Machine Translation between Language Stages:
Extracting historical grammar from a parallel diachronic corpus of Polish

Amir Zeldes
Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
az-omega@013.net

Abstract

This paper explores methods for the extrapolation of correspondences in a small parallel diachronic corpus taken from the Modern and Middle Polish Bible, in an attempt to answer the question "can historical grammar and lexica be derived directly from a corpus?" The problem of extracting this data is approached from a machine translation point of view: by envisioning texts from different periods as language models for their respective language stages, and historical grammar as a translation model mapping one language stage onto another. This notion is explored using automatic extraction of morphological, lexical and syntactic correspondences.

1 Introduction

Research in historical linguistics is more limited to written data than in other linguistic disciplines: we simply have no sources except for texts. The main presupposition in corpus linguistics, that conclusions can be drawn from sample data that apply to the state of affairs in the abstract language, is thus already made. This data is however usually subjected to a great deal of interpretation, especially since it forms only a selective and in essence accidentally preserved cross-section of a language stage which is not representative of an entire language. As Labov (1994: 11) puts it: "Historical Linguistics can then be thought of as the art of making the best use of bad data". But given that historical research is corpus-based, what insights can we get directly from the corpus with the least amount of interpretation? In this paper I will attempt to use correspondences in a diachronic corpus to automatically extract and quantify meaningful historical phenomena with as few theory-dependent presuppositions as possible. These results can provide an unbiased, data-driven complement to human observation and intuition, since rather than testing an existing model, correspondences are extrapolated directly from the data. We may then see in how far these results match known descriptions of historical change, and why.

Our approach to finding which historical phenomena correspond to which between language stages will draw on machine translation (MT) techniques. While the application of MT to historical linguistics may seem odd at first, it is not altogether unnatural. MT is similar to historical linguistics since it too is concerned with correspondences between two languages. Like the diachronic researcher, who looks for regularities governing the relationship between language stages, an MT system needs to describe the relationship between two languages in order to create a translation for every input. MT is therefore suitable for answering questions of the sort: "which Y in language B does X in language A correspond to? Under what circumstances, and how often?". Statistical machine translation (SMT) is an MT technique which relies on a parallel corpus to deduce correspondences between languages, rather than using preconfigured translation rules. A basic SMT system is

comprised of two parts: a "translation model", which tells us how likely it is that a certain element will appear in the target language given that another element appears in the source language; and a "language model", which describes the probability of certain items and sequences of items appearing in the target language. These two models are then used together to produce a translation: the most likely sets of target items are identified by the translation model, and different arrangements of the items in these sets are evaluated by the language model to find the optimal output (see Somers, to appear). Translation models are based on the idea that items recurring in aligned sections of a parallel corpus are more likely to be translations of each other. This can be illustrated with the following German-English parallel sentences:

I eat apples : Ich esse Äpfel
I eat oranges : Ich esse Orangen

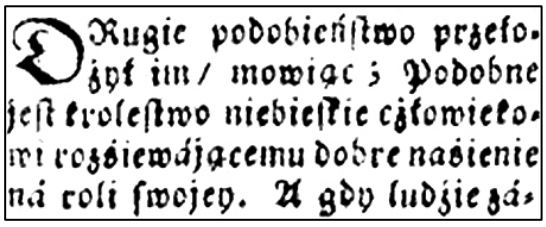
We can deduce that "I eat" translates "Ich esse", since the appearance of these items is correlated (more complex approaches also deduce from differences between parallel pairs that "apples" is a translation of "Äpfel", and "oranges" of "Orangen" (Somers, 1999; Cicekli and Güvenir, 2001), but these will be left aside for now).

While we are not interested in translating text automatically from older language stages into newer ones, we can learn from both models. Language models are the characteristic mono-lingual distributions of words, collocations and other items or sequences. A translation model between two language stages is like a historical lexicon when applied to words or expressions, and like a historical grammar when applied to correspondences between constructions or grammatical properties in an annotated corpus. Using automated techniques, correspondences between hundreds of lexical items and constructions can be easily located, which would be difficult to do manually. However caution is as always warranted: finding some phenomenon in an old text and a different one under similar circumstances in a new text does not mean that one element has 'replaced' the other in the usual sense. Often two or more constructions or words compete for extended periods, having subtly different meanings and usage (Rissanen, to appear; Labov, 1994: 27). Language change can thus be seen as a process characterized by variation or variability in the significatory value (in the structuralist sense) of different signs in a related field (*cf*. Curzan, to appear). It is only in this sense that a new attested form replaces an old one: by being used in a corpus, one is chosen over the other within such a field, effectively taking part in the constant renegotiation of the linguistic value of the field and the items in it.

The next section briefly presents the corpus created for this study, followed by a short discussion of the validity of Bible corpus-based inferences. Section 3 then examines the parallel distribution of Polish nominal inflectional suffixes. Section 4 offers a quantitative study of lexical change in verbal stems, roots and prefixes based on automatic translation pair extraction. Section 5 concludes with examples of data-based parallel syntactic pattern identification.

2 The corpus

For this study a small parallel corpus drawn from the Polish Bible translations was created, containing two translations of the Gospel of Matthew. The older translation was taken from the Protestant Gdansk Bible (Biblia Gdańska), first printed in 1606 (the New Testament) and then in 1632 (New and Old Testament). Since the original



Drugie podobieństwo przełożył im, mówiąc: Podobne jest królestwo niebieskie człowiekowi, rozsiewającemu dobre nasienie na roli swojej.

