l+θ/ → /lθ/ as in *health*), borrowing of loan words with a particular cluster (e.g. /ns/ as in *commence*, and a wealth of other Romance loan words); sound change such as cluster simplification (/mb/ → /m/ as in *bomb*), devoicing (/ld/ → /lt/ as in *spelt*, in certain dialects), schwa loss /nəd/ → /nd/ as in *sinned*), and schwa epenthesis (e.g. /lm/ → /ləm/ as in *film*, in certain dialects). The issue of the implementation of the change in status and prevalence at the level of an individual cluster, i.e. why a particular mechanism is applied to a particular cluster

² Thus, morphonotactics could provide tools for explaining long-term developments which otherwise defy a unified account, such as long-term reduction of consonantal complexity in English. This might be particularly true in cases where sociolinguistic explanations fail.

at a particular point in time, however, is beyond the scope of our investigation. We are taking a broader perspective and asking why such mechanisms are applied to particular clusters and not to others.

More specifically, we address the question of how morphological signaling and analogy interact and determine the diachronic evolution of consonant clusters by means of a mathematical dynamical-systems model. In a nutshell, the model describes the growth (or decline) of a population of cluster items that belong to a specific cluster type (like e.g. word-final /nd/). More precisely, this population of cluster items is divided into two subpopulations: morphonotactic instances and lexical instances of that cluster type, respectively. The dynamics of this population of cluster items is determined by various factors such as the likelihoods of transmitting or memorizing a cluster. Most importantly, the respective effects that morphonotactic signaling and analogy have on the growth of the cluster population are built into the model. The model then allows us to investigate the diachronic evolution of a specific cluster type with respect to its distribution of morphonotactic and lexical instances. If this is done for a large number of cluster types, one can simulate the diachrony of entire cluster inventories. These simulated developments finally are compared against historical data, in order to check the validity of the model.

Our research questions are the following:

- (1) How do the opposed forces deriving from the semiotic signaling function of clusters on the one hand and analogy effects on the other interact and contribute to the diachronic evolution and synchronic distribution of consonant clusters?
- (2) How are the clusters which can act as both lexical and morphonotactic spread over the lexicon in written/spoken, synchronic/diachronic data?
- (3) How do the results of mathematical modeling and simulations of the evolutionary dynamics of consonant clusters compare to empirically observable distributions?

The paper is structured as follows. Section 2 introduces the Semiotic Utility Function, which is a formal expression of the relationship between the degree to which a cluster is morphonotactic and its ability to signal morphological complexity. It formalizes the predictions of morphonotactics (Dressler and Dziu-

balska-Kołodziejczyk 2006: 83) for the shape of this function (2.1), discusses the role of analogy (2.2), and compares and contrasts the predictions of morphonotactics and of analogy for the distribution of ambiguous and unambiguous clusters in the lexicon (2.3). It ends with the description of the Polish and English data used in the study (2.4). Section 3 introduces our modeling approach, while Section 4 discusses the applications of the model to diachronic English data (4.1.) and synchronic Polish data (4.2). Section 5 summarizes our conclusions.

2. Opposing forces in morphonotactics

2.1. Morphonotactics and predictions about semiotic utility

Dressler and Dziubalska-Kołodziejczyk (2006: 83) state that “[p]rototypical morphonotactic clusters [...] have the function of co-signaling the existence of a morphological rule, morphonotactic default clusters [...] fulfill this function less adequately, while phonotactic clusters [...] cannot fulfill this function [...]”. As a formalization of this claim, we propose the Semiotic Utility Function. Semiotic utility is the degree to which a given cluster is capable of signaling morphological complexity. That the Semiotic Utility Function must be a decreasing function follows from the statement cited above. However, Dressler and Dziubalska-Kołodziejczyk’s claim is agnostic about the shape of the function (linear, concave or convex). The various possible shapes of the function are presented in Figure 1. Consider a consonant cluster that occurs only across morpheme boundaries, i.e. purely morphonotactically. While a convex shape indicates that by adding items to the lexicon which include the same cluster morpheme internally, the complexity-signaling ability of the morphonotactic cluster is substantially compromised, a concave shape means that adding the same items to the lexicon would not have a huge impact on the cluster’s semiotic utility.

We suggest that the curvature of the Semiotic Utility Function may be linked to the extent to which a specific language makes use of inflectional morphology. Languages clearly differ in the extent to which they rely on morphological coding versus word order. Polish relies on morphology much more than English does, as is clear from Sadeniemi et al.’s (2008) study employing both a Kolmogorov complexity-based approach and a morpheme-level comparison to measure the morphological complexity of languages.

Arguably, languages that tend to express grammatical functions by morphemes are expected to be more sensitive to lexical items that feature a usually

morphonotactic cluster morpheme internally, than languages that express the same grammatical functions analytically are, because the latter do not have to rely that much on proper recognition of morphemes. That Polish relies on clear boundaries between morphemes more than English does is suggested by the observation relating to language acquisition stating that “the segmentation of bound morphemes may be important much earlier for learners of highly inflecting languages [...]” (Peters 1997). Indeed, the hypothesis that “the richer the morphology of the language they are acquiring is, the faster children will develop morphology” (Dressler 2007: 3) is supported by the studies on the development on morphology in typologically disparate languages collected in Laaha and Gillis (2007), where morphology has been found to develop faster in children acquiring Slavic languages (Croatian and Russian) than Germanic languages (German and Dutch). This is in line with the idea that recognizing morphological boundaries is more important for speakers of languages with more synthetic morphology, such as Polish, than for speakers of more analytic languages, such as English. Thus, we will assume that synthetic languages exhibit more convex Semiotic Utility Functions, while analytic languages feature more concave ones.³ We will come back to this in Section 4, when we apply our approach to English (representing a more analytic language) and Polish (being a more synthetic one). Indeed, as we shall see, the shape of this function will play a crucial role in the diachronic development of consonant clusters.

