Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence

Christian Bentz
University of Tübingen
DFG Center for Advanced Study
Rümelinstraße 23
Tübingen, 72074, Germany
`chris@christianbentz.de`

Aleksandrs Berdicevskis
UiT The Arctic University of Norway
Department of Language and Culture
Postbox 6050 Langnes
9037 Tromsø, Norway
`aleksandrs.berdicevskis@uit.no`

Abstract

The morphological complexity of languages differs widely and changes over time. Pathways of change are often driven by the interplay of multiple competing factors, and are hard to disentangle. We here focus on a paradigmatic scenario of language change: the reduction of morphological complexity from Latin towards the Romance languages. To establish a causal explanation for this phenomenon, we employ three lines of evidence: 1) analyses of parallel corpora to measure the complexity of words in actual language production, 2) applications of NLP tools to further tease apart the contribution of inflectional morphology to word complexity, and 3) experimental data from artificial language learning, which illustrate the learning pressures at play when morphology simplifies. These three lines of evidence converge to show that pressures associated with imperfect language learning are good candidates to causally explain the reduction in morphological complexity in the Latin-to-Romance scenario. More generally, we argue that combining corpus, computational and experimental evidence is the way forward in historical linguistics and linguistic typology.

1 Introduction

Languages inevitably change over time. Sounds are added or removed from phoneme inventories, morphological markers are grammaticalized or lost, word orders permute in historical variants. Causal explanations for these phenomena are often complex – or lacking all together – since they have to cope with the interplay of learning and usage in a multitude of social settings, at different times, in different places.

We here focus on a prominent change from Latin towards the Modern Romance languages: the systematic loss of *morphological markers* in all descendant languages of the common proto-language over hundreds and thousands of years. Latin marked grammatical functions by means of inflectional variants of the same word root, thus displaying complex word forms. For example, the Latin word for "brother" *frater* was inflected to yield *fratres*, *fratribus*, *fratris*, *fratrum*, *fratri*, *fratre*, etc. according to singular/plural and case distinctions. This complexity is considerably reduced in Modern Romance languages, where there are often simpler singular/plural distinctions as in Italian *fratello/fratelli*, French *frère/frères*, and Spanish *hermano/hermanos*.

A theory currently gaining ground at the interface of historical linguistics, linguistic typology and sociolinguistics maintains that reduction in morphological complexity might be driven by learning pressures, namely *imperfect learning* by non-native adults (McWhorter, 2002; McWhorter, 2011; Trudgill, 2011; Wray and Grace, 2007). Adults learning a foreign language often lack the breadth of exposure to the target language that a native speaker would have. Thus, they only partially learn the range of inflectional variants of a word, and omit morphological markers in their language production (Papadopoulou et al., 2011; Haznedar, 2006; Gürel, 2000). If non-native speakers represent a considerable part of the overall speaker population, they might drive the language towards morphological simplification.

This line of reasoning was recently backed by quantitative analyses. Across different language families and areas it was shown that languages spoken by more people (a proxy for the proportion of non-native

speakers) tend to have lower morphological complexity (Lupyan and Dale, 2010), that languages with more non-native speakers have less complex morphological case marking (Bentz and Winter, 2013), and that languages with more non-native speakers have fewer word forms more generally (Bentz et al., 2015).

Our hypothesis is that imperfect language learning might also explain the loss of inflections from Classical Latin towards the Romance languages. These formed as the Roman empire expanded into the European continent, and later evolved into modern day Romance languages. In the process of expansion, Vulgar Latin varieties must have "recruited" considerable numbers of non-native speakers, which might have reduced the range of word forms in usage across the whole population of speakers (Herman, 2000; Bentz and Christiansen, 2010; Bentz and Christiansen, 2013). Over several generations, this mechanism can lead to considerable loss of morphological marking. We here present three lines of evidence to give such language change hypotheses an empirical and quantitative foundation.

1. A growing number of *diachronic and synchronic corpora* (see Cysouw & Wälchli (2007)) are available to measure patterns of change, rather than using single, isolated examples. Typological analyses based on corpora have the advantage of reflecting actual language production and usage, rather than expert judgement only. They are reproducible and transparent. In line with a range of earlier studies (Juola, 1998; Milin et al., 2009; Moscoso del Prado, 2011; Bentz et al., accepted; Ehret and Szmrecsanyi, 2016; Wälchli, 2012; Wälchli, 2014), we here apply corpus-based methods to measure morphological complexity.