Fig. 1: The same text (Matthew 13:24) in a facsimile of the 1632 edition next to the digital text of the Gdansk Bible (reproduced from scans of the Württembergische Landesbibliothek Stuttgart, available at http://www.bibliagdanska.pl).

text is not available electronically, a concession had to be made to use the modern edition of the text, which is available on-line (originally obtained from http://www.biblia.com.pl/, now available on the Polish Wikisource at http://pl.wikisource.org/wiki/Biblia_Gda%C5%84ska). This edition has undergone two revisions (the more recent being the Warsaw revision of 1881), affecting its orthography, punctuation, and in a few cases some inflectional endings (e.g. whether Hebrew proper names inflect or remain indeclinable), which makes it unreliable for the study of orthography/phonology, but otherwise suitable for a variety of linguistic studies of the text. Figure 1 shows a passage from the 1632 edition alongside the electronic version to illustrate its relative faithfulness. The newer translation was taken from the 1990 edition of the Warsaw Bible (Biblia Warszawska), first published in 1975, which is the text which finally replaced the archaic Bible of Gdansk as the standard Polish Protestant Bible (available e.g. at http://www.bapost.ok.info.pl/nt/).

The corpus was tagged and lemmatized using a tagging programme called Polimorph (see Zeldes, 2006), which was expanded to handle the older language. Disambiguating the older text was facilitated by projecting annotations from the modern text. An advantage of this tagger is that it outputs the morphological suffixes used to identify a form, and these can be annotated in the corpus (this will be taken advantage of in section 3). The suffixes follow a morphophonological notation along the lines used in Swan's (2002) grammar. This means that allomorphs of the same suffix are represented using one variant (e.g. /y/ for both allophones <i> = [i] and <y> = [y]). Some of the suffixes appearing below have a prefixed capital R followed by a number (R1-R4). These symbols indicate which, if any, mutation the suffix may cause in the stem to which it is attached. For example, two different suffixes containing the phoneme /y/ mark the forms <ciężk**i**> 'heavy (nom. sg. masc.)' and <ciężc**y**> 'heavy (nom. pl. masc. personal)'. The first suffix, which palatalizes the stem's final /k/ into /k'/, is notated as R4y#, while the second, which mutates the /k/ into an affricate /c/, is notated as R1y# (for a complete account of these operators see Zeldes (2006)).

The entire parallel corpus with both texts contains a little over 46,000 tokens, in 1,071 aligned verses. The small size in terms of a normal, mono-lingual corpus is partly made necessary by the lack of reliable training data for tagging the older language, meaning annotation must be manually proofread. On the other hand, this also ensures high quality tagging, and the size has been shown to be sufficient for the application of many statistical and especially MT techniques, which often achieve various tasks at good success rates with well below 1,000 example pairs (Somers, 1999: 119-121; see also Nurmi, 2002 for an evaluation of monolingual research with a relatively small corpus). This is mainly possible thanks to the interdependency between the two texts, which allows drawing founded conclusions from comparably little data, provided annotation quality is high and the parallelism is faithful.

According to Fung (1998: 2), algorithms for extracting correspondences from parallel corpora depend on the following characteristics:

- Words have one sense per corpus.
- Words have a single translation per corpus.
- There are no missing translations in the target document.
- The frequencies of words and their translations are comparable.
- The positions of words and their translations are comparable.

These properties seem to generally hold with regard to the Bible text, which is typically translated very painstakingly and completely, and is also semantically relatively homogeneous, reducing polysemy. The similarity of the language stages also contributes to proximity of word order and comparable frequencies. We are thus in a good position to use SMT on a parallel diachronic Bible corpus.

While not optimal for many purposes, Bible corpora have been widely used in historical linguistics long before the advent of computer technology, not only because of the text's theological and cultural significance, but simply because the Bible (and in particular the Gospels) is one of the earliest sizable texts documented for many (especially European) languages. The Bible also holds major attractions for modern corpus linguistics (Resnik *et al*., 1999): the digital text is freely available in an unparalleled variety of languages, and has been repeatedly updated in various periods, making it ideal for comparative and diachronic studies (see also Cysouw and Wälchli, to appear). The dependable consistency of verse alignment between corpora is both effortless and more accurate than many automatic alignments – misalignment occurs in only a handful of cases (Resnik *et al*. 1999: 135) compared to average success rates between 90-95 percent for automated alignments (admittedly on sentence alignment tasks, more fine-grained than verse alignment, see Simard *et al*., 2000: 54-55). The care taken in translating the Bible also makes omissions relatively unlikely. The main objections to using the Bible for linguistics are probably that (*cf*. Resnik *et al*., 1999):

1. it is a translated text especially prone to loan translations/foreign constructions which preserve the language of the source text (often itself a translation);
2. it is a semantically very marked text, whose special religious content bears only a limited similarity to the 'general language';
3. biblical language is conservative, and therefore unsuitable for historical study.

The first two points are not independent of each other: many expressions that can be traced back to loan translations form part of the style of biblical language. As a consequence, once a loan construction has been accepted into the language through the text, it often becomes part of that language's native inventory, a fact which speakers are usually unaware of. Are the expressions *God fearing* or *to fear God* valid English phrases, or the everyday German word *hartnäckig* 'obstinate, stubborn'? These all represent loan translations reaching as far back as Biblical Hebrew, where יָרֵא אֶת-יְהוָה 'to fear God' had the sense 'to be devout', and קְשֵׁה-עֹרֶף 'hard naped' meant 'refusing to bow' and hence 'stubborn' (on Polish Biblical phraseology see Koziara, 2001). While modern biblical languages owe their existence at least in part to a sort of 'translationese', the naturalization of many of these forms is hard to ignore.