2.2. Analogy and mutual support

There are analogy-based lines of argumentation for a mutually beneficial relationship between morphonotactic and lexical clusters. The general idea is that the clusters that occur morphonotactically support lexical occurrences, or that vice versa, the presence of a certain cluster in a given language lexically supports its morphonotactic occurrences. Hogg and McCully (1987: 47) point to a possible interaction of the former type, saying that “the type of syllable structure found in a simplex word such as *wind* (/waɪnd/) has been protected through analogy with inflected forms such as *weaned*”. In another study that corroborates analogical support from morphonotactic to lexical consonant clusters, Bau-

³ Crucially, this assumption is made *a priori* and independently from other phonological features of a language. Rather, the shape of this function is assumed to be conditioned by morphosyntactic properties of that language.

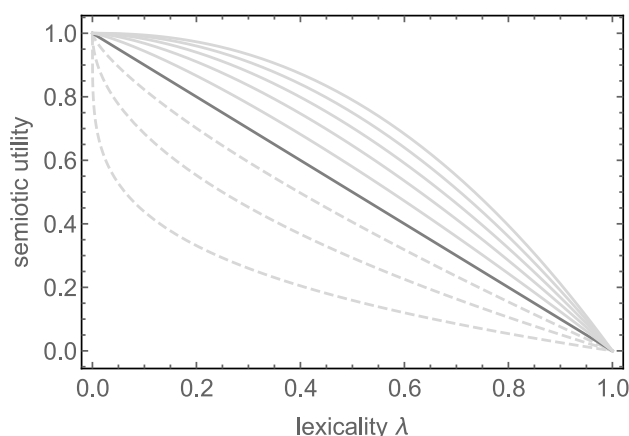


Figure 1. Semiotic utility as a decreasing function of lexicality λ , where λ is a measure of the number of simple word types in the lexicon a cluster occurs in. It can be linear (solid dark gray; each additional item decreases the cluster's semiotic signaling function to an equal extent), convex (dashed light gray; additional lexical items substantially decrease the signaling function of a primarily morphonotactic cluster) or concave (solid light gray; additional lexical items decrease the signaling function of a primarily morphonotactic cluster not to a large extent).

mann et al. (2015) suggest that the diachronic development of the English consonant cluster inventory reveals frequency effects among lexical and morphonotactic clusters.

On the other hand, based on a data driven study of English noun-noun compounds, Martin (2007: 99) claims that “the categorial phonotactic restrictions that hold within morphemes also hold gradiently across morpheme boundaries”, and thus indirectly supports the latter direction, in that morphonotactic consonant clusters benefit from being lexically licensed by their phonotactic counterparts. In total, this gives us a bidirectional mutual relationship between morphonotactic and lexical clusters.

2.3. Comparing the predictions

The two approaches to the relationship between morphonotactic and lexical clusters, the morphonotactic and the analogy-based approach, make contradicto-

ry predictions regarding the frequency distribution of cluster types. Following the logic of morphonotactics, clusters either specialize for the lexical domain or for the morphonotactic domain. Since the ambiguous clusters are not apt to signal morphological complexity, they should be susceptible to loss through language change. The surviving lexicon is expected to be populated by high proportions of purely morphological clusters, which signal morphological complexity reliably, and of purely lexical clusters, which do not signal morphological complexity at all. Thus, morphonotactics predicts a V-shaped distribution of morphonotactic and lexical clusters throughout the lexicon.

If we follow the analogy-based view of the relationship between morphonotactic and lexical clusters, on the other hand, we would predict a Λ -shaped distribution. If the two kinds of clusters are assumed to reinforce one another, then ambiguous clusters should be favored in language change. Such conditions would leave the lexicon populated with high rates of clusters which are morphonotactic and lexical at the same time. The two predictions are represented schematically in Figure 2.

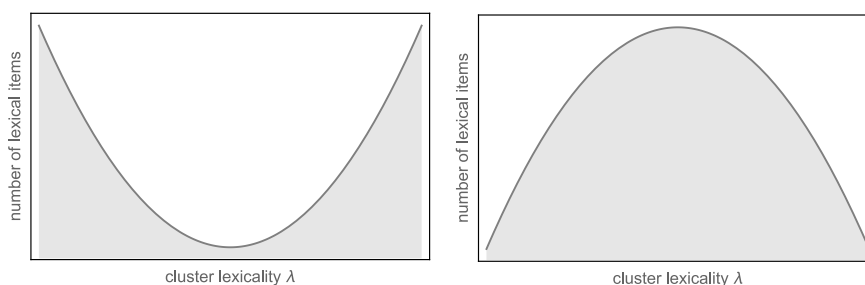


Figure 2. Schematic representation of the frequency distributions of cluster types with respect to the degree of lexicality λ which is a measure of the number of lexical items containing a given cluster type. On the left: Under the morphonotactic approach, a lack of ambiguous cluster types is expected, while most clusters are either primarily lexical or primarily morphonotactic. On the right: In the presence of mutually supporting analogical effects, most cluster types are expected to be ambiguous, i.e. medially lexical.

2.4. Data

For the present study we used three data sets. The first set comprises English diachronic written data. It is the current version of the database created by the

Evolution of Consonant Clusters in English (ecce) research project (www.ecce.univie.ac.at). It is a morphologically coded database of all final consonant clusters in the Penn-Helsinki Parsed Corpus of Middle (and Early Modern) English (Kroch and Taylor 2000; Kroch et al. 2004), containing approximately 240,000 tokens, spanning the 12th to the 17th century. The second set comprises Polish synchronic dictionary data, namely all initial clusters, yielding 232 cluster types spread over 2,137 words (Zydorowicz et al., forthcoming). The third set comprises Polish synchronic spoken data. It comprises all consonant clusters appearing in a corpus of spoken Polish, containing 375 cluster types distributed over 10,051 tokens (Kaźmierski 2015).