2. NLP tools allow us to automatically and efficiently analyze large collections of texts. This is here illustrated with *lemmatization*, i.e. neutralization of inflected word forms to their base forms, also called *lemmas*. Thus we can tease apart the effect of inflections from other factors influencing the complexity of words.

3. Psycholinguistic experiments elicit the learning pressures that drive language change. So-called *iterated learning* experiments are particularly helpful to understand multiple factors shaping information encoding strategies in artificial languages (Kirby et al., 2008; Kirby et al., 2015). We here reanalyse data gathered in an artificial language learning experiment where inflectional marking is transmitted over several generations of "normal" and "imperfect" learners (Berdicevskis and Semenuks, forthcoming).

We would argue more generally that an integration of corpus, computational and experimental evidence is a valid strategy for understanding changes in any other set of languages and their phonological, morphological and syntactic features.

2   Methods

2.1   Corpora

To control for constant content across languages, we use two sets of parallel texts: 1) the *Parallel Bible Corpus* (PBC) (Mayer and Cysouw, 2014),[1] and 2) the *Universal Declaration of Human Rights* (UDHR) in unicode.[2] Details about the corpora can be seen in Table 1. The general advantage of the PBC is that it is bigger in terms of numbers of word tokens per language (ca. 280K), compared to the UDHR (ca. 1.8K). They represent two different registers: religious writing and transcribed speeches. This is important to ensure that the trends observed extrapolate to different text types. In our analyses, we focus on the Romance languages available in these corpora.

2.2   Estimating the Complexity of Words

We apply an information-theoretic measure of word complexity. Imagine a language that repeats the same word over and over again. This language is maximally redundant, each instance (token) of the same word (type) does essentially not store any information, hence this language has minimum word

---

[1]Last accessed on 09/03/2016

[2]http://unicode.org/udhr/

Parallel Corpus	Size	∅ Size	Texts	Lang.
PBC	$\approx 420M$	$\approx 280K$	1471	1083
UDHR	$\approx 650K$	$\approx 1.8K$	356	333

Table 1: Information on the parallel corpora used.

complexity. In contrast, a language with an infinite number of different words, expressing an infinite number of concepts, packs a lot of information into words, i.e. has maximum word complexity. Natural languages range in between these extremes (Bentz et al., 2015), displaying a variety of distributions of word tokens over word types. Such differences in type-token distributions can be measured by calculating their entropies. The classic Shannon entropy (Shannon and Weaver, 1949) is defined as

$$H = -K \sum_{i=1}^{r} p_i \, log_2(p_i). \tag{1}$$

Where K is a positive constant determining the unit of measurement (which is bits for K=1 and log to the base 2), r is the number of ranks (or different word types) in a word frequency distribution, and p_i is the probability of occurrence of a word of ith rank.

According to the maximum likelihood account, the probability p_i is simply the frequency of a type divided by the overall number of tokens in a text. However, it has been shown that this method underestimates the entropy, especially for small texts (Hausser and Strimmer, 2009). To estimate entropies reliably, the *James-Stein shrinkage* estimator (Hausser and Strimmer, 2009) is used here instead. According to this approach the estimated probability per type is

$$\hat{p}_i^{shrink} = \lambda \hat{p}_i^{target} + (1 - \lambda)\hat{p}_i^{ML}, \tag{2}$$

where $\lambda \in [0, 1]$ is the shrinkage intensity and $\hat{p}_i^{target}$ is the so-called "shrinkage target". Hausser & Strimmer (2009) suggest to use the maximum entropy distribution as a target, i.e. $\hat{p}_i^{target} = \frac{1}{V}$. This yields

$$\hat{p}_i^{shrink} = \frac{\lambda}{V} + (1 - \lambda)\hat{p}_i^{ML}. \tag{3}$$

The idea here is that the estimated probability $\hat{p}_i^{shrink}$ consists of two additive components, λ/V and $(1 - \lambda)\hat{p}_i^{ML}$ respectively. In the full shrinkage case ($\lambda = 1$), Equation 3 yields $\hat{p}_i^{shrink} = 1/V$, i.e. the maximum entropy. In the no shrinkage case ($\lambda = 0$), Equation 3 yields $\hat{p}_i^{shrink} = \hat{p}_i^{ML}$, i.e. the ML estimation that is biased towards low entropy. Given empirical data, the true probability is very likely to lie somewhere in between these two cases and hence $0 < \lambda < 1$. The optimal shrinkage can be found analytically. Finally, the probability p_i^{shrink} plugged into the original entropy equation yields

$$\hat{H}^{shrink} = -K \sum_{i=1}^{r} \hat{p}_i^{shrink} log_2(\hat{p}_i^{shrink}). \tag{4}$$