Additionally, although the Bible (and in fact any text) has some idiosyncratic properties, it still shows considerable overlap with "general language". Resnik *et al*. (1999: 147) compared the vocabulary of the Modern English New International Version of the Bible with the control vocabulary list used to write definitions in the Longman Contemporary Dictionary of the English Language, which is meant to represent the core vocabulary of the language most suitable for learners. The Bible corpus contained around 80 percent of the lemmas on the 2,200 word list, thus showing that the Bible's vocabulary did in fact cover central areas of modern language. That said, it remains important to avoid what has been termed the "God's truth fallacy" (Rissanen, 1989), which essentially means reliance on a corpus as representative of an entire language, while disregarding its limitation in belonging to a certain time, place, genre and author. In the end this situation is partly inevitable for some older languages, since our data is rather limited, and often religiously motivated. This mandates greater care to limit our statements as applying to a particular sub-language. "Biblical Polish" or the biblical language of any other standard language can be accepted as a sub-language, insofar as it is recognized by speakers as belonging to their language and interacts with standard language as well[1].

The third objection has been partly addressed already, in that possible conservatism in biblical language is immediately part of the characteristics of the sub-language about which we make statements. Furthermore, conservatism has some advantages for historical research: if a new version of a conservative text was forced to alter some element or construction, it is all the more likely that it was really no longer tolerated or comprehensible in contemporary language. Those elements that were changed may thus indicate central points in historical grammar and lexicography.

3 Nominal suffix changes and distributions in a parallel corpus

A parallel diachronic corpus with morphological suffix annotation provides two interdependent morphological language models. Since the language stages are closely related, we can expect to find many forms where the same lemmas are used in parallel with the same grammatical functions, but possibly with different suffixes. We can thus identify changes in suffixal morphology by searching for tokens with identical lemmas and grammatical analyses (case, gender, number, *etc*.), but different suffixes. Such pairs are made possible by the tagger used to prepare the corpus, which uses a dictionary that does not specify the list of permissible suffixes for each lemma. Instead, it accepts any suffix which may be used to create a regular form of any lemma as a possibility for analysis (this is comparable to an English tagger accepting a regularized form <oxes> for the 'correct' plural form <oxen>, see Zeldes (2006) for more details).

Table 1 lists the results of a query for all different suffix pairs in the parallel corpus which mark the same form of the same common noun lemma. For the analyses note that Polish distinguishes three masculine genders: personal or 'virile' (MP), animate (MA) and inanimate (MI); M means any one of these. All of the alternations in the table correspond to historical developments in Polish nominal morphology

[1] Scriptural language has generally been very influential in shaping many standard literary languages, *cf*. the influence of Luther's Bible on the development of standard German (Wolf, 1996). In a text with such normative influence as the Bible, where the choice of each item in the corpus makes it the *de facto* normative bearer of the Biblical meaning invested in a particular passage, the case is all the stronger for regarding a single text as a sublanguage in itself.

	Analysis	Suffix Pairs		Examples		Sense
1	acc pl MP	R4y#	ów#	*anioły*	*aniołów*	angels
		R4e#	ów#	*króle*	*królów*	kings
		R4e#	R4y#	*nauczyciele*	*nauczycieli*	teachers
		#	R4y#	*sług*	*sługi*	slaves
2	gen pl MP	ów#	#	*poganów*	*pogan*	heathens
3	gen sg MI	a#	u#	*podołka*	*podołku*	hem
4	inst pl M/N	R4y#	ami#	*duchy*	*duchami*	spirits
5	inst pl MI	mi#	ami#	*kijmi*	*kijami*	clubs
6	inst pl N	R4yma#	ami#	*uszyma*	*uszami*	ears
7	nom pl MP	owie#	R4e#	*wężowie*	*węże*	snakes
		owie#	R4y#	*narodowie*	*narody*	peoples
8	nom/acc pl F	R4y#	e#	*nocy*	*noce*	nights
9	acc sg MP	#	a#	*(wyjść za) mąż*	*męża*	husband

Tab. 1: Variant suffix pairs in nominal morphology.

(though the last entry, due to the expression *wyjść za mąż* 'to marry', contains a fossilized accusative zero suffix # which is not transparent). For example, the fluctuation of the genitive masculine singular between the suffixes a# and u# on row 3 is part of a known trend to make animate masculine nouns have the genitive in a#, and inanimates in u# (Rospond, 2003: 126-127). This process is still ongoing in contemporary Polish, with endings changing in both directions, though considerable groups of exceptions persist (Swan, 2002: 72-73). The reality of such phenomena can be studied in distributional data from the corpora. It gives only a weak indication of this process here – the distribution of the suffixes is similar in both texts (Figure 2). The older corpus shows a majority of a# genitives (upper white sections), but a higher frequency of u# genitives in the inanimate masculine (the two sections labelled MI). The new corpus shows much the same distribution, with a perhaps slightly larger proportion of inanimate u# genitives (110:70 or 61.1 percent, instead of 92:76 or 54.7 percent). Nonetheless, inanimate a# genitives remain quite wide-spread. The other a# genitives form two groups, the large group of genitives marking persons (MP) and the small group marking animate genitives (MA).

