Studies in Polish Linguistics
vol. 16 (2021), issue 3, pp. 145–162
doi:10.4467/23005920SPL.21.007.14261
www.ejournals.eu/SPL

Rafał L. Górski https://orcid.org/0000-0003-4727-2639

Institute of Polish Language, Polish Academy of Sciences

Dynamics of Language Change: The Case of Polish *barzo > bardzo*[*]

Abstract

The paper discusses the benefits and shortcomings of modelling a language change with logistic regression, an approach often called the Piotrowski-Altmann law. It is shown with an example of an isolated change, which occurred in Middle Polish, namely *barzo* > *bardzo.* The study is based on a historical corpus of Polish consisting of several hundreds of texts with over 12 million running words. Logistic regression based on the entire dataset shows relatively high goodness of fit, still there are some data points, especially close to the end of the process, which are quite far removed from the idealised trajectory. In the article, the author seeks to answer the question: to what extent the quality of the corpus affects the model. An experiment was conducted: a number of texts were randomly removed in order to create a smaller corpus, containing 90%, 75% and 50% of the texts of the entire set. Since such procedure is repeated 200 times, it is possible to compare the distribution of the scores indicating the goodness of fit of the model. It turns out that the smaller the corpus, the more diverse the goodness of fit, and in some rare cases it is even better than its counterpart for a larger corpus. Still the larger the corpus, the scores indicating goodness of fit tend to be higher.

Keywords

historical linguistics, language change, Middle Polish, corpus linguistics, Piotrowski's law, logistic regression

[*] This research was supported financially by a Poland's National Science Centre (NCN) grant (project number 2013/11/B/HS2/02795). I would like to express my gratitude to the anonymous SPL reviewers for their numerous insightful comments, which made me rethink and improve my argumentation. Also, I am grateful to Rafał Szeptyński for his most valuable hints pertaining to the phonetic mechanisms, which drove the change, tackled with here.

Abstrakt
W artykule omówiono korzyści płynące z modelowania zmiany językowej za pomocą regresji logistycznej, a także ograniczenia tej metody. Fakt, że zmiana taka powinna dać się opisać we wspomniany sposób, jest nazywany prawem Piotrowskiego-Altmanna. Ilustrujemy to przykładem izolowanej zmiany, jaka wystąpiła w języku średniopolskim, a mianowicie przejściem *barzo* > *bardzo*. Dane pozyskano z historycznego korpusu języka polskiego składającego się z kilkuset tekstów i liczącego około 12 milionów słów. Regresja logistyczna oparta na całym zbiorze danych wykazuje dobre dopasowanie, wciąż jednak istnieją pewne punkty, szczególnie pod koniec procesu, które są dość daleko od wyidealizowanej trajektorii. W artykule autor stara się odpowiedzieć na pytanie, w jakim stopniu jakość korpusu wpływa na model. W tym celu przeprowadzano eksperyment: z istniejącego korpusu usuwana jest losowo pewna liczba tekstów, tak aby stworzyć mniejsze korpusy zawierające 90%, 75% i 50% tekstów korpusu wyjściowego. Ponieważ taką procedurę powtarza się 200 razy, możliwe jest porównanie rozkładu wyników wskazujących na dopasowanie modelu. Wyniki wskazują, że im mniejszy korpus, tym większy rozrzut miary dobroci dopasowania, w skrajnych wypadkach nawet lepszy niż dla pełnego korpusu. Większe korpusy dają jednak na ogół lepsze wyniki dopasowania.

Słowa kluczowe
językoznawstwo historyczne, zmiana językowa, okres średniopolski, językoznawstwo korpusowe, prawo Piotrowskiego, regresja logistyczna

1. Introduction

The aim of this paper is twofold. First it is an attempt to model the dynamics of a certain isolated linguistic change in Polish, namely *barzo* > *bardzo* 'very'. This phenomenon, one of the minor diachronic processes in the Middle Polish period, is mentioned in historical grammars (e.g. Klemensiewicz 1965). If we revisit this change, it is not because it was overlooked in historical linguistics, but rather because we want to show how the use of machine-readable corpus and statistical techniques can deepen our understanding of the process. Górski et al. (2019) model this change among other changes, which occurred in the Middle Polish period and compare their dynamics. In this article we examine the course of this change in detail.

The second aim is to explore, with the example of the aforementioned language change, the extent to which the actual data can be idealised. In historical linguistics the availability of a certain text is a matter of chance. However, it is well known that the results are never better than the corpus itself. The older the epoch, the less documents have survived. A number of documents is unknown to researchers, even less are available in electronic format. Moreover, there is a certain bias – texts, which are appreciated for any reason, be it literary quality or historical importance, are more likely to make their way to a corpus. Thus we want to examine how the contents of the corpus affect the overall picture of a change. Or to be more precise: what is the extent (if any) to which the idealised model changes when an actual corpus is diminished by randomly removing a number of texts.