A preliminary inspection of the data allows for some conclusions regarding the predictions following from morphonotactics and from analogy for the distribution of cluster types. To begin with, the Polish data follow a V-shaped distribution. This is the case both for spoken and for dictionary data, with no appreciable difference in distribution between the two sets (comparison of 5-bin distributions; distribution-difference effect size estimated at Cramer's $V = .18 \pm .08$ (95% CI); $\chi^2 = 19.35$). Thus, we find no evidence of substantial analogy effects on the stability of consonant clusters in Polish. Additionally, the lack of difference between spoken and written data supports the validity of using written evidence for phonological research. This is welcome news not least in view of our diachronic English data, which by necessity, are written only. English diachronic data form a suitable testing ground for comparing the effects of morphonotactics and analogy. In Middle through early Modern English, large numbers of consonant clusters were created due to schwa loss. Schwa loss itself is assumed to have been phonologically conditioned, and not sensitive to morphology. However, if morphonotactics and/or analogy are indeed relevant for the diachronic survival of clusters, the reflexes of these forces should be seen in the English data. Indeed, for the English data set, we observe diachronic change in the distribution of morphonotactic and lexical clusters. Comparing the frequency distribution of clusters with increasing ratio of lexical tokens for the earliest stage (EME) with that of a later stage (EmodE), we notice a transition from a V-shaped bimodal to a W-shaped trimodal distribution (see Figure 3). This indicates a combined influence of morphology and analogy. While analogy not playing a role would give a V-shaped bimodal distribution and analogy playing a dominating role would give a Λ -shaped unimodal distribution, a W-shaped distribution is a combination of the two, with analogy exerting some, but not all-powerful, influence.

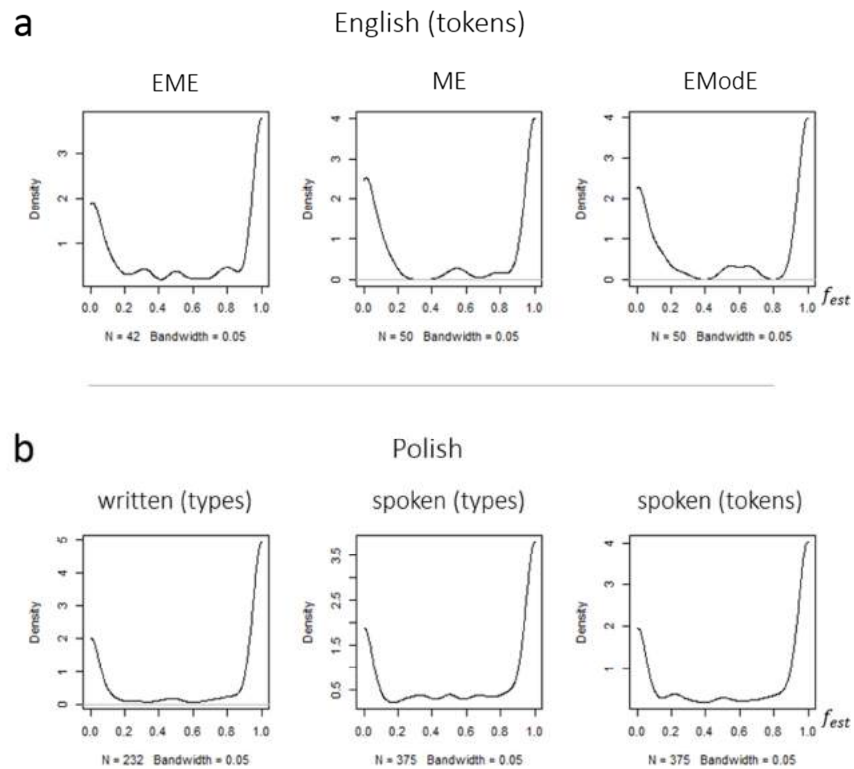


Figure 3. Estimated density distributions for the English and Polish consonant cluster inventories (Gaussian kernel; fixed bandwidth $h = .05$). (a) Diachronic development of the distribution of fractions of lexical cluster tokens $f_{est} = L_{est}/(M_{est} + L_{est})$ from Early Middle English (EME) to Early Modern English (EModE), where L_{est} and M_{est} are the estimated token frequencies of lexical and morphonotactic clusters, respectively. Diachronically, the distribution establishes a trimodal W-shape. (b) Distributions of f_{est} -values in terms of type frequency (written and spoken) and token frequency (spoken only). All distributions are V-shaped and bimodal. The difference between the former two distributions (comparison of 5 equally sized bins) is small at an effect size of $V = .18 \pm .08$ (95% CI); $\chi^2 = 19.35$.

Thus, the distributional shape for Polish is consistent with the predictions of morphonotactics, and the distributional shape for English indicates a joined influence of both morphonotactics and analogy. The modeling presented in the

following section tackles the issue of how such distributional shapes come about.

3. Modelling approach

3.1. Modelling the interaction of consonant clusters

As outlined above, analogy effects will, in the long run, lead to configurations with large numbers of ambiguous consonant cluster types, with respect morphonotacticity, while morphonotactics predicts distributions characterized by large numbers of purely morphonotactic cluster types on the one hand and purely lexical cluster types on the other hand. However, in general it is difficult to assess what the development looks like when we face an interaction between these two pressures. The long-term dynamics that result from such a complex interaction are by no means trivial.

Thus, mathematical modeling suggests itself. It allows for a formalization of the interacting forces, and, at least, qualitative predictions. There might be other factors, like the precise social structure of the speaker community, that influence the dynamics quantitatively; however, they would probably not change the overall behavior of the dynamics. The predictions of a mathematical model then are – not only but also – as qualitative in nature as the diachronic predictions that are derived from hypotheses about analogy and/or morphonotactics. The model does nevertheless provide a more elaborate and complete picture.

In the present study we opt for a Lotka-Volterra type structured-population dynamical system. This model family is chosen not because of its quantitative accuracy, but due to its flexibility and simplicity. It is well studied and allows for a straightforward analysis of different types of interactions, such as competition, parasite-host configurations, or – as in this paper – mutualism, i.e. bidirectionally supporting effects (Hofbauer and Sigmund 1998: 28–29).

The term *structured population* refers to the fact that the population of instances of a given consonant cluster type, e.g. /nd/, is composed of two different subpopulations, namely morphonotactic instances and lexical instances of that cluster type. What the mathematical model does is that it simply tracks the frequencies of morphonotactic and lexical instances in time. We do not model the dynamics of uttered tokens on the one hand and speakers knowing a certain cluster on the other hand explicitly. Rather, we assume for simplicity that the abundances of uttered instances (that can be retrieved from a corpus) reflect the

cognitive entrenchment of linguistic constituents in the speaker population (see e.g. Pierrehumbert 2001 and Wedel 2006 for exemplar based accounts; but also Yang 2000 for an approach in a generative framework and Nowak 2000 and Solé 2011 for related models of word dynamics). The two subpopulations interact according to the forces introduced in 2.1 and 2.2, and the parameters that specify the growth of the subpopulations reflect these interactions. In the following, these parameters will be described and motivated. For more formal details and a rigorous analysis of the model see Authors (#####) as well as the Appendix.