A clearer change can be seen in the instrumental plural suffixes of the neuter and masculine genders (rows 4-6). Here we see the loss of the old dual form (row 6), the irregular suffix mi# being thematized into the regular suffix ami# from the pronominal and feminine nominal declensions (row 5), and the replacement of the old

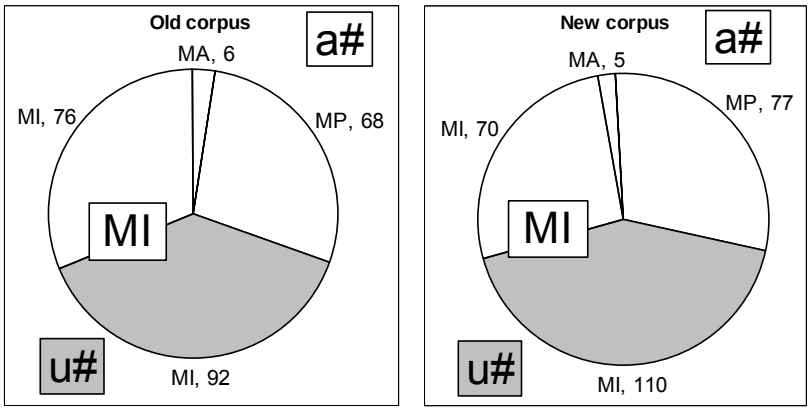


Fig. 2: Distribution of masculine genitive suffixes.

regular R4y#, also by the new ami# (row 4). The spread of ami# at the expense of the other suffixes occurred over the course of the 17th century (Wiśniewska, 1994: 110-111), as reflected by the different corpus distributions (Figure 3).

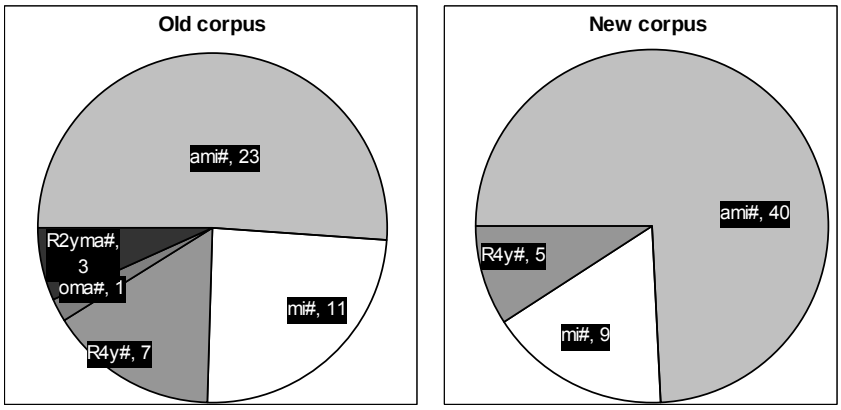


Fig. 3: Distribution of masculine and neuter instrumental plural suffixes.

The graphs in Figure 3 confirm the disappearance of the dual suffixes oma# and R2yma#, which appear only on the left. The last productive days of these suffixes were probably in the 16th century, but use in nouns signifying natural duals such as hands, eyes *etc*. was still the norm well into the 18th century (Klemensiewicz, 1999: 304). The forms are now considered archaic (Swan, 2002: 119). Otherwise the graphs show only a slight decrease in mi# and R4y#, however, an examination of the actual instances of modern R4y# shows it to be limited to a petrified use in the fixed expression *tymi słowy* 'with these words'; in other words the suffix was only retained where it was lexicalized (Wiśniewska, 1994: 110). The examination of the parallel corpus can thus automatically identify and, subsequently, quantitatively substantiate the existence of historical processes in suffixal morphology; however attention must always be given to the underlying data, which must be examined in order to ensure no artefacts are being produced by other factors.

Another way of investigating inflectional morphology is to examine not which suffixes signify a grammatical category (e.g. genitive or instrumental masculine), but rather, in the spirit of Jespersen's Systematic Grammar (1924: 30-57), to ask what roles each suffix plays in the language. Again we will limit ourselves here to the suffixes of common nouns. Figure 4 gives the frequency of each of the major suffixes (very rare irregular suffixes have not been considered) in both corpora, and how often they express which cases. The overall similarity of the distribution despite the limited size of the corpus stems from the fact that both texts share essentially the same content, but there are some subtle differences. For example, in the new corpus, besides signaling genitive, the suffix ów# often marks accusatives (the white part of the seventh bar from the top), but this is not so in the old corpus. This is related to the first two examples in row 1 of Table 1, which show some other suffixes for the accusative plural being replaced by this suffix, which originally signaled genitive only. The items showing this accusative suffix all signify male humans, for reasons which are well known (Klemensiewicz *et al*., 1955: 271-272, 281-282): animate masculine singulars used the same form as the genitive also for the accusative already in the oldest Slavic monuments. The development was originally motivated by the accusative singular form becoming identical to the nominative due to sound change, which disrupted subject identification, mainly in transitive sentences with human

male subjects. The anomaly of using genitive-accusatives in the singular but not in the plural was thus resolved. We can also notice that the suffixes marking the instrumental are all unambiguous (solid black bars) except for R4y#, which has only a tiny black sliver. This explains the pressure to lose this ending as mentioned above.