Obviously, a historical linguist has no access to the linguistic competence of a native speaker, which is an indispensable source of empirical data. Access to the previous stages of a language can be gained in two ways. First is the observation of the performance of native speakers who have lived in the past. However, written records cover a much shorter period than the linguists would like to explore. Insight to the pre-literary era can be gained by exploring the system and lexis of a language. The comparative method takes as a starting point the axiom that the linguistic sign is fully arbitrary, it can be concluded that systematic correspondences between words in several languages cannot be a matter of chance. The linguist seeks for words similar both in meaning and in form, in order to find regularities. This leads to observations such as "in language (dialect) A, sound X in a given context regularly corresponds to sound Y in language (or dialect) B." This observation, in turn, often allows for reconstruction of the past of these two languages. For example, if sound X corresponds to Y in a number of languages, it is more likely that that in the past of the language there was a change Y > X. Mutatis mutandis this holds for morphemes.

Another method of exploring the stages of language not attested by written sources is a method called internal reconstruction. Its greatest advantage over the comparative method outlined above is the fact it does not require comparisons between languages. Again, if one sign has several variants conditioned by context, it is probably a single form, which underwent a change. E.g. if we compare Latin *amicus* 'friend' with *inimicus* 'enemy', which is derived from the former by prefixation, we observe that /i/ in the latter word corresponds to /a/ in the former. With some further assumptions, which we are not tackling with here, we can draw a conclusion that here we observe a change /a/ > /i/ in the non-initial syllable.

Now, in both methods outlined there is an underlying assumption that not only regular similarities but also differences are not random, moreover, all such differences are caused by a regular historical process. Though these assumptions in general are supported by very strong evidence, they are not without issues.

In contrast, the philological method involves screening old texts representing former stages of language in order to examine elements which are subject to a change. It is not the abstract system or lexicon which is under scrutiny but rather texts, which can be viewed as performances of native speakers. This method by definition allows for exploring only those stages of language development which are attested by written testimonies, which is, of course, a serious limitation.

However, if we reconcile ourselves with the limitation of our research to the epochs which have produced written sources, the philological method shows several advantages. First is its credibility. Even if a single occurrence is a very dubious witness, one cannot deny the burden of proof of a series of attestations occurring in multiple texts.

Furthermore, let us focus on other features of the said method, which are of greater importance for our reasoning. First is the possibility of very precise dating. As long as the documents bear a certain date (be it a day, a year or a maybe a decade), a precise chronology can be established. In contrast, the comparative method as well as the method of internal reconstruction allow only for relative chronology (i.e. the process A follows the process B). The second feature is that the texts yield the researcher with quantitative data.[1]

Modern corpus linguistics can be regarded as an application of the philological method to synchronic studies. However, compared to traditional historical linguistics, the corpus methodology makes a much wider use of quantitative argumentation, often quite advanced in its nature. In a way we regard it as feedback: what has originated in historical linguistics, evolved in synchronic studies and reverted – with much more sophisticated tools – back to its place of origin, into diachrony.

A serious limitation for such advanced methods is imposed by the paucity of texts documenting older periods. Nonetheless, once we reach an epoch when the texts become more abundant, an extensive use of more advanced statistical techniques becomes possible. As for Polish, the textual testimonies for the so-called Old Polish period, which is dated by some linguists up to 1500 or (which is a wider opinion) 1543, are very scarce both in terms of the number of texts as well as running words.[2] Moreover, in the literary legacy of this period there is virtually nothing but religious and legal writing. After mid-16th the abundance of written material enables reliable quantitative studies.

It is worth mentioning that due to the abundance of data on the one hand, and extensive use of statistical methods on the other, the interest of historical linguists shifted to the more recent epochs, which in the "pre-electronic era" had seemed too similar to the modern stages of language to deserve attention. Though qualitatively the changes may be scarce, quantitatively they are more serious than it might seem at the first glance. A good example of such a study of Polish is Derwojedowa et al. (2016).

[1] We gloss over two other very important kinds of data which are provided only by texts, namely the context which licenses the phenomenon and the sociolinguistic determinates of text. These are, however, of no importance to this study.

[2] The Old Polish Corpus consists of 17 texts with ca 500 000 running words. The corpus covers all known continuous texts up to 1500.

Now, as we have already stated above, the combination of two kinds of data provided by a corpus (that is the precise chronology as well as frequency of occurrences of certain forms) allows for a study of the dynamics of the diachronic process, which is under the scrutiny of this paper.

2. Modelling a change

A language change can follow two different scenarios. The first scenario anticipates an emergence of a new, previously not existing entity; a good example is the phoneme /f/ (absent in pre-literary era) emerged in Old Polish or the *going to* future in English. Here, the innovation parasites on the old system, e.g. each use of *going to* in a text diminishes the frequency of other markers of future tense, but does not replace them entirely; moreover the two forms coexist peacefully, since they are not totally synonymous and the user of the language makes use of both of them. In the other scenario the innovation cannibalises its recessive counterpart, that is, finally the innovation completely displaces the older form.

Now, a language change – as described in a handbook of historical grammar – seems to be a phenomenon which happens in a moment, say a phoneme in a particular context is replaced by another phoneme. However, common sense tells us that replacing one phoneme, form, or construction by another one must be a gradual process. It starts within a small, probably geographically and socially restricted community, but (in a way somewhat similar to an epidemic) the speakers who are exposed to the innovation start to replicate it in their speech. The more people adopt the change, the greater the chance of exposure to the innovation for those still adhering to the recessive form.