$\hat{H}^{shrink}$ is a robust approximation of the word entropy calculated from a text collection. It reflects the shape of type-token distributions. A long-tailed distribution will have higher $\hat{H}^{shrink}$, i.e. higher overall word complexity, while a short tailed distribution will have lower $\hat{H}^{shrink}$, i.e. lower overall word complexity. We use the R package *entropy* for shrinkage entropy estimations (Hausser and Strimmer, 2014).

Language	Tokens	unknown	%
Latin	11427	266	2.5
Italian	15314	888	5.8
French	17602	983	5.6
Spanish	15581	907	5.8

Table 2: Information on number and percentage of tokens unknown to the TreeTagger in a combined corpus of the PBC and UDHR. Note that only verses of the PBC which are parallel across several hundred languages were taken into account here. This explains the relatively low number of tokens.

2.3 Lemmatization and Inflectional Complexity

Note that the overall complexity of words is driven by a range of factors. Consider the example of the Latin lexeme *frater* "brother" again, which is frequently used in the Bible. As is illustrated in Figure 1, in Latin we get a whole range of inflectional variants, while in Italian, French and Spanish this range is reduced to the singular/plural forms. However, inflection is but one process besides derivation, compounding, contraction and others that shape type-token distributions (Bentz et al., accepted).
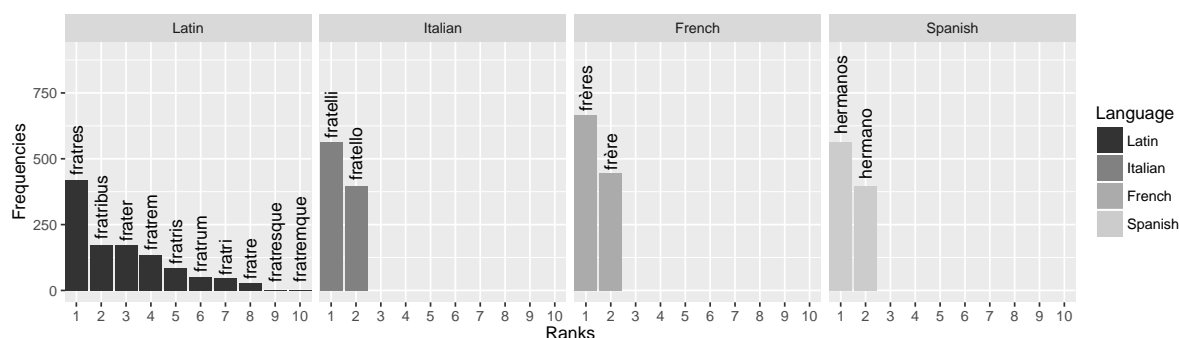
Figure 1: Reduction of inflectional variants from Latin to the Romance languages. Here exemplified with the occurrence of the word for "brother" in the Latin, Italian, French and Spanish texts of the PBC.

To tease apart *inflectional complexity* (C_{infl}) we can estimate the difference in entropy before (H_{raw}) and after lemmatization (H_{lem}):

$$C_{infl} = \Delta H = H_{raw} - H_{lem} \tag{5}$$

This requires automatic (or manual) neutralization of inflectional variants of the same word root in a corpus, i.e. lemmatization (see Moscoso del Prado (2011) for a similar account). In the following, this process is outlined for both the natural and artificial languages used here.

Natural Languages: Word types of some natural languages can be lemmatized using the *TreeTagger* (Schmid, 1994). It first associates the respective word type with a POS tag (and case indication if relevant), and then derives the most likely lemma, thus outputting *wordType/POS/lemma*. For example, for the Latin proper noun *fratrem* "brother.ACC.SG" the TreeTagger outputs: *fratrem/N:acc/frater*. Likewise, for an inflected verb such as *creavit* "create.3P.SG.PST" it outputs *creavit/V:IND/creo* (where *creo* "create.1P.SG.PRES" is taken as the default lemma instead of the infinitive *creare*).