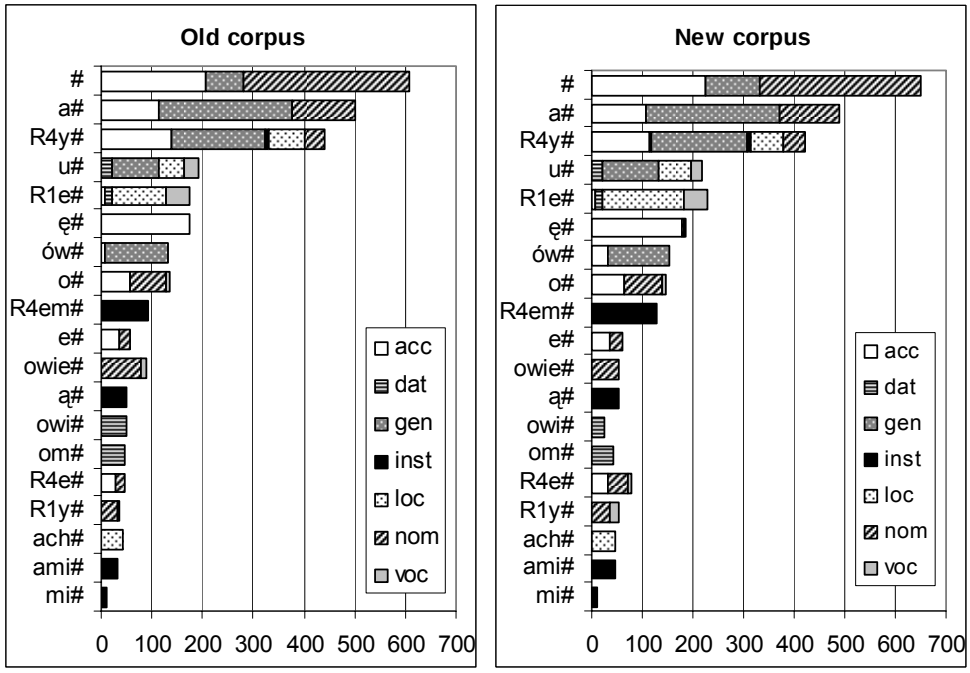


Fig. 4: Distribution of nominal suffixes and cases in the old and new corpora.

These examples are but a few of the statements that can be made using parallel distributions. The parallel corpus allows us to examine phenomena in our text quantitatively, and even to point them out using queries to compare the two corpora automatically. We next turn to automatic extraction of lexical correspondences between the corpora, and its application for examining changes in verbal prefixation.

4 Parallel lexical extraction and a study of verbal prefixation

In this section, I will apply cooccurrence measures used in SMT translation models and the construction of parallel terminologies, in order to model relationships between lexical items in both language stages. Items can be word forms, lemmas, or even morphological or syntactic features depending on the research question being asked. Here we will examine lemmas, as well as collocations, sometimes referred to as multi-word units or idioms, as item candidates. Collocations are understood as sequences of multiple tokens whose semantic/syntactic properties cannot be predicted from their components (Evert, 2004: 17), and which, more importantly for our purposes, may have their own corresponding translations independently of their components. To identify collocations in each corpus we use the *z*-score (see Figure 5), a well established measure which has the advantage of applying to both contiguous

$$p = \frac{B}{N - A} \qquad E_{(c)} = p \cdot A \cdot S \qquad z = \frac{C - E_{(c)}}{\sqrt{E_{(c)}(1-p)}} \qquad MI3 = \log_2 \frac{C^3 \cdot N}{A \cdot B}$$

Fig. 5: The z-score for 2 items appearing A and B times in total and C times together in a span of S items in a corpus of N items; and MI3 for two items appearing separately in A and B pairs respectively, and in C pairs together among N possible pairs (see Oakes, 1998: 163-166, 170-172).

and non-contiguous tokens (for an evaluation of the *z*-score against other measures see Pearce (2002)). Next we test correlations between items in parallel sections of the corpus. For this we have used Daille's (1995: 36-37) Cubic Association Ratio (sometimes called mutual information cubed, or MI3) which subjectively seems to perform well, though other measures tested, such as Log Likelihood (Dunning, 1993), have shown very similar results. MI3 gives a score between plus and minus infinity of how likely we are to find *b* in a parallel section given *a* appears in the source section.

This results in a table listing the association strength between each two lemmas or collocations that appear in parallel aligned sections, though a cut-off point for significant matches must be chosen empirically on a case by case basis. Matching items may be identical, and collocations may match with single lemmas, as shown in Table 2. The lemmas *przedni* 'front, fore' and *kapłan* 'priest' are recognized as collocates, since the former appears in this corpus only in the phrase *przedniejszy kapłan* 'foremost priest'; in the new corpus this is replaced by *arcykapłan* 'archpriest'. Also, although they are important for translation purposes, punctuation and other token-types with a frequency of more than 1 percent in either corpus (e.g. 'function words' like 'and' *etc*.) are of no interest at this point, and their entries are eliminated.

a (old corp.)	b (new corp.)	Sense	A	B	C	MI3
przedni kapłan	*arcykapłan*	chief priest	441	237	19	13.931
słowo	*słowo*	word	343	585	24	14.001

Tab. 2: Matching lemmas and collocations between corpora.

The following examination of changes in verbal lemmas is an example study using correspondences established in this way. Verbal lemmas can exhibit several types of correspondence between the corpora: the lemma may remain unchanged or it may be replaced by either a non-verbal lemma (possibly also a collocation) or another verbal lemma. If there is a verb-verb correspondence, we can check which parts of the lemma are replaced: prefixes, the root (i.e. the abstract morpheme from which verb stems are formed through vowel gradation and suffixation), or suffixes (see Figure 6), any combination of these, or even, as we shall see, the rules for combining them. In this example we will attempt to automatically find verb substitutions retaining the same root, but with a different prefix. The importance of such changes is in that they affect a very large portion of the lexicon by renegotiating the linguistic value of both the prefix in question in all lemmas that exhibit it, and all the other prefixes in

	Old Corpus		New Corpus	
Prefix change:	*na-śmiać*	:	*wy-śmiać*	'ridicule'
	at-laugh		out-laugh	
Root change:	*wy-gnać*	:	*wy-pędzić*	'drive out'
	out-chase		out-rush	
Suffix change:	*za-bieżeć*	:	*za-biec*	'run across' (both from root *bieg*)
	beyond-run		beyond-run	