In the second of the above-outlined scenarios, in mathematical terms, the probability *(p)* of finding an innovative form and a recessive form (denoted as *i* and *r*, respectively) in the corpus is

$$p(i) = 1-p(r),$$

which also implies that

$$p(r) = 1-p(i).$$

Consequently, the joint frequency of the two forms might remain constant over the centuries, while their mutual proportions usually vary to a significant degree.

Such a change is a perfect example of a phenomenon which can be captured by logistic regression. This statistical technique is used to model quite a number of phenomena from such diverse fields as demography, epidemiology, and environmental biology.

Gabriel Altmann, a Slovak mathematician with linguistic interests proposed a formula which defines a curve describing the diachronic process (Altmann 1983). This formula can be interpreted as a variant of logistic regression. We are not going to dive into the mathematical details, however we should underline that the formula contains arbitrary parameters which should be adjusted, so as to make the trajectory of the curve as close to the empirical data as possible. There are mathematical means allowing for such a perfect adjustment. For this study, we have used a function available in the statistical programme R. Also, all the plots below are produced by this programme. This curve should be interpreted as a probability of encountering an innovation (or reversely the recessive form) in the texts produced in a given moment. Note that the curve resembles an elongated letter "s", therefore it is often referred to as an "s-curve".

As already stated, the curve is an idealisation – in practice its trajectory always departs to a larger or lesser extent from the data provided by the corpus. Again, there are means to estimate the goodness of fit. We use Nagelkerke's R^2 measure, which has the advantage of being easily interpreted: 0 should be interpreted as a total lack of adequacy, whereas 1 means a perfect agreement of the empirical data and the model.

The first to observe that the course of a diachronic process resembles an *s*-shape curve was a Russian linguist Raimund Piotrowskij, therefore this kind of modelling is often called after him the Piotrowski's law, or due to Altmann's modifications Piotrowski-Altmann law. Originally, Piotrowski's works focused on the history of Russian word-formation and French articles. However this model was probably most widely used in the study of chronology of the percolation of lexical borrowings (to name but a few of a large volume of studies: Best 2013; Stachowski 2016; Gnatchuk 2015). The bulk of loanwords tend to percolate from donor to the host language for a certain, restricted time span. At the beginning of this process, the number of loanwords is limited, however – with the raise of the cultural attractiveness or intensity of contacts with the donor language – it more and more rapidly increases, only to gradually slow down when the donor is no more appealing to the community of the host language. While the language change and growth of lexis are very distinct processes, they can be well described by the same mathematical formula.[3]

[3] A survey of applications of the so-called Piotrowski's law is provided in Leopold (2005), Best (2016) and Stachowski (2020).

3. The corpus

The study is based on a diachronic corpus of Polish of ca 12 million running words,[4] which covers a period between 1380 and 1850 (Górski et al. 2019). This corpus was compiled for a larger study on the dynamics of changes in Middle and early Modern Polish. However, since first traces of the innovative form are found in the 16th century and last attestation of the recessive form in late 18th century, in this study we did not make use of the Old Polish and early Modern Polish data. Due to a limited amount of data, the corpus had to be opportunistic, meaning that neither the temporal coverage nor balance of genres was reliably controlled for. Obviously, in the case of early historical data the number of available texts is by all means limited, therefore it is virtually impossible to guarantee the balance of such a corpus. As always in historical linguistics the older the time period to be covered by a corpus, the bigger the expected bias. Still we assume that for our purposes, the bias will not distort the results. Since our study is aimed at examining the proportion between two forms – the recessive and the innovative one – and not in the actual word occurrences over time, uneven temporal coverage of the corpus should not affect the relative proportions between respective words to a significant extent.

A more serious issue is an uneven representation of certain text genres in particular epochs, e.g. one should keep in mind the overrepresentation of religious treatises in the late Middle Ages, as well as the overrepresentation of *belles lettres* in the 19th century. Regardless of several reasons such as bias in any diachronic corpus, we hardly believe it can ever be reliably corrected for. However, since different genres do not affect morphology to the same extent as they affect lexis, we assume that the grammatical change under our scrutiny will be reliably reflected in our corpus anyway.

The texts collected in the corpus might be, and usually are, of different size. We did not attempt trimming long texts, though. Firstly, it is widely agreed upon in corpus linguistics that the texts should not be sampled but rather included as a whole, since each part of a text has its own peculiarities. What is more important however, with the scarcity of historical data, it would be imprudent to let large amounts of already-acquired data be wasted. At the same time, the cost we have to accept when taking entire texts rather than sampling cannot be neglected. Namely, as we will see, a long text which is more conservative (or more progressive) than its contemporaries can skew the results by shifting the curve away from the general tendency.

[4] A large part of this corpus contains same texts as the Baroque Corpus (KorBa, cf. Gruszczyński et al. 2020). We would like to thank the team of the KorBa project for making their resources available for our purposes before the official launch of the corpus.

On the other hand, even a long text may contain a restricted number of the forms we are interested in.