First, the model incorporates the most fundamental parameters responsible for the spread of linguistic items: on the one hand reproduction, i.e. the combination of perception, learning and production, and on the other hand disappearance due to restricted memory or speaker death. We furthermore assume that the growth of linguistic items is bounded from above, since the number of linguistic items within memorized utterances, the number of utterances per speaker, and the number of speakers are obviously limited. These factors are cognitive, physiological, biological or simply physical in nature. Social biases are not explicitly incorporated, although this could easily be done (e.g. by increasing the reproduction rate of a certain class of items). This is, however, not the focus of this paper. In the mathematical model, reproduction is represented by a strictly positive growth rate r , meaning that per time unit r items are produced per already existing item. That is, if for example a consonant cluster occurs in an utterance, it will on average lead to the utterance of r new instances of the same type. In the absence of other populations of linguistic items, growth is bounded by a so-called carrying capacity K , which represents the maximal number of slots in memorized utterances, in which linguistic items of interest could be found. It deserves a more elaborate discussion. As suggested above, the amount of items should be limited for multiple reasons. We assume that the reproduction rate decreases linearly as the population size approaches K . This means that the growth of the population depends (a) on the number of its reproducing members and (b) on the number of yet available slots in which its members can be placed. This type of regulating population growth is referred to as logistic growth and regarded as the null model of density regulation (Otto and Day 2007: 75). Logistic behavior corresponds to S-shaped diachronic trajectories, which have been frequently observed in diachronic linguistics (Kroch 1989; Denison 2003; Blythe and Croft 2012) and applied to models of language or word spread (Solé 2011: 169–170; Nowak 2000). Finally, disappearance is formalized by a strictly positive parameter d so that per time unit, d items are removed per existing item (cf.

Pierrehumbert 2001). This is plausible, since memorized utterances can be forgotten or simply disappear due to speaker death.

In our model, the parameters characterizing the growth of morphonotactic consonant clusters can differ from that of their lexical counterparts. In particular, we assume that the morphonotactic growth rate (r_M) is larger than the lexical growth rate (r_L), and that the lexical growth rate depends on the number of lexical items (i.e. morphemes and words) which a consonant cluster can occur in. This amount shall be expressed by a parameter λ , which is – for convenience – normalized to the unit interval. Note in particular that λ is a normalized measure of type frequency rather than token frequency. We furthermore assume that – in line with morphonotactics – the reproduction rate of the morphonotactic subpopulation is supposed to be decreasing in λ , which represents the decreasing semiotic utility for signaling morphological complexity of the cluster. This will play a key role in the evolutionary dynamics of the system.

Finally, we build mutually supporting analogy into the model. We assume that analogical transfer between morphonotactic and lexical consonant clusters is proportional to the product of the two population sizes. In the theory of dynamical systems this is referred to as the law of mass action (Heesterbeek 2005). If one population has a positive impact on another population, it is more likely that the latter benefits from the former if both populations are large than if one of them is rare. This is plausible also from a linguistic point of view. Analogical transfer between two populations of sufficiently similar linguistic items takes place if they co-occur in a language at a given point in time. We formalize this as a strictly positive interaction parameter a , meaning that per co-occurrence event of morphonotactic and lexical consonant clusters a new clusters of each type are produced. Note that this allows the populations to exceed K . This is reasonable, since via analogy the chance of memorizing an item due to the presence of another item can be larger than in cases where the latter is absent.

3.2. Dynamics of consonant clusters

Under the assumption of the analogical-interaction parameter a and the disappearance parameter d being sufficiently small one can show that the model outlined in the previous subsection converges to a stable mix of morphonotactic and lexical consonant clusters (Hofbauer and Sigmund 1998: 28–29; see also Figure A1). The question now is how this mix changes if the number of lexical items containing a cluster represented by the parameter λ varies and, even more

importantly, whether λ gets larger or smaller, i.e. whether a cluster type evolves in such a way that it occurs in more and more lexical items (e.g. by interconsonantal schwa deletion as in ME *godes* to ModE *gods*), or whether its occurrence in lexical items gets reduced (e.g. by cluster reduction as in *womb* from ME to ModE).

The model allows us to answer this question. Given a cluster-type specific λ used in the language and a new and different value λ' that comes into usage, one can determine if the new variant – which stands for a configuration in which the cluster occurs in more or less lexical items – replaces the old variant, or if no change occurs. The latter is the case, if an innovation like a hypothetical cluster reduction of *st* to, say, *s* in *vast* or *gist* is not supported. By this logic, we can determine the evolutionary trajectory of λ by means of a sequence of innovation-invasion-substitution events (cf Appendix A.1).⁴

As described in the previous section, the dynamics of the system depend on the parameter λ in that the reproduction rate of the morphotactic subpopulation is assumed to decrease in this parameter due to decreasing semiotic utility. This is a direct consequence of how semiotic utility is operationalized (as described in 2.1). Since languages differ in the extent to which they make use of morphology, we suppose that the precise shape of the Semiotic Utility Function is language specific. As outlined in 2.1, our conjecture is that languages with elaborate morphological systems are more sensitive in the sense of morphotactics, i.e. the semiotic utility of a morphotactic cluster will already be to a large extent decreased if it also appears non-morphotactically in a small number of lexical items. Thus the Semiotic Utility Function is assumed to be convex (dashed curves in Figure 1). On the other hand, languages making less use of inflectional or derivational morphology are supposed to be more tolerant to ambiguous clusters, so that their dynamics are better described by employing concave Semiotic Utility Functions (solid light-gray curves in Figure 1; see also Appendix A.2).⁵

⁴ Note that this differs from the sequence of innovation/actuation and propagation proposed by Croft (2000: 37–38, 98) in that in the present approach, innovation is not necessarily functionally driven while propagation is not exclusively socially but rather also functionally driven (cf. Ritt 2004: 74–78).