Fig. 6: Examples of corresponding verb pairs with different parts substituted.

opposition to it (particularly the other prefix involved in the substitution in that instance). We will look for verbal lemmas in the old corpus which show the best (but not necessarily only) correspondence with another, non-identical verbal lemma. We then compute Levenshtein Distance (LD) between the lemmas, which checks how

many character insert, delete or replace operations are required to transform one string into another. Items with zero LD are identical, and represent non-change of a lemma. High LD is characteristic of total replacement of a lemma, probably including the root, while low values signify a partial change. We also need to check where the difference between lemmas is: at the left of the strings (possible prefix change), at the right (stem change), or both. We therefore define two sets of functions: *LeftChangeIndex* and *RightChangeIndex*, which return for each lemma the amount of characters left in a string once the first difference has been detected, starting from the left and from the right respectively; and *LeftIdentIndex* and *RightIdentIndex*, which return the amount of identical characters on either end of the strings. The criterion for a prefix change, for example, is set at a *RightChangeIndex* < 5 in both lemmas (the space occupied by the prefixes), and *LeftIdentIndex* > 2, which is required in order to show an identical verb stem (consisting of at least 3 characters – a syllable onset, vowel and the infinitive ending <ć>/<c>). A stem change, by contrast, is found with *LeftIdentIndex* between 0 and 5 and *RightIdentIndex* < 3 (assuming identical prefixes, and a possibly identical suffix). Finally, low LD and *LeftChangeIndex* but high *RightChangeIndex* identify a stem change from the same root. Table 3 illustrates this classification.

a (old)	b (new)	Sense	MI3	LD	LId	RId	aLC	aRC	bLC	bRC
obwarować	*zabezpieczyć*	guard	10.357	10	0	1	9	8	12	11
pełnić	*spełniać*	fulfil	10.575	2	0	1	6	5	8	7
pogrześć	*pogrzebać*	bury	12.365	2	6	1	2	7	3	8
zadziwić	*zdziwić*	amaze	11.302	1	1	6	7	2	6	1
wziąć	*wziąć*	take	15.513	0	5	5	0	0	0	0

Tab. 3: Examples of verb change types with string comparison measures. High LD indicates total replacement, LD=0 means the lemma was retained. Low LD signals partial change, in prefix and suffix, suffix only or prefix only (middle 3 examples).

If multiple parallels have identical MI3 scores but one is a pair of identical lemmas, we assume this is the best parallel (our null hypothesis is that no change has occurred). It is important to make explicit that changes in prefixes *etc*. may modify the meaning of a verb such that a pair is only parallel in a particular use or sense. In treating a pair as parallel we assume a measure of semantic uniformity between our texts by virtue of their forming a parallel corpus: if the pair provides the statistically soundest match, we note that one item was chosen over the other in this context, be it for reasons of a total ousting of the old form through language change or merely stylistic variation between competing items. In this case study we will be interested in questions regarding the relations between different kinds of change that occur in the expression of what is in essence the same semantic content, even though in another context certain replacements may not have occurred, or different ones may have instead (*cf*. Rissanen, to appear).

The results in Table 4 show prefix substitutions using the query described above; results are extrapolated directly from the corpus with no human intervention. The suggested different prefixes (marked in bold) are extracted automatically by taking the first *RightChangeIndex* number of characters on the left of the respective lemma field. The 58 results can be divided into 3 groups. The last six results represent errors which stem from similar looking verbs not actually exhibiting a prefix change. One pair *nałamać* ‘crack’ : *dołamać* ‘break’ does differ in prefix only, but the match is incorrect: though similar in meaning and appearing together in the text, the correct match as far as the parallel text is concerned is found with an equal score further up

the list: *nałamać* : *nadłamać*. The pair *smęcić* : *smucić*, both 'mourn' actually represents the same word etymologically, but the latter form is due to Czech influence, with /u/ instead of the nasal vowel; nonetheless no prefix change is involved. *zmiłować* : *zlitować* 'pity' is a correct pair with coincidentally similar endings and no prefix change. The remaining errors are match errors. The other 52 verb pairs truly differ only in prefixes (including the borderline case *dufać* : *zaufać* 'to believe, trust', where the verbs are related, and the new lemma has added a prefix, but the old lemma's initial 'd' is not a transparent prefix), producing an accuracy rating of 52/58 ≈ 90 percent. As for recall, missing prefix changes must stem from either a missing translation pair in the correspondence table (because a better match could be found instead, *etc*.), in which case the match was not well attested in the corpus and can therefore be safely left out; or from a prefix change not identified by the string comparison criteria. A manual examination of all verb to verb correspondences has revealed only 1 such case, the pair *po**prze**wracać* : *po**wy**wracać* 'to overturn', which is due to a string-internal change of the second prefix being missed. Recall is thus 52/53 ≈ 98 percent, for an F-score of: $F = 2 \cdot Pr \cdot Rc / (Pr + Rc) \approx 94$ percent.