The first step which is to be undertaken when a language change is modelled by logistic regression is to split the data into "slices," or particular time windows in which the proportion of the recessive and innovative forms is being measured. In fact, this means that the input corpus has to be divided into a series of chronologically ordered subcorpora. Depending on the corpus, a natural time window size could range from one day (in the case of press corpora) to, say, one century. On the one hand we face data scarcity, on the other hand the goal is to obtain as many data points as possible while keeping the noise at a moderate level. We deal with two mutually exclusive needs here: we want the corpus to be as fine-grained as possible, and at the same time we want the subcorpora to be as large as possible in order to provide the data as credible as possible. Dividing the corpus into one-century blocks would give very few yet reliable data points, whereas using one-year chunks would result in hundreds of data points (quite a number of them actually with no data!), yet affected by noise.

The division into chronologically ordered subcorpora is a delicate question, since the choice has to be made arbitrarily. Moreover, assume that one divides a corpus into subcorpora of 20 years. Now assume there are three texts, say written in 1602, 1618 and 1622. The first two texts fall into one "slice," the third to another, even if the first two are separated from each other by 16 years, whereas the second and the third one by 4. Note that the larger the time spans covered by the subcorpora, the bigger the unwanted effect.

In order to avoid the abovementioned issues, we have involved a "moving window" procedure, in which the subsequent "slices" were excerpted with an overlap. Not only does it allow for more data points, but it also diminishes the effect of Procrustean bed of setting arbitrary borderlines between subcorpora. In the aforementioned example, the text from 1618 would still fall into the subcorpus with the text from 1602, but additionally into the second subcorpus together with the text from 1622. The advantages of the "moving window" procedure by far surpasses its downsides, which include the fact that a single outlier affects more than one "slice," and thus more than one data point. Below, we present the results obtained for 20-years' windows, with a 10-year overlap contrasted with non-overlapping subcorpora of 10 and 20 years.

An important caveat is in place here. We inevitably fall into the common pitfall of the philological method in historical linguistics: since the only available material are written attestations, we trace in fact changes in orthography rather than observing them directly in actual sounds (cf. Campbell 1998: 333). One cannot deny, however, that orthography has always been following phonetics, even if we have no clear hint how close the relation has been. For this reason, we believe that the changes observed in orthography do reflect, to a significant extent, actual language phenomena. Moreover, since spelling is more conservative than pronunciation, we assume that the change of spelling can be treated as a *terminus ante quem.* We have to rely on philologists: we have to believe the editor as to the faithfulness of the electronic text.

This said, we should underline that the orthography is much more diverse than *barzo/bardzo.* The graphical variants include *bárdzo, barziej, barziey, barźiey, bárzieij, bárziey* as well as *bárźiey* etc. also the markers of superlative are *na-* and *naj-*, *nay-*, *náy-*. It is worth mentioning that the KorBa corpus notes altogether 88 variants of spelling (including the marker of superlative). All the variants were taken into account in the queries.

Now, Osiewicz (2015) suggested that the variation in spelling of *albo* and *abo* might be caused by typesetting – a letter was added or removed in order to extend or shorten a line in print. Still – both variants did exist in language, however the choice of one of them was dictated not only by the idiolect of the author but also by technicalities of the print. We cannot exclude that to some extent the same phenomenon played a role in the choice between *bardzo* and *barzo*, especially if we take into account that there is quite a number of texts where both forms occur.

4. The change *barzo > bardzo*

What we are dealing with is an isolated change, which can phonologically be described as the change of voiced spirant /z/ into affricate /d͡z/, which is reflected in the spelling <barzo> and <bardzo> (and their abovementioned variants) respectively. It is not a *lautgesetz,* but rather an isolated phenomenon restricted to this very lexeme, with a very small amount of parallels among other words.

The mechanism of this change remains unclear and – to our best knowledge – it still awaits a good explanation. Łoś (1922: 148) calls it "a spontaneous change."[5] Boryś (2005) suggests that the change is caused by dissimilation. There is no doubt that the sequence of a trill and a voiced alveolar spirant (i.e. /rz/) in Polish is very rare. Rafał Szeptyński (personal communication) in turn specifies that the affricate helped to avoid a change /barzo/ > /baʐo/ (a form attested in dialects, cf. Leszczyński 1978). The speakers resisted the change to /baʐo/ either in order to keep the phonetic shape of the word closer to the initial form or because /baʐo/ was less prestigious.

[5] "Poza tem mamy jeszcze *ʒ*, które powstało spontanicznie z dawniejszego *z*" (Apart from it, we have also ʒ, which emerged from older *z* (where /ʒ/stands for /dz/, RG)).

The corpus attests 9553 occurrences of the recessive form and 3793 of the innovative one. There are altogether 22 occurrences of the innovative *bardzo* before 1600, compared to 2128 attestations of *barzo* in the same period. Since all attestations of the innovation come from texts where the recessive forms prevail, it might be the case that – at least some of them – are mistakes of the editor. We encounter the earliest attestation in *Rozmyślanie przemyskie* (ante 1510); nevertheless if we compare this single occurrence with 240 occurrences of *barzo* in this very text, we must not pay any particular attention to this finding. Similar is true for chronologically consequent texts, namely *Rozmowy, które miał król Salomon mądry z Marchołtem grubym a sprośnym* by Jan of Koszyczki, *Żołtarz Dawida proroka*, a translation of the Psalter by Walenty Wróbel. Note a large "hillock" around 1650. It is caused by Jan Chryzostom Pasek, author of bulky memoirs. With 246 occurrences he is responsible for a 94% share of innovative forms in the subcorpus 1650–1660. Should we remove Pasek's text from the corpus, the innovative form would make no more than 15% of all uses of the word in question. Of course we do not want to manipulate the data, but rather to show to what extent one author who differs from his contemporaries can affect the results. Similarly, Benedykt Chmielowski's encyclopaedia *Nowe Ateny* covers some 2/3 of all attestations of the innovative form in the years 1740–1760. They distort the otherwise rather neat overall picture.