⁵ It should not be unmentioned that, as a Reviewer of this paper points out correctly, one could think of more complex shapes than just uniformly convex, linear, or concave functions, such as sigmoid or piecewise sigmoid shapes. While this is undoubtedly true (and psycholinguistically indeed likely), implementing more complicated shapes would at the same time increase the complexity of the model significantly: the functions which we use for modelling decreasing semiotic utility

Thus, on the evolutionary time-scale our model predicts the diachronic development of the number of lexical items a given cluster type occurs in specified by λ which in turn determines the numbers of morphonotactic and lexical cluster instances, which shall be labelled M_λ and L_λ , respectively. From these values the fraction of lexical instances can be calculated according to $f_\lambda = L_\lambda / (M_\lambda + L_\lambda)$ for each point in time, where $f_\lambda = 0, 0.5$, and 1 correspond to purely morphonotactic, maximally ambiguous, and purely lexical configurations, respectively (see Authors #####). In this way, the model shows the trajectory of the relative frequency of lexical instances of a given cluster type. In the subsequent section it will be shown by means of simulations that the long term dynamics of this relative frequency depends on the language specific shape of the Semiotic Utility Function. We will see that, depending on this shape, clusters can either evolve towards the boundaries (i.e. so that they become either purely morphonotactic or purely lexical) or towards a fair mix of morphonotactic and lexical cluster instances. Note that the system solely models the evolution of single cluster types on the scale from purely morphonotactic to purely lexical. It does not account for the complete loss of consonant clusters once they have reached either of the two boundaries.

4. Applications of the model

4.1. Diachrony of English

The model described above needs to be tested against actual language data. We will start with an analysis of the diachronic development of the English inventory of word final consonant clusters from the beginning of the ME period (12th cent.) up to the EModE period (17th cent.). The respective distributions have been described in 2.4 and Figure 3a. We expect the predictions of the proposed model to be consistent with the distributions found in the 12th-century and 17th-century data.

We proceed as follows. First, the fractions of lexical instances $f_{est} = L_{est} / (M_{est} + L_{est})$ are estimated for each consonant cluster in the 12th-century

are power functions of the form $\lambda \rightarrow 1 - \lambda^c$ (see Appendix A.1), while sigmoid functions usually involve exponential terms. The function family chosen in this approach provides a reasonably cheap way of accounting for a large number of functional relationships. Thus, the present approach strikes a balance between formal complexity and explanatory power.

cluster inventory in order to determine the distribution of f_{est} values. From this empirically estimated initial distribution a random sample of $N = 80$ values⁶ is drawn (empirical initial distribution in Figure 4a). This set serves as a starting point for the simulation (simulated initial distribution in Figure 4a). Since English is regarded as moderately inflectional and derivational as compared to other languages such as German or Slavic languages (Szmrecsanyi 2012), we assume the Semiotic Utility Function to be non-convex.⁷ For each initial value in the simulated initial distribution, a sequence of repeated innovation-invasion-substitution events is simulated according to the structured-population dynamical system outlined in the previous section. In order to make the model more realistic, we allow for uniformly random fluctuations. The respective trajectories are shown in the middle panel in Figure 4a. In general, we see that trajectories that start close to the boundaries evolve towards even more extreme f_λ values, while trajectories initiating in more or less ambiguous configurations evolve towards a fair mix of morphonotactic and lexical instances. After several simulation steps, the resulting distribution of f_λ values is determined and compared to the empirical distribution of f_{est} from the 17th century data. It can be seen that the simulated final distribution and the empirical final distribution both display a considerable number of ambiguous clusters (histograms in the right panel of Figure 4a).

That is, under the assumption that an English speaker's ability to process morphonotactic consonant clusters is not to a large extent reduced by the occurrence of the same cluster type within a small number of lexical items, the model predicts that the English consonant-cluster inventory evolves in such a way that there are (a) a large number of purely morphonotactic consonant clusters, (b) a large number of purely lexical consonant clusters, and (c) a substantial number of ambiguous consonant clusters that show a fair share of morphonotactic and lexical instances (but see Appendix A.2 for more formal information about the diachronic reflex of the Semiotic Utility Function's curvature on the set of possible evolutionary outcomes). This is consistent with the trimodal distribution we find in the English data.

⁶ The sample size is motivated by a pre-simulation power analysis which revealed that – if actually present – significant medium differences in terms of Cramer's V between the simulated and the empirical final 5-bin distributions, respectively, could be detected at a power of .80.

⁷ In particular, note that according to quantitative analyses conducted by Szmrecsanyi (2012: Fig. 3) no monotonous trend of the syntheticity and analyticity of English is identifiable in the ME period. This supports our notion that the curvature of the English semiotic utility function stayed constant from the 12th to the 17th century.

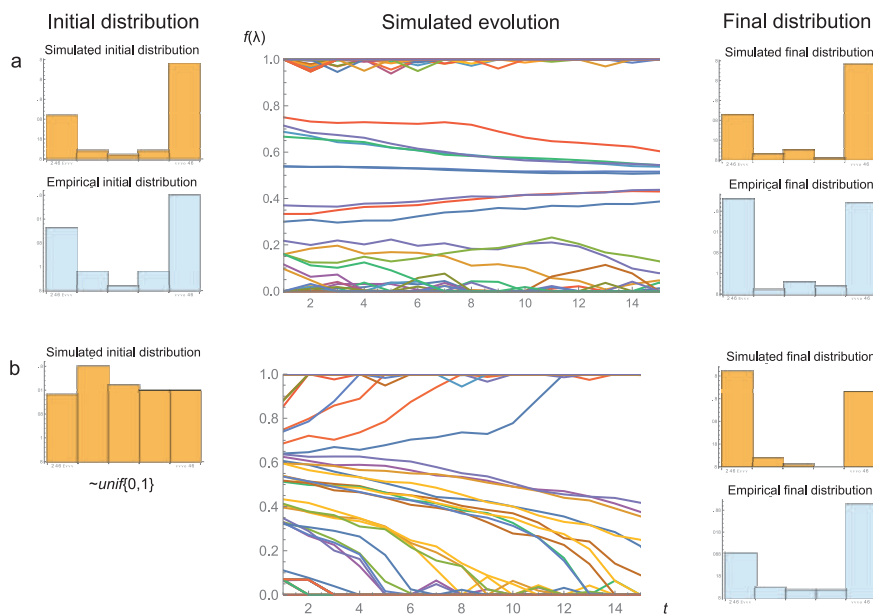


Figure 4. Simulating the evolution of the (a) English and (b) the consonant-cluster inventory, respectively. Orange histograms represent simulated distributions while blue histograms represent empirically measured distributions. The repeated invasion-analysis simulations ($N = 80$) are based on (a) a model with a non-convex Semiotic Utility Function in the case of English and (b) a model with a convex Semiotic Utility Function in the case of Polish ($a = 0.01$, $d = 2$, $K = 50$, $r_M = r_L = 10$, random drift $\sim \text{unif}\{\pm 0.01\}$).