The TreeTagger is based on a statistical model trained on samples of manually lemmatized text. It provides high accuracy on words already seen in the training set (close to 100%). The words that are unknown to the tagger result in higher error rates. Table 2 shows the percentage of unknown word types for each language. We use the PBC for lemmatization, since it is the bigger corpus in number of tokens. The Romance languages that can be lemmatized using the TreeTagger are: Latin, Italian, French, and Spanish.

Artificial Languages: The morphological structure of the artificial language used in our analyses is outlined in Section 2.4, and illustrated in Appendix 6.1. Note that descendant languages usually preserve much of this structure. We lemmatize the artificial languages by the following rules:

1. Only if two word forms occur at the same place in the utterance (i.e. first or second), can they be neutralized to the same lemma.

2. Moreover, word forms have to adhere to the following similarity criteria to be neutralized to the same lemma:

   - Words that occur at the first place are "nouns", they always denote an entity. For nouns, the similarity criterion is the normalized Levenshtein distance between the two forms, which has to be smaller than 0.50. In the initial languages, noun stems consisted of three letters, with a plural marker being a one-letter ending. However, more complicated systematic ways to mark number emerged in some descendant languages, which is why the threshold is 0.50 instead of 0.25.
   - Words that occur at the second place are "verbs", they always denote an event. For verbs, the first letter has to be the same. In the initial languages, verb forms always consisted of two letters, the first of which was a stem, the second an agreement marker. In the descendant languages, the stem letter is usually preserved, while the agreement marker can undergo various changes.

3. The shortest form of a given paradigm is chosen as the lemma. For example, the nouns *seg* and *segl* (PL.) are lemmatized to *seg*. If there are several forms of equal length, the most frequent one of these is chosen as the lemma. If there is a tie, the first form the algorithm comes across is chosen.[3]

2.4 Iterated Learning Experiments

The experimental data are taken from an *iterated learning experiment* (Berdicevskis and Semenuks, forthcoming). In the experiment, an artificial miniature-language called "epsilon" is learned and transmitted from one participant to the next in an online setting. 15 isomorphic variants of epsilon are created to be transmitted in 15 separate chains. "Isomorphic" here means that the grammatical structure of the variants is the same, but the vocabulary is different. Thus, phrases are built based on a selection of two nouns (e.g. *seg*, *fuv*) and three verb roots: e.g. *m-* "to fall apart", *r-* "to grow antlers", *b-* "to fly". Morphological features include number marking on nouns (e.g. SG: -∅, PL: -*l*) and agreement on verbs (e.g. agreement with *seg*: -*o*, agreement with *fuv*: -*i*). Phrases consisting of nouns and verbs have to be learned based on visual scenes of moving objects. For example, the phrase *segl bo* would be paired with a picture where several *seg* objects fly. Overall, this leads to 15 epsilon variants made up of 16 possible phrases matching 16 possible scenes (see Appendix 6.1).

Transmission chains consist of 10 generations (one participant per generation). There are 45 transmission chains, 15 for the *normal* learning condition (there are no imperfect learners in the population), 15 for what is called a *temporarily interrupted* condition (there are imperfect learners in generations 2-4) and 15 for a *permanently interrupted* condition (there are imperfect learners in generations 2-10). Note that the same 15 epsilon variants were used for all three conditions.

Imperfect (non-native) learning is simulated via less exposure. "Native" learners have six training blocks to learn the artificial language epsilon, whereas "non-native" imperfect learners have only half of the training blocks.

For further analyses, we collapse the 15 transmission chains per condition to one "corpus" consisting of the word tokens produced in each generation. This yields around 280 tokens (noun and verb forms) used in each generation of 10 learners. Finally, there were 450 participants (45 chains times 10 generations), all of them native speakers of Russian.

---

[3]Sometimes an utterance does not contain a verb, even though there is an event occurring on the stimulus image. We do not posit empty tokens for these cases, which means that there is some variation in the corpus size across languages, both for the lemmatized and non-lemmatized versions. If a participant produced one verb-less utterance, it means that their output language will lack one word form which it might have had if the participant chose to name every event.