This number of innovative forms gradually raises in the course of the 17th century, around 1700 they are approximately as numerous as the recessive ones, and by the end of the 18th century *bardzo* definitely replaces *barzo.* The last attestation of the recessive form, which we traced is in *Monitor na Rok Pański 1772* by Ignacy Krasicki.

The two forms coexisted in the output of speakers of Polish over two centuries, often even in a single text. Let us repeat: it lasted over two hundred years to replace the recessive form with the innovative one.

Having all this in mind, let us have a birds-eye look at Figure 1. Until 1600, the number of innovative forms is close to 0. It can be easily noticed that at the very beginning the innovative form is rare, but its proportion in the overall number of the occurrences of the lexeme is gradually raising. It is not the case that each subsequent subcorpus would show a higher proportion of the *bardzo* compared to *barzo.* On the contrary, several subcorpora show a lower proportion of the innovative form, than those which represent an immediately preceding period. However, the general pattern revealed by the data points is clear: the old form is gradually replaced by the new one.

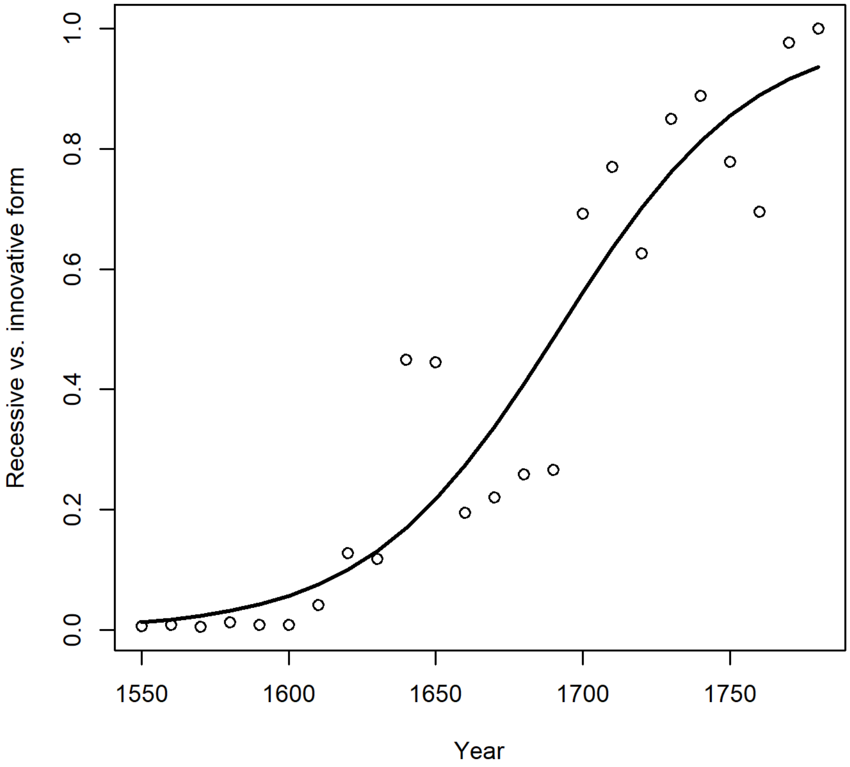

Fig. 1. The course of change of *barzo* > *bardzo*

Still, can we generalise over the actual data? That is, can we gloss over local raises and drops in order to discover a more general path of the change? Logistic regression allows for drawing a smooth curve, which is in fact such a generalisation. Of course this line sometimes departs from actual data, nevertheless this should not hold us off from making use of this statistical technique. On the contrary, as already mentioned, we do want a generalisation over actual data. Since the goodness of fit for this very model is high there is no reason to reject this idealisation. Nevertheless, we still should bear in mind that we are idealizing upon the figures provided by the corpus. Needless to say, the model reflects the actual change within the linguistic community only to the extent that the corpus is its good representation.

Now, let us compare the "moving window" approach with the standard division of the corpus into 20 non-overlapping subcorpora. Note that in the latter case each subcorpus covers 10 rather than 20 years.