4.2. Synchronic analysis of Polish data

In contrast to English, the Polish consonant-cluster inventory exhibits a bimodal distribution. Cluster types tend to be either unambiguously morphonotactic or unambiguously lexical, while ambiguous consonant clusters are relatively rare (see 2.4 and Figure 3b). While we have collected solely synchronic Polish data, which renders the direct comparison of simulated and empirically observed diachronic trajectories impossible, we can still test the proposed model against the Polish data.

Instead of using a historical Polish cluster-inventory distribution as initial state for the simulation, we are even stricter and test whether the Present Day

Polish distribution would be approximated when starting with a completely random distribution. Thus, the initial values f_λ for the simulation are randomly drawn from a uniform distribution on the unit interval $unif(0,1)$. Note, that this does not and should not in any way reflect the distribution of the Polish consonant-cluster inventory at any point in time. Rather, in the lack of historical data we stay completely agnostic about the history of Polish and test our model against the least informed initial state.

As in 4.1, the precise shape of the Semiotic Utility Function must be specified. In contrast to what was assumed for the English case, a convex semiotic signaling function was chosen for Polish. This means that already small numbers of lexical items containing a consonant cluster substantially decrease the semiotic signaling function of its morphonotactic version. This can be motivated by the inflectional and derivational richness of Polish.

The evolution from the assumed initial state towards the Polish-like simulated final distribution is shown in Figure 4b. After $t = 15$ simulation steps, the resulting distribution of f_λ values is bimodal, i.e. the cluster-inventory evolves in such a way that clusters are predominantly either purely morphonotactic or purely lexical, while ambiguous consonant clusters have almost disappeared (cf. Appendix A.2). This is as expected and coincides with the configuration found in the synchronic Polish data.

5. Conclusion

The complexity-signaling ability of morphonotactic consonant clusters and analogy effects between morphonotactic and lexical consonant-cluster instances constitute two cognitive forces which are crucial to the replication of consonant clusters. They are counteracting in the sense that the complexity-signaling ability – in other words: the semiotic utility – of morphonotactic clusters is handicapped by the appearance of lexical clusters, while analogy constitutes a mutual relationship between the two cluster categories. The interaction of these two forces is complex and its diachronic reflexes are nontrivial. However, it can be addressed by means of a mathematical model which tracks the respective abundances of the morphonotactic and lexical versions of a given cluster type, and allows for an analysis of the diachronic development of its degree of morphonotacticity. We have shown in this paper that the way in which the semiotic utility of consonant clusters is operationalized is crucial for the diachronic long-term development.

It has been shown that the development of the inventory of English word-final consonant clusters from the 12th to the 17th century is best captured by the model, if a non-convex Semiotic Utility Function is assumed (cf. Appendix A.2). This means that the semiotic utility of a consonant cluster type is not substantially diminished by the presence of lexical instantiations of that cluster. Crucially, the implementation of analogy is necessary for obtaining a trimodal W-shaped distribution. In the absence of analogical effects between morphotactic and lexical consonant clusters, the stable amount of ambiguous English clusters could not be accounted for. We take this to support the hypotheses about the relevance of analogy in the reproduction of morphotactic consonant clusters, which have been outlined in 2.2 (see Appendix A.3 for a slightly more formal argument).

In contrast, Polish synchronic data exhibit a bimodal V-shaped distribution in which ambiguous consonant clusters are considerably less prominent. In the model, this can be explained by a very convex Semiotic Utility Function, which in turn means that the reproduction of morphotactic consonant clusters is impeded significantly by the appearance of lexical items that contain a lexical consonant cluster. It is worth pointing out that the model shows that even if there are analogical effects, this can be compensated by a very convex Semiotic Utility Function, so that finally the amount of ambiguous consonant clusters gets reduced.

While the theory of morphotactics suggests that the Semiotic Utility Function is decreasing, this paper stresses – in addition – the relevance of its curvature. The shape of the Semiotic Utility Function thus is modeled as to be language specific. As a motivation, we hypothesize that it is related to the extent to which a language makes use of inflectional or derivational morphology, i.e. its syntheticity. Speakers who are exposed to elaborated morphological systems are supposed to be challenged by slightly ambiguous clusters to a larger extent than speakers of less synthetic languages. We take this as a starting point for further (i) experimental research which tries to evaluate the language specific shapes of the Semiotic Utility Function and (ii) quantitative and comparative corpus research on the respective (mor)photactic consonant-cluster inventories.

Appendix

A.1 Mathematical outline of the model and its evolutionary analysis

In the following, the model which underlies the simulation described in Section 3 and 4 shall be outlined in a bit more mathematical detail. The dynamics of the densities of morphonotactic consonant clusters M and lexical or phonotactic consonant clusters L are modelled by means of a two-dimensional dynamical system in continuous time, which belongs to the Lotka-Volterra model family. The changes in the respective densities, i.e. the first derivatives of M and L with respect to time, are given by the system

$$\begin{pmatrix} \dot{M} \\ \dot{L} \end{pmatrix} = A(\lambda) \cdot \begin{pmatrix} M \\ L \end{pmatrix},$$

where

$$A(\lambda) := \begin{pmatrix} r_M u(\lambda) \left(1 - \frac{M}{K}\right) - d & a \\ a & r_L \lambda \left(1 - \frac{L}{K}\right) - d \end{pmatrix}$$

is a 2 by 2 matrix depending on $\lambda \in [0,1]$. The parameters r_M, r_L (i.e. the reproduction rates r specific to M and L , respectively), K, d and a are assumed to be real and strictly positive. They are motivated in Section 3. Semiotic utility is formalized by a strictly decreasing function $u: \lambda \rightarrow 1 - \lambda^c$ on the unit interval, where the exponent $c > 0$ determines the curvature of u . For $c > 1$ the function is concave and for $c < 1$ it is convex (Figure 1). If d and a are sufficiently small, the system exhibits a single stable population-dynamical attractor $(\hat{M}_\lambda, \hat{L}_\lambda)$ (Hofbauer and Sigmund 1998: 29). For most functions u an explicit analytical representation of this equilibrium would be too large and complicated, so that it cannot be shown in this paper. The ecological dynamics shall, however, be illustrated in Figure A1.