The parallel corpus can thus automatically deliver a fairly reliable list of parallel verbs differing only in prefixes. However, the second group of 11 verbs (marked gray), which exhibit an alternation between having some prefix and no prefix, can all be ascribed to grammatical, and not lexical differences – the prefixed form is perfective, the unprefixed is imperfective. In these cases the new text uses a

Old lemma	New lemma	Sense	MI3
*wyni**j**ść*	*wy**j**ść*	go out	15.24
*wni**j**ść*	*we**j**ść*	go in	14.35
paść	*u**p**aść*	fall	13.57
*po**c**ząć*	*za**c**ząć*	begin	13.22
*s**k**ryć*	*u**k**ryć*	hide	12.80
stawić	*wy**s**tawić*	stand (vt.)	12.70
*u**w**inąć*	*o**w**inąć*	wrap	12.69
wzrosnąć	*wy**r**osnąć*	grow	12.25
*o**b**udzić*	*z**b**udzić*	wake up	12.11
*przy**d**ać*	*do**d**ać*	add	12.08
*za**ś**piewać*	*od**ś**piewać*	sing	11.88
*po**ś**więcać*	*u**ś**więcać*	consecrate	11.76
*po**p**rzedzić*	*wy**p**rzedzić*	precede	11.73
*na**ś**miać*	*wy**ś**miać*	ridicule	11.69
*na**g**otować*	*przy**g**otować*	prepare	11.62
*o**s**ławić*	*znie**s**ławić*	dishonor	11.52
*na**r**odzić*	*u**r**odzić*	be born	11.34
*za**d**ziwić*	*z**d**ziwić*	marvel	11.30
*pa**d**ać*	*s**p**adać*	fall	11.25
*na**ł**amać*	*nad**ł**amać*	crack	11.08
*u**b**ić*	*z**b**ić*	beat up	11.08
*s**t**rudzić*	*u**t**rudzić*	tire	10.95
*s**p**ytać*	*za**p**ytać*	ask	10.93
*u**s**iąść*	*za**s**iąść*	sit down	10.85
*na**ś**miewać*	*wy**ś**miewać*	ridicule	10.82
*o**k**rywać*	*przy**k**rywać*	cover	10.58
*przy**ł**ączyć*	*po**ł**ączyć*	join	10.46
*u**m**ieść*	*wy**m**ieść*	sweep	10.37
*w**s**iać*	*za**s**iać*	sow	10.37

Old lemma	New lemma	Sense	MI3
*d**u**fać*	*za**u**fać*	believe	10.35
*u**r**osnąć*	*pod**r**osnąć*	grow	10.00
*z**w**ołać*	*przy**w**ołać*	convene	9.90
*we**j**rzeć*	*spo**j**rzeć*	glance	9.89
*wy**r**ozumieć*	*z**r**ozumieć*	understand	9.81
*od**n**ieść*	*za**n**ieść*	carry	9.68
*o**t**rząsnąć*	*s**t**rząsnąć*	shake off	9.61
*za**w**ołać*	*przy**w**ołać*	call	9.53
*roz**s**zerzać*	*po**s**zerzać*	widen	9.37
*z**g**otować*	*przy**g**otować*	prepare	9.08
*za**m**yślać*	*roz**m**yślać*	ponder	8.69
*prze**k**linać*	*za**k**linać*	curse	8.31
*po**k**alać*	*k**a**lać*	defile	15.77
*ż**ą**dać*	*za**ż**ądać*	desire	13.98
*za**p**ieczętować*	*p**i**eczętować*	seal	12.28
*d**r**żeć*	*za**d**rżeć*	tremble	12.25
*m**a**czać*	*u**m**aczać*	wet	11.88
*z**r**ozumieć*	*r**o**zumieć*	understand	11.71
*w**i**ać*	*po**w**iać*	blow	11.20
*p**o**dobać*	*s**p**odobać*	please	10.70
*t**r**ząść*	*za**t**rząść*	shake	10.19
*m**i**eszkać*	*za**m**ieszkać*	dwell	9.59
*p**y**tać*	*za**p**ytać*	ask	9.59
*zmi**ł**ować*	*zli**t**ować*	pity	13.01
*sm**ę**cić*	*sm**u**cić*	mourn	12.36
*prag**n**ąć*	*ła**k**nąć*	desire/hunger	12.13
*na**ł**amać*	*do**ł**amać*	crack/break	11.08
*na**s**adzić*	*o**g**rodzić*	plant/fence	9.52
*szpe**c**ić*	*poś**c**ić*	deface/fast	8.97

Tab. 4: Query results for verbal prefix changes.

construction with a different aspect of the same verb, which entails substituting the lemma for one with the appropriate aspect. This incidentally reveals that the perfective form is probably showing the 'default' perfectivizing prefix, with minimal semantic influence, which can be of lexicographic interest in itself (on default prefixes and aspectual pair types see Włodarczyk and Włodarczyk, 2001). These pairs can be omitted by specifying that the aspect of both lemmas must match (this is possible since the corpus is tagged for aspect). The remaining 41 pairs exhibit various interesting historical phenomena of variation in verbal prefixation:

1. Use of prefixed perfective verbs instead of unprefixed, inherently perfective ones: *paść* : *upaść* 'fall', *stawić* : *wystawić* 'stand s.t. out, deploy'
2. Use of prefixed verbs with specialized senses vs. more general or polysemous verbs: *stawić* : *wystawić* 'stand s.t. out, deploy' (*stawić* has more senses outside this context), and conversely *wyrozumieć* : *zrozumieć* 'understand' (*wyrozumieć* has a more specific sense of 'fully understanding').
3. Different choices of default perfectivizing prefixes, which are still in competition today, e.g. *obudzić* : *zbudzić* 'rouse', *spytać* : *zapytać* 'ask', *etc*.
4. Change of directional prefixes, e.g. with *wy-* 'out' focusing on resultativity: *umieść* : *wymieść* 'to sweep', *naśmiać* : *wyśmiać* 'to ridicule', *poprzedzić* : *wyprzedzić* 'to outpace, precede'
5. Change in morphotactics in prefixation of verbs with initial vowel: *wynijść* : *wyjść* 'go out', *wnijść* : *wejść* 'go in'.[2]

Although these phenomena can thus be detected automatically using the previously discussed techniques, it remains clear that they can only be interpreted with knowledge external to this corpus. This is especially true regarding the nature of 'replacements', since as already mentioned, a recognized pair does not mean that one form completely replaced the other over time, only that one was chosen over the other in the modern text in a parallel narrative content. This opens up interesting possibilities for synchronic comparison of the distributions of paired items in non-parallel monolingual corpora, both modern and historical. We must also consider that this coarse procedure does not take multiple senses into account: if multiple senses of a verb undergo different changes, only the most frequent case will be picked up, since matches have been made per lemma. This could be remedied by using sense tagging and ranking of multiple significant correspondences.