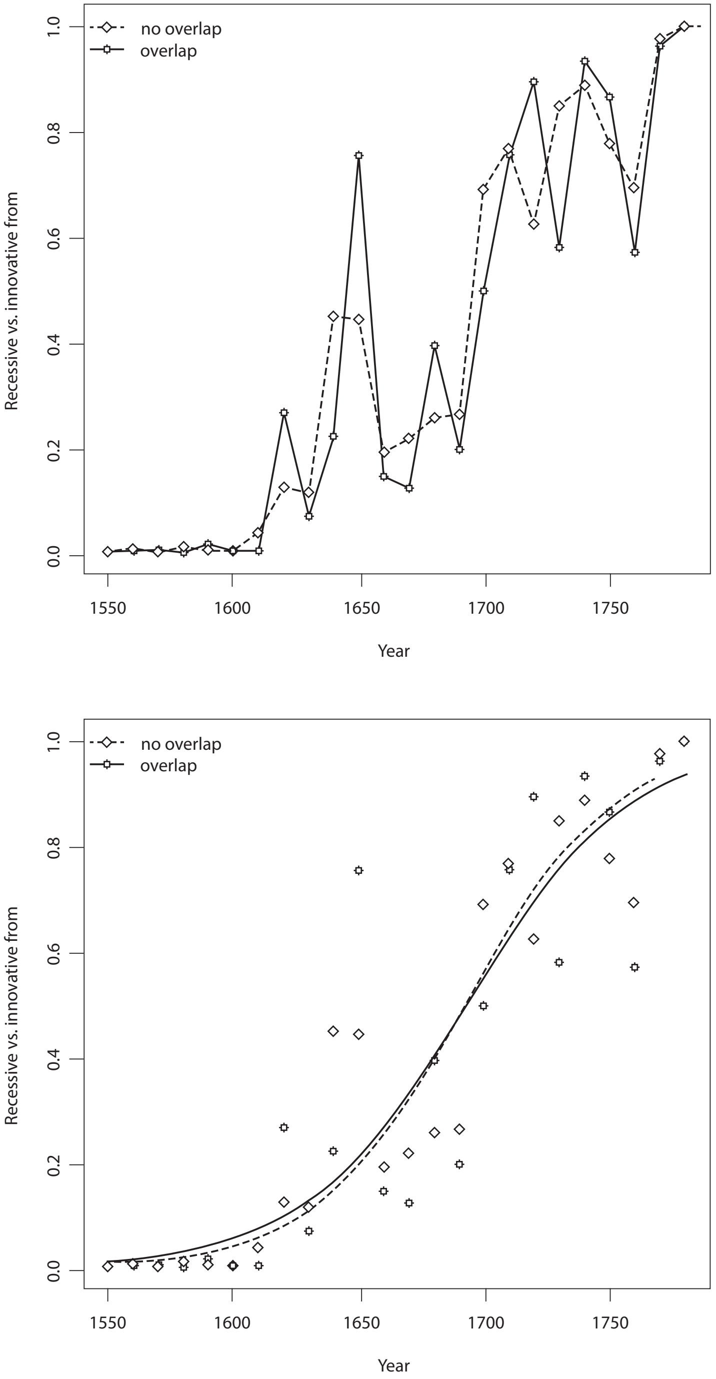

Fig. 2. The course of change of *barzo* > *bardzo*, with 20 non-overlapping and overlapping subcorpora

Although in case of non-overlapping subcorpora the points are much more scattered across the chart, the curves based on these two datasets coincide almost perfectly. Nontheless the goodness of fit for non-overlapping subcorpora drops from 0.922 (in case of the moving window) to 0.822, which is still a very good fit.[6]

5. The change in the light of various corpora

Let us underline once again: logistic regression is an idealisation over actual data. We expect it to capture a whole picture rather than some of its peculiarities, that is local "revolutions" or "counter-revolutions" – writers (especially prolific ones), who go against the tide and are more conservative (or innovative) than their contemporaries. However, it is naïve to expect, that any statistical technique will cover the shortcomings of the data themselves. On the contrary, if we had an ideal corpus, consisting of the entire book production of the past, the trajectory of the curve would certainly be different and – there is no doubt – closer to the actual data.

Whereas the availability of texts limits any corpus and a historical corpus can consist only of books which survived throughout the centuries, still we can artificially compile a worse, that is less complete, corpus. As already said, it is a matter of chance that one or another text is part of our empirical base. This sheer chance can be simulated, namely we can randomly remove a number of texts from our collection.

How does such a smaller and probably less representative empirical basis affect the results? In order to answer this question we have conducted an experiment. We randomly removed one tenth, a quarter, and a half of the texts, thus we obtained a corpus which contains 90%, 75%, and 50% of the texts of the entire collection, or in raw numbers 276, 414 and 497 items. Since we want to estimate this impact in a more systematic way, each procedure was repeated 200 times, thus we "compiled" 600 different corpora. The texts were removed in a purely random way, that is we did not control neither for the chronological coverage, nor for its size. It is quite possible, that this 50% of texts make much more than half of the corpus in terms of running words.

It is difficult to visually compare 600 curves. However, apart from the shape of the s-curve, there is one more factor which is of interest to us, namely the goodness of fit. We can ask the question, whether – regardless

[6] This figure is lower than quoted in Górski et al. (2019), because we take into account only texts dated between 1500 and 1800, whereas the cited book the change is examined vis-à-vis the entire corpus. In this case a larger number of data points representing recessive form or innovative form exclusively increases the goodness of fit.