The successful invasion of a new configuration of consonant clusters characterized by λ' can be shown to depend on

$$s(\lambda, \lambda') := \mu(A(\lambda'))|_{(M_{\lambda'}, L_{\lambda'}) = (\hat{M}_\lambda, \hat{L}_\lambda)},$$

where μ is the stability modulus (i.e. the largest Eigenvalue of the Jacobian matrix of the system) and where (M_λ, L_λ) shall denote the pair of the densities of morphonotactic and lexical consonant clusters depending on the number of lexi-

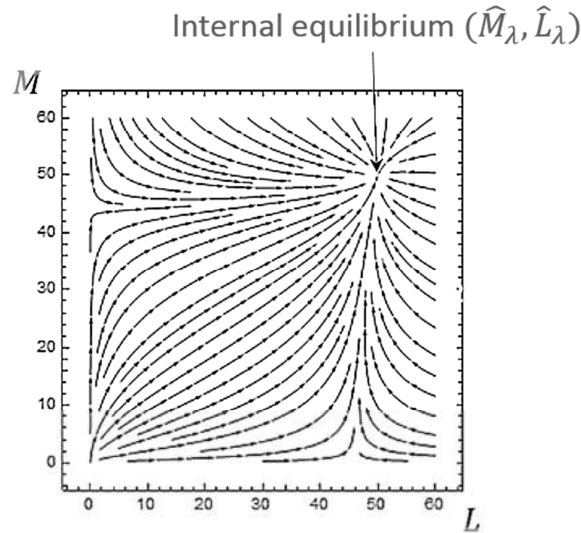


Figure A1. Phase portrait of the ecological dynamics of the structured population. The horizontal axis measures the number of lexical instances while the vertical axis measures the number of morphonotactic items of a specific cluster type. If death rates and analogy rates are sufficiently small, the population always converges to a stable ecological equilibrium consisting of a mixed population of morphonotactic and lexical clusters. In this example there are as many lexical items as morphonotactic ones.

cal items represented by λ . See Geritz et al. (1998), Dercole and Rinaldi (2008: ch. 2.8), and Hoyle and Bowers (2008) for more details. If the difference between λ and λ' is sufficiently small,

$$s(\lambda, \lambda') > 0$$

implies that the new configuration replaces the old one (Geritz et al. 2002) and converges to a new ecological equilibrium $(M_{\lambda'}, L_{\lambda'})$ (as in Figure A1). Otherwise λ stays the same. Thus, by checking the latter inequality condition for each pair of successive values λ and λ' , one can keep track of the change in the number of lexical items a cluster occurs in. This is what is done at each step of the simulation. Furthermore, we allow for random fluctuations drawn from a uniform distribution at each simulation step. The sequence of all simulation steps finally constitutes the long-term evolutionary trajectory for each initial value of λ .

A.2 Analytical results: possible evolutionary outcomes

Certain qualitative properties of the evolutionary dynamics of λ can, in fact, be predicted analytically. In particular, it can be shown that the set of possible evolutionary outcomes differ depending on the curvature of the Semiotic Utility Function. For concave, linear, and mildly convex Semiotic Utility Functions, λ will either approach an intermediate value (a so-called mixed evolutionarily stable strategy, ESS) or one of the two boundaries of the unit interval. In contrast, if the Semiotic Utility Function is sufficiently convex, then the mixed ESS disappears so that λ always approaches either 0 or 1, depending on its initial position. A rigorous derivation of these results is beyond the scope of this paper, but see Hoyle & Bowers (2008) and Rűffler et al. (2004) for some discussion on the relationship between trade-off curvatures and possible evolutionary outcomes.

For the applications discussed in Section 4, this has the following consequence. If the Semiotic Utility Function is not strongly convex (as in the English case), then clusters with intermediate λ values will always move towards the middle of the unit interval so that an accumulation of ambiguous clusters get visible after a sufficient number of simulation steps. In contrast, if the Semiotic Utility Function is strongly convex, the final distribution will always end up being bimodal with a large numbers of non-ambiguous clusters, as long as not all clusters enter the simulation in the same basin of attraction of either 0 or 1.

A.3 Evolutionary optimization in non-interacting populations

The above described model is based on the assumption that morphonotactic and lexical cluster items interact via analogy. In the model, this is represented by the condition that a is strictly positive. Indeed, if it were zero, i.e. if there was no interaction between the two populations of clusters, then the evolutionary dynamics of λ would be much less complicated. If $a = 1$, then the matrix A becomes reducible so that the corresponding stability modulus only depends on one of the two diagonal entries (Horn & Johnson 1985). Evolutionarily, this inevitably leads to an optimization of λ towards either 0 or 1 (Metz and Diekmann 2008). In this scenario, mixed strategies, i.e. ambiguous clusters, will always disappear in the long run.

Acknowledgements

The research reported on in this paper was supported by Österreichischer Austauschdienst grant No. PL8/2014 and by Austrian Science Fund (FWF) grant No. P27592-G18.