3 Results

3.1 Corpus Analyses

First, we want to measure the exact difference in word complexities between Latin and the Romance languages. To this end, we use the shrinkage entropy estimation method (Section 2.2) applied to our parallel corpora. As is illustrated in Figure 2 for both the PBC and the UDHR, Modern Romance languages have systematically lower word entropies. While Latin has a word entropy of ~ 10.5 (PBC) and ~ 8.5 (UDHR) respectively, Modern Romance languages fall below 9.75 and 8.2. Note that for some languages we get different translations (e.g. 6 translations into Portuguese in the PBC), and hence a range of entropy values. This variation is indicated by 95% confidence intervals in Figure 2 and also Figure 3. Overall, these analyses illustrate that in the ca. 2000 years since Latin was last spoken as a native language, the word entropy of its daughter languages systematically declined by around 10-15%.[4]
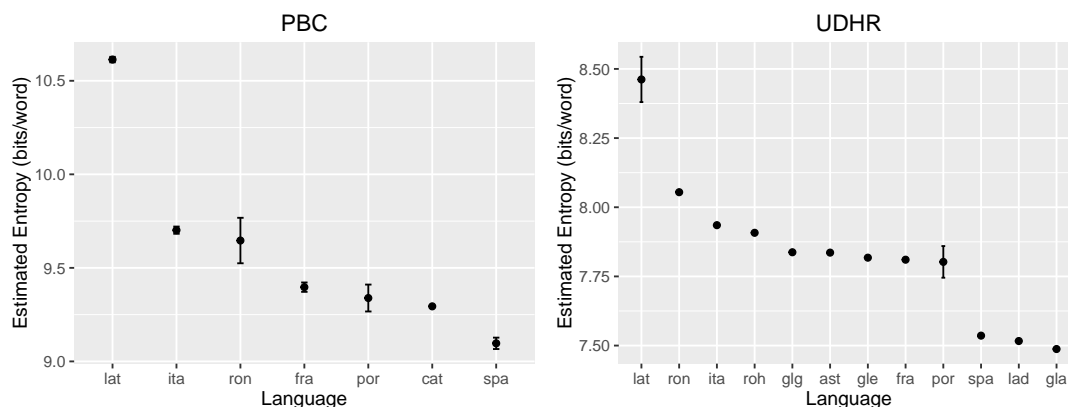
Figure 2: Reduction of word entropy from Latin to the Romance languages in two parallel corpora (PBC and UDHR). ISO 639-3 codes are given on the x-axis: Latin (lat), Italian (ita), Romanian (ron), French (fra), Portuguese (por), Catalan, (cat), Spanish (spa), Romansh (roh), Galician (glg), Asturian (ast), Ladino (lad).

3.2 Inflectional Complexity Reduction: Latin to Romance

Secondly, to pin down the reduction in inflectional complexity, we apply the lemmatization method outlined in Section 2.3. The results of this analysis for Latin, Italian, French and Spanish can be seen in Figure 3. Grey dots indicate the word complexities for the raw texts (H_{raw}), black dots indicate the word complexities after lemmatization (H_{lem}). The difference between these indicates the inflectional complexity of each language, or in other words, how much information is stored in inflectional variants of word roots. This is highest for Latin ($C_{infl} = \Delta H \sim 2.5$), and systematically lower for Italian, French and Spanish ($C_{infl} = \Delta H \sim 1.5$). Hence, the complexity of inflectional marking has systematically dropped in the 2000 years between Latin and its descendant languages. Namely, around 1 bit of information – formerly stored in inflectional marking – is now either lost or replaced by another level of encoding (Ehret and Szmrecsanyi, 2016; Koplenig et al., 2016; Moscoso del Prado, 2011).

3.3 Iterated Learning Experiments

The missing link to explain the entropy reduction in natural languages is the actual behaviour of language learners and users. Their impact on word entropy is illustrated here with data gathered from epsilon. Figure 4 gives an overview of the entropy change in the aggregated epsilon variants over 10 generations of transmission. The left panel shows the word entropy change in the three different learning conditions (normal, temporarily interrupted, permanently interrupted), while the right panel shows lemma entropy change (words neutralized for inflections).

---

[4]Note that there are two non-Romance languages in the UDHR sample: Irish Gaelic (gle) and Scottish Gaelic (gla).
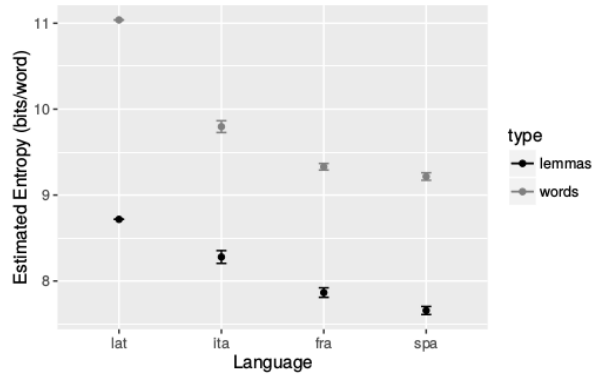
Figure 3: Differences in entropies before and after lemmatization (*words* and *lemmas*) in Latin, Italian, French and Spanish.