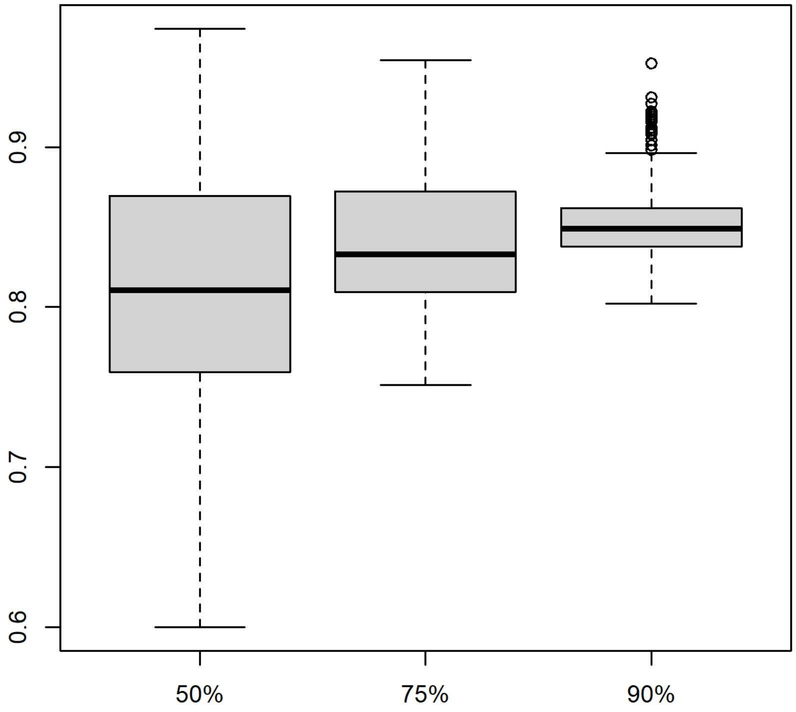

Fig. 3. The distribution of Nagelkerke's R^2 for corpora consisting of 50%, 75% and 90% of the texts of the entire corpus

of the trajectory – a smaller corpus tends to yield a more dramatic discrepancy between the actual and the idealised data. Recall, that we estimate this discrepancy by the Nagelkerke's R^2 measure. The closer it is to 0, the larger is this discrepancy. And vice versa, the closer to 1, the better the idealised curve represents the empirical data points. Since we have 600 different corpora, we can compare the distribution of this score for each "corpus size."

The boxplots in Figure 3 show the distribution of the R^2 score for the corpora of a given size. Not surprisingly, the range of values for the corpus where half of the texts were removed is the largest. The two other boxes show much smaller interquartile range, i.e. difference between the second and third quartiles (note, that by definition 50% of the observations lay between these values). This is quite obvious, since corpora "compiled" by removing a quarter of texts are more similar one to another than those where only every second text found its way to the corpus. This is even more true in the case of corpora where only every tenth text was discarded. No wonder – we draw a number of texts form a fixed set; now, when we draw every second text we compile corpora which differ one from another to a much larger extent, than the corpora compiled by drawing 90% of items. Simply, the in the latter case the variation between the corpora was much smaller compared to the former, therefore the variation of the goodness of fit is also less diverse.

Much more important is that most of the scores for the "three-quarters" corpus lay above the median for the halved corpus. This is also true, when we compare the other two datasets. In other words, in case of a larger corpus the chance that the idealised curve resembles the actual data to a little extent is much smaller, than it is in a corpus of a limited size.

However, this simple experiment shows that it is not always the case that accuracy increases with the size of corpus. Though the experiment shows that this situation is unlikely, still it is possible that with more texts added, the goodness of fit drops. Moreover, the smallest corpus yields not only a poor (0.6, what is really low), but also the best performance (0.97). A certain configuration of texts gives a fit much better than in case of a large corpus. This is because the removed texts are those, which swim against the tide. Finally, one should keep in mind that the minimal corpus contains 276 texts, what is already a considerable size, still, its reliability is rather low.

6. Conclusion

The course of change, which we are tackling with in this paper is a good example of a diachronic process which can be modelled via logistic regression. The rather high goodness of fit assures us that the idealisation follows the actual data quite well. This is not always the case, in several languages including Polish a number of changes hardly fit to this model (cf. Górski et al. 2019).

Still, is there any "linguistic added value," apart from a neat mathematical model? What more can a historical linguist learn from it? Or better, what more can it tell us about diachronic processes? And above all – why should we quantify a language change anyway?

First, the figures are always an interesting comment to any linguistic phenomenon. A change is driven both by language-internal and social factors. Both – as far as possible – call for explanation. Now, the latter (maybe with the exception of language contact) are accessible only via philological method. One of the social factors, which should be described is the resistance of the language community to a change, or to put it more precisely – how long did it take to fully accept the innovative form. This can be estimated by measuring the interval between the first attestation of the innovative form and the last occurrence of the recessive one. However, such a simplified approach can be very misleading if only because late attestations are often intentional archaisms, while first attestations in our dataset are so rare, that they do not allow to draw any certain conclusions. Moreover – what we are interested in here, is the behaviour of the entire linguistic community, not of individuals. Therefore we are not so much interested in single attestations, as in proportions of the utterances containing the old and the new form produced by a number of authors in a certain time-span.

Moreover, it seems that the most important benefit of the approach which we have taken is that the model is able to pick up regularities at a higher level of generalization. The discrepancies between the s-curve and the actual data can be regarded as a factor undermining the descriptive power of the model. However, instead of focusing attention on particular local proportions of the innovation, we see the process as whole. Though we do not observe a steady, incessant growth of the share of the innovative form in each subsequent corpus, this does not undermine the value of the model. Recall that each point on the s-shape curve is the probability of encountering the innovative (or reversely the recessive) form. Now, this probability is based on the entire data set, not only on a local value. This is especially true for the data points heavily affected by a single prolific author, who diverges from his contemporaries, such as Pasek or Chmielowski.