References

- Baumann, A., C. Prömer and N. Ritt. 2015. "Diachronic reflexes of frequency effects among word final (mor)phonotactic consonant clusters in Middle and Early Modern English". Paper presented at the 3rd International Workshop on Phonotactics and Phonotactic Modeling.
- Blythe, R.A. and W. Croft. 2012. "S-curves and the mechanism of propagation in language change". *Language* 88(2). 269–304.
- Croft, W. 2000. *Explaining language change. An evolutionary approach*. Harlow, New York: Longman.
- Denison, D. 2003. "Log(ist)ic and simplistic S-curves". In: Hickey, R. (ed.), *Motives for language change*. Cambridge: Cambridge University Press. 54–70.
- Donohue, M., R. Hetherington, J. McElvenny and V. Dawson. 2013. "World phonotactics database". Department of Linguistics, The Australian National University. <<http://phonotactics.anu.edu.au>>. Last accessed: 16 Sep 2015.
- Dressler, W.U. 2007. "Early development of nominal and verbal morphology from a typological perspective". In: Laaha, S. and S. Gillis (eds.), *Typological perspectives on the acquisition of noun and verb morphology. (Antwerp Papers in Linguistics 112.)* 3–9.
- Dressler, W.U., K. Dziubalska-Kořaczyk and L. Pestal. 2010. "Change and variation in morphonotactics". *Folia Linguistica Historica* 31. 51–67.
- Dressler, W.U. and K. Dziubalska-Kořaczyk. 2006. "Proposing morphonotactics". *Wiener Linguistische Gazette* 73. 1–19.
- Dziubalska-Kořaczyk, K. and P. Zydorowicz. 2014. "The production of high-frequency clusters by native and non-native users of Polish". *Concordia Working Papers in Applied Linguistics* 5. 130–144.
- Geritz, S.A.H., E. Kisdi, G. Meszřena and J.A.J. Metz. 1998. "Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree". *Evolutionary Ecology* 12. 35–57.
- Geritz, S.A.H., M. Gyllenberg, F.J. Jacobs and K. Parvinen. 2002. "Invasion dynamics and attractor inheritance". *Journal of Mathematical Biology* 44. 548–560.
- Heesterbeek, J. 2005. "The law of mass-action in epidemiology: A historical perspective". In: Cuddington, K. and B. Beisner (eds.), *Ecological paradigms lost: Routes of theory change*. Amsterdam: Academic Press. 81–104.
- Hockema, S.A. 2006. "Finding words in speech: An investigation of American English". *Language Learning and Development* 2(2). 119–146.

- Hofbauer, J. and K. Sigmund. 1998. *Evolutionary games and population dynamics*. Cambridge: Cambridge University Press.
- Hogg, R. and C.B. McCully. 1987. *Metrical Phonology: A course book*. Cambridge: Cambridge University Press.
- Horn, R. and C. Johnson. 1985. *Matrix analysis*. Cambridge: Cambridge University Press.
- Hoyle, A. and R.G. Bowers. 2008. "Can possible evolutionary outcomes be determined directly from the population dynamics?" *Theoretical Population Biology* 74. 311–323.
- Jarosz, G., S. Calamaro and J. Zentz. Submitted. "Input frequency and the acquisition of syllable structure in Polish". *Language Acquisition*.
<http://people.umass.edu/jarosz/jarosz_etal_submitted.pdf>
- Kaźmierski, K. 2015. "Consonant clusters in casual speech in Polish: Testing the predictions of NAD and Beats-and-Binding phonology". Paper presented at the 23rd Manchester Phonology Meeting.
- Kroch, A. 1989. "Reflexes of grammar in patterns of language change". *Language Variation and Change* 1. 199–244.
- Kroch, A. and A. Taylor. 2000. "Penn-Helsinki Parsed Corpus of Middle English". <<http://www.ling.upenn.edu/hist-corpora/>>.
- Kroch, A., B. Santorini and L. Delfs. 2004. "Penn-Helsinki Parsed Corpus of Early Modern English". <<http://www.ling.upenn.edu/hist-corpora/>>.
- Laaha, S. and S. Gillis (eds.). 2007. *Typological perspectives on the acquisition of noun and verb morphology*. (*Antwerp Papers in Linguistics* 112.)
- Madelska, L. 2005. *Słownik wariantywności fonetycznej współczesnej polszczyzny* [A dictionary of phonetic variability in contemporary Polish]. Kraków: Collegium Columbinum.
- Martin, A. 2007. The evolving lexicon. (PhD dissertation, UCLA, Los Angeles.)
- Metz, J.A.J., S.D. Mylius and O. Diekmann. 2008. "When does evolution optimize?" *Evolutionary Ecology Research* 10. 629–654.
- Nowak, M.A. 2000. "The basic reproductive ratio of a word, the maximum size of a lexicon". *Journal of Theoretical Biology* 204 (2). 179–189.
- Otto, S. and T. Day. 2007. *A biologist's guide to mathematical modeling in ecology and evolution*. Princeton: Princeton University Press.
- Peters, A.M. 1997. "Language typology, prosody and the acquisition of grammatical morphemes". In: Slobin, D.I. (ed.), *The crosslinguistic study of language acquisition* (vol. 5). Hillsdale, NJ: Lawrence Erlbaum Associates. 136–197.
- Pierrehumbert, J. 2001. "Exemplar dynamics: Word frequency, lenition, and contrast". In: Bybee, J. and P.J. Hopper (eds.), *Frequency effects and the emergence of lexical structure*. Amsterdam: John Benjamins. 137–157.
- Rueffler, C., T. Van Dooren and J.A.J. Metz. 2004. "Adaptive walks on changing landscapes: Levins' approach extended". *Theoretical Population Biology* 65. 165–178.
- Sadeniemi, M., K. Kettunen, T. Lindh-Knuutila and T. Honkela. 2008. "Complexity of European Union languages: A comparative approach". *Journal of Quantitative Linguistics* 15(2). 185–211.
- Shockey, L. 2003. *Sound patterns of spoken English*. Malden, MA: Blackwell.
- Solé, R. 2011. *Phase transitions*. Princeton: Princeton University Press.

- Szmrecsanyi, B. 2012. "Analyticity and syntheticity in the history of English". In: Nevalainen, T. and E. Closs Traugott (eds.), *The Oxford handbook of the history of English*. Oxford: Oxford University Press. 654–665.
- Wedel, A. 2006. "Exemplar models, evolution and language change". *The Linguistic Review* 23. 247–274.
- Yang, C. 2000. "Internal and external forces in language change". *Language Variation and Change* 12. 231–250.
- Zydorowicz, P., P. Orzechowska, M. Jankowski, K. Dziubalska-Kołaczyk, P. Wierchoń and D. Pietrala. Forthcoming. *Phonotactics and morphonotactics of Polish and English. Theory, description, tools and applications*.

Address for correspondence:

Andreas Baumann
Institut für Anglistik und Amerikanistik
Universität Wien
Campus d. Universität Wien
Spitalgasse 2-4/Hof 8.3
1090 Wien
Austria
andreas.baumann@univie.ac.at