Focusing on the left panel first: in the normal condition (black), the entropy slightly decreases (from 7.23 to 7.15, i.e. ca. -1.1%) over 10 generations. For the temporarily interrupted condition (light grey), the word entropy decreases more sharply by 7.23 to 6.93 (-4.1%). In the permanently interrupted condition (dark grey), it also continuously drops from 7.23 to 6.95, i.e. by -3.9%. The right panel further illustrates – as we would expect – that entropy drops for lemmas compared to words, namely from 7.23 to 6.23, i.e. by 1 bit or -14% (in generation 0). However, for lemmas there is less of a systematic pattern in entropy change over 10 generations. In fact, for the normal and permanently interrupted conditions there is almost no change at all. Only for the temporarily interrupted condition does the lemma entropy drop somewhat from 6.23 to 6.09 (-2.2%).

In other words, the entropy drop in word forms over 10 generations of learning is mainly due to loss of morphological markers (e.g. losing plural marking *-l*, or agreement marking *-o* and *-i*), rather than a change in the base vocabulary (e.g. replacing vocabulary or using the noun *seg* where *fuv* should actually be used).
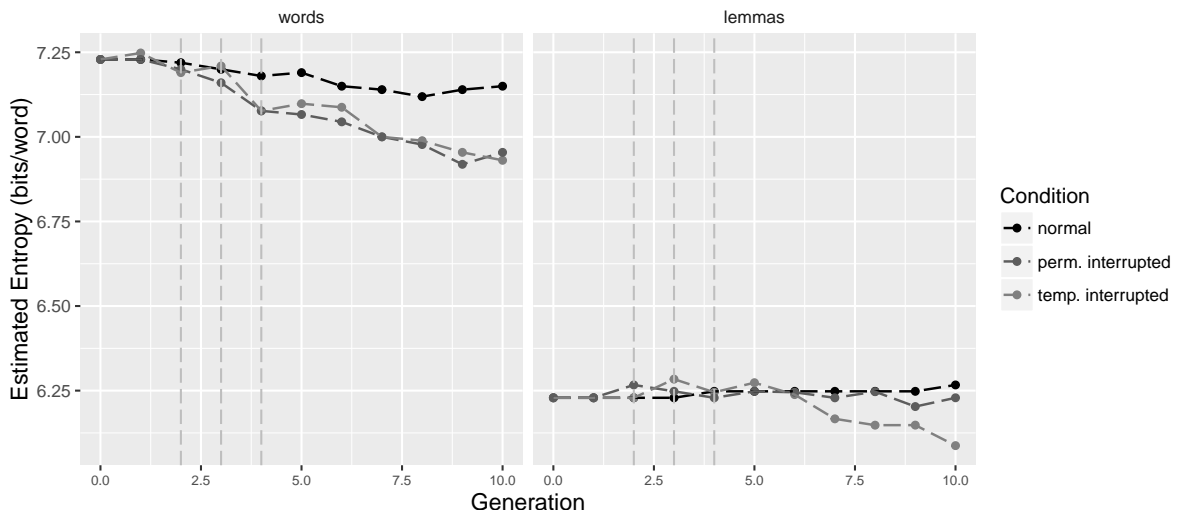
Figure 4: Entropy changes in the artificial language epsilon over 10 generations for both the original words produced by participants (left panel), and the lemmatized version (right panel). The languages were either transmitted via normal exposure (black), with temporary interruption (light grey), or with permanent interruption (dark grey). The vertical dashed lines indicate the generations of interruption in the "temporarily" interrupted condition.

4 Discussion

The word entropy/complexity of *all* Modern Romance languages represented in the PBC and UDHR parallel texts is lower than that of Classical Latin, which represents an earlier stage of the Romance genus. Namely, across 10 different descendants the word entropy is reduced by ca. 10-15%. Since this is a trend found in two independent parallel corpora, it is very unlikely due to effects of register or style of translation.