In the very process, which we were tackling with the goodness of fit is relatively high, but what if the actual shape heavily departs from the empirical data? In our example the actual data are relatively close to the modelled curve, even if some data points are quite far from the idealisation. But what if the model has little to do with the empirical data? Is this model useless in such a case? Before we answer this question, let us consider the reasons for such a potential discrepancy. The first answer that comes to mind is to blame the corpus. In fact – as shown in the experiment – in case of smaller corpora the chance that the model poorly fits the data is large. Should a larger corpus be available, the goodness of fit would have been better. Even if a larger corpus contains more authors who stand out from their times, adhere to a dialect rather than the standard language etc., all these peculiarities cancel each other out. And indeed, the experiment shows that the larger the corpus, the goodness of fit tends to be better. What is less obvious, high goodness of fit may be caused not by a very representative resource, but rather by its underrepresentativeness, i.e. the fact that those authors, which are more (or less) conservative then their contemporaries, are not represented in the corpus.

It is also possible that there are some external factors which distort the ideal course of change. To name but a few: some dialects or genres adhere rather to one of the forms. When in a particular time-span such a variety is overrepresented in writing, it is not without effect on the overall picture.

This suspicion is particularly justified when the data pertaining another change, gathered from the very same corpus yield a much higher goodness of fit. In any case poor goodness of fit calls for explanation. There are numerous factors which can speed up or slow down the process. And this may be another benefit for a historical linguist. The model allows one to quickly estimate whether the process as whole requires further investigation or certain data points, which depart from it to a larger extent than the others, require closer inspection. Finally, there is one more reason for such a kind of modelling a language change, namely it provides a good means of visualisation of the dynamics of a diachronic process, especially when several of them are compared.

References

Altmann Gabriel (1983). Das Piotrowski-Gesetz und Seine Verallgemeinerungen. In Best, Kohlhase (eds.) 54–90.

Best Karl-Heinz, Kohlhase Jörg (eds.) (1983). *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte*, Göttingen: Edition Herodot.

Best Karl-Heinz (1983). Zum Morphologischen Wandel Einiger Deutscher Verben. In Best, Kohlhase (eds.) 107–118.

Best Karl-Heinz (2013). Iranismen im Deutschen. *Glottometrics* 26, 1–8.

Boryś Wiesław (2005). *Słownik etymologiczny języka polskiego.* Kraków: Wydawnictwo Literackie.

Best Karl-Heinz (2016). Bibliography – Piotrowski's law. *Glottotheory* 7(1), 89–93.

Campbell Lyle (1998). *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Derwojedowa Magdalena, Kieraś Witold, Bilińska Joanna, Kwiecień Monika (2016). Dynamika zmian fleksyjnych i ortograficznych między reformami 1830–1918. *Język Polski* XCVI(1), 24–35.

Górski Rafał L., Król Magdalena, Eder Maciej (2019). *Zmiana w języku. Studia kwantytatywno-korpusowe.* Kraków: IJP PAN.

Gnatchuk Hanna (2015). Anglicisms in the Austrian Newspaper 'Kleine Zeitung.' *Glottometrics* 31, 38–49.

Gruszczyński Włodzimierz, Adamiec Dorota, Bronikowska Renata, Wieczorek Aleksandra (2020). Elektroniczny korpus tekstów polskich z XVII i XVIII w. – problemy teoretyczne i warsztatowe. *Poradnik Językowy* 8, 32–51.

Klemensiewicz Zenon (1965). *Historia języka polskiego*. Warszawa: PWN.

Leopold Edda (2005). Das Piotrowski-Gesetz. In *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch/An International Handbook*, Reinhard Köhler, Gabriel Altmann, Rajmund Piotrowski (eds.), 627–633. Berlin–New York: de Gruyter.

Leszczyński Zenon (1978). *Kierunki zmian w grupach spółgłoskowych typu* Sr *oraz* rS *w świetle geografii językowej*. Wrocław: Zakład Narodowy im. Ossolińskich.

Łoś Jan (1922). *Gramatyka polska*, part 1: *Głosownia historyczna*. Lwów: Wydawnictwo Zakładu Narodowego im. Ossolińskich.

Osiewicz Marek (2015). O możliwości technicznego uwarunkowania oboczności *albo*//*abo* w drukach polskich z XVI wieku. *Poznańskie Studia Polonistyczne. Seria Językoznawcza* 22(1), 185‑202.

Stachowski Kamil (2016). German loanwords in Polish and remarks on the Piotrowski-Altmann Law. In *Issues in Quantitative Linguistics 4*, Emmerich Kelih, Róisín Knight, Ján Mačutek, Andrew Wilson (eds.), 237–259. Lüdenscheid: RAM-Verlag.

Stachowski Kamil (2020). Piotrowski-Altmann Law: State of the art. *Glottotheory* 11(1), 3–14.

Rafał L. Górski
Institute of Polish Language, Polish Academy of Sciences
Al. Adama Mickiewicza 31
31-120 Kraków
[rafal.gorski(at)ijp.pan.pl]