We can conduct similar studies of verbs exhibiting stem substitution and prefix retention, stem modification (suffixes, vowel gradation), or total substitution, using different queries on the 'translation model'. Figure 7 indicates the distribution of these change types in the corpus. For 34 percent of verbs in the old corpus no consistent parallel could be found in the new corpus (usually because a lemma is rare and/or has multiple 'translations'), and 2 percent were consistently replaced by non-verbs. For the remaining 64 percent, the results show that a very large portion of verbal lemmas (76 percent) have remained unchanged, which also reaffirms the reliability of the

[2] The rule inserting /n/ between the prefix and vowel was generalized from two common prefixes which preserved an old /n/ in this position, *cf*. Old Church Slavonic *vŭn* 'in-' and *sŭn* 'with-'. When /n/ after /ŭ/ was dropped in closed syllables, the prefixes exhibited two forms: with /n/ before a vowel and no /n/ elsewhere. Other prefixes adopted this behavior, resulting in forms like *vyn-* 'out-', from the prefix *vy-*, which originally had no /n/. The old forms here are the direct descendants of these, whereas Modern Polish has done away with this rule, combining all prefixes with no intermediate /n/.

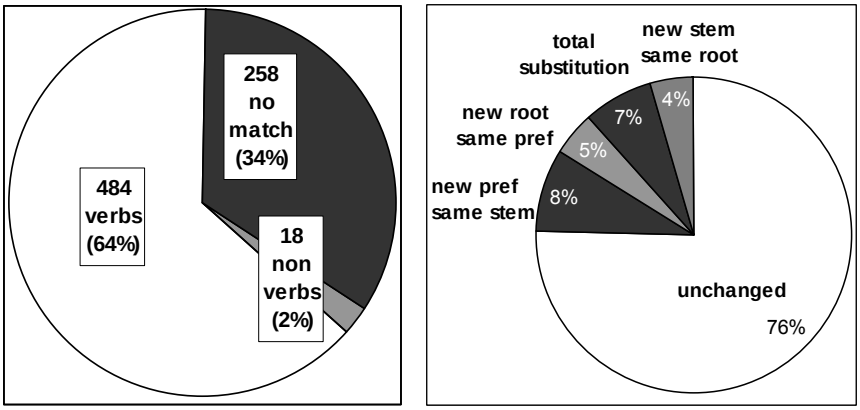


Fig. 7: Verb correspondences and distribution of verb replacement types.

matching process, since identical pairs are almost certainly correct. Of those changed, most cases are not of complete substitution: either a stem, root or a prefix is usually retained. Stem, root or prefix-only changes show that prefix and root semantics are often separable, with the sense of the new verb still requiring either one or the other. Thus a verb with an outwards motion will be prefixed with 'out' even if the new root is unrelated, and vice versa, another prefix may be chosen to mark the completion or particular perspective of a verb, but the semantics of the root will maintain its use. This is more likely to be the case in productive, transparent prefixation, which is especially common in motion verbs (on the central role of motion verbs in the development of Polish prefixation see Śmiech, 1986). We may then speak of complex verbs with composite semantics, and results from queries on prefix retention and stem change seem to substantiate this view.

A parallel corpus can reveal very specific facts about the relationship between forms in the texts it comprises, and in a quantifiable way difficult to obtain otherwise. Even if the text is conservative, we can learn that the same story can be told four centuries apart with only a quarter of consistently parallel lemmas being replaced; and even then, more often than not some elements of the older form are retained, which is of interest in the study of related verbs with different prefixes across the Slavic and Indo-European languages in general. Interpretation of the data should however be integrated into our linguistic knowledge from other sources, including other corpora. This opens up the opportunity for use of larger, non-parallel corpora, against which we can compare the particular kinds of answers parallel corpora can give us.

5 Syntactic change and parallel patterns

The same principles applied to the study of lexical items can also be adapted to the study of grammatical categories and syntagms. If we consider different kinds of items other than lemmas, we can look for significant correlations between syntactic structures between the corpora. Since our corpus is not parsed, syntactic structures will have to be defined in terms of flat, recurring patterns of tokens. This level of abstraction is not ideal, but the rich case system in Polish often makes establishing subject, object, congruent attributes *etc*. possible even without a parse. In principle, however, a parsed corpus could be used to identify structures more accurately, and their occurrences in aligned sections could be correlated in the same way. As a simple example we may consider the development of copulas used with the passive participle. We can search for passive participles by using part-of-speech (POS) information but discarding lemma information, and then see what lemmas occur next to these

participles in each corpus. The query in Table 5 searches for correlated parallel lemmas that occur in bigrams containing passive participles (excluding punctuation and conjunctions). It retrieves only two significant matches.

a (old corpus)	b (new corpus)	A	B	C	MI3
być	być	188	164	80	13.07
być	zostać	188	31	18	9.018

Tab. 5: Correlated lemmas next to passive participles