Instead, this pattern derives to a large extent from the loss of morphological marking witnessed in the ca. 2000 years since Romance languages evolved from Vulgar Latin. This is most clearly illustrated by means of lemmatization. Systematically neutralizing inflectional variants reduces the word entropy in Latin by around 2.5 bits, but in Spanish, French and Italian only by around 1.5 bits. Put differently, there is 1 bit less information stored in word forms of these three Modern Romance languages compared to Latin. Note that this information is not necessarily entirely lost, but potentially traded off for other means of information encoding beyond the word, e.g. word order (Ehret and Szmrecsanyi, 2016; Koplenig et al., 2016; Moscoso del Prado, 2011).

A potential caveat of our approach is that we infer grammatical structures used in spoken language production from analysing written texts. Classical Latin, as represented in the PBC and UDHR texts, is generally not considered the spoken proto-language of Modern Romance languages. Hence, a careful interpretation is that our analyses hold for written Latin and written Modern Romance. Written records are probably more conservative than spoken varieties, and reflect the spoken languages used at an earlier stage. Having said that, a mechanism for morphological loss in (written) language usage is starting to emerge from artificial language learning experiments. The data examined here illustrate that morphological marking can be learned and successfully transmitted under sufficient exposure. That is, the word type entropy of the artificial language epsilon was largely maintained in the "normal" condition of learning and transmission. However, when learning pressure is increased by reducing the exposure, imperfect learning effects kick in, and morphological distinctions are lost, which causes the word entropy to drop by around 4%, depending on the number of generations of imperfect learners.

These numbers seem relatively small, but over several centuries and millennia they can accumulate to considerable changes in the morphological structure of languages. Remarkably, the percentages of reduction from artificial languages also make sense – as an approximation – in the natural language context. If we assume generations of 30 years for a timespan of 2000 years, we arrive at ca. 66 generations. The word entropy reduction in epsilon (for the temporarily and permanently interrupted conditions) is approximately $\frac{0.3}{10} = 0.03$ bits per generation. If we multiply this by 66 we predict a reduction of ~ 2 bits, which is around -18%. This is close to – but somewhat higher than – the 10% to 15% word entropy reduction we actually find across 10 Modern Romance languages.

5 Conclusion

We have reported three lines of evidence – based on natural language corpora, NLP tools, and experimental data – to support the hypothesis that changes in morphological complexity from Latin to Romance languages were driven by imperfect learning scenarios. This suggests more generally that integrating corpora, computational tools, and experiments is a worthwhile strategy to model and explain complex scenarios of language change.

First, corpus data of historical and synchronic varieties are a source to help us observe general trends, rather than cherry-picking examples fitting our hypotheses. This was here illustrated by measuring the exact word entropy reduction between Latin and Modern Romance languages. Second, computational tools, such as entropy estimators and lemmatizers, allow us to quantify and further tease apart the effects in question. In our case study, we established that morphological loss is the main driver for entropy reduction. Third, psycholinguistic experiments elicit the potential learning pressures at play under different scenarios of language transmission. The systematic loss of morphological distinctions in the artificial language epsilon, driven by "non-natives", i.e. learners with less exposure, helps to understand the exact mechanisms of change. They might be subtle when looked at in isolation, but can have considerable effects when accumulated over time.

Thus, it is conceivable that elaborate corpus analyses in conjunction with iterated learning experiments (with several, separate points of interruption, and with artificial languages closely modelled after specific natural languages) could – for the first time – give us a model of language change "in real time".

Acknowledgements

References

Christian Bentz and Morten H Christiansen. 2010. Linguistic adaptation at work? The change of word order and case system from Latin to the Romance languages. In Thomas C. Scott-Phillips, Monica Tamariz, Erica A. Cartmill, and James R Hurford, editors, *The evolution of language. Proceedings of the 8th international conference (EVOLANG8)*, pages 26–33, Singapore. World Scientific.

Christian Bentz and Morten H Christiansen. 2013. Linguistic Adaptation: The trade-off between case marking and fixed word orders in Germanic and Romance languages. In *East Flows the Great River: Festschrift in Honor of Prof. William S-Y. Wang's 80th Birthday*, pages 46–58.

Christian Bentz and Bodo Winter. 2013. Languages with more second language speakers tend to lose nominal case. *Language Dynamics and Change*, 3:1–27.

Christian Bentz, Annemarie Verkerk, Douwe Kiela, Felix Hill, and Paula Buttery. 2015. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE*, 10(6):e0128254.

Christian Bentz, Tanja Samardžić, Dimitrios Alikaniotis, and Paula Buttery. accepted. Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. *Journal of Quantitative Linguistics*.

Aleksandrs Berdicevskis and Arturs Semenuks. forthcoming. Imperfect language learning eliminates morphological overspecification: experimental evidence.

Michael Cysouw and Bernhard Wälchli. 2007. Parallel texts. Using translational equivalents in linguistic typology. *Sprachtypologie & Universalienforschung STUF*, 60.2.

Katharina Ehret and Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler and Guido Seiler, editors, *Complexity and Isolation*. de Gruyter, Berlin.

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

Jean Hausser and Korbinian Strimmer. 2009. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research*, 10:1469–1484.

Jean Hausser and Korbinian Strimmer, 2014. *entropy: Estimation of Entropy, Mutual Information and Related Quantities*. R package version 1.2.1.

Belma Haznedar. 2006. Persistent problems with case morphology in l2 acquisition. *Interfaces in multilingualism: Acquisition and representation*, pages 179–206.

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

Patrick Juola. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Mueller-Spitzer. 2016. The statistical tradeoff between word order and word structure: large-scale evidence for the principle of least effort. *arXiv preprint arXiv:1608.03587*.

Gary Lupyan and Rick Dale. 2010. Language structure is partly determined by social structure. *PloS ONE*, 5(1):e8559, January.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 3158–3163. European Language Resources Association (ELRA).

John H McWhorter. 2002. What happened to English? *Diachronica*, 19(2):217–272.

John H McWhorter. 2011. *Linguistic simplicity and complexitiy: Why do languages undress?* Mouton de Gruyter, Boston.

Petar Milin, Victor Kuperman, Aleksandar Kostic, and R Harald Baayen. 2009. Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. *Analogy in grammar: Form and acquisition*, pages 214–252.

F Moscoso del Prado. 2011. The mirage of morphological complexity. In *Proc. of the 33rd Annual Conference of the Cognitive Science Society*, pages 3524–3529.

D. Papadopoulou, S. Varlokosta, V. Spyropoulos, H. Kaili, S. Prokou, and a. Revithiadou. 2011. Case morphology and word order in second language Turkish: Evidence from Greek learners. *Second Language Research*, 27(2):173–204, February.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, volume 12, pages 44–49.

Claude E. Shannon and Warren Weaver. 1949. *The mathematical theory of communication*. The University of Illinois Press, Urbana.

Peter Trudgill. 2011. *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford University Press, Oxford.

Bernhard Wälchli. 2012. Indirect measurement in morphological typology. In A Ender, A Leemann, and Bernhard Wälchli, editors, *Methods in contemporary linguistics*, pages 69–92. De Gruyter Mouton, Berlin.

Bernhard Wälchli. 2014. Algorithmic typology and going from known to similar unknown categories within and across languages. *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, 28:355.

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

6 Appendices

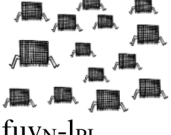## 6.1 Appendix A. Meaning space used in iterated learning experiment.

		event: none	event: fall apart	event: grow antlers	event: fly
agent: round animal	number: singular	seg_N	seg_N $mv\text{-}o_{AGR}$	seg_N $rv\text{-}o_{AGR}$	seg_N $bv\text{-}o_{AGR}$
	number: plural	$seg_N\text{-}l_{PL}$	$seg_N\text{-}l_{PL}$ $mv\text{-}o_{AGR}$	$seg_N\text{-}l_{PL}$ $rv\text{-}o_{AGR}$	$seg_N\text{-}l_{PL}$ $bv\text{-}o_{AGR}$
agent: square animal	number: singular	fuv_N	fuv_N $mv\text{-}i_{AGR}$	fuv_N $rv\text{-}i_{AGR}$	fuv_N $bv\text{-}i_{AGR}$
	number: plural	$fuv_N\text{-}l_{PL}$	$fuv_N\text{-}l_{PL}$ $mv\text{-}i_{AGR}$	$fuv_N\text{-}l_{PL}$ $rv\text{-}i_{AGR}$	$fuv_N\text{-}l_{PL}$ $bv\text{-}i_{AGR}$

Figure 5: The meaning space of one of the initial input languages with the corresponding signals. Subscript N denotes noun stems, V - verb stems, PL - plural marker, AGR - agreement marker. Morphemes are hyphenated for illustration purposes. Reproduced with permission from Berdicevskis and Semenuks (under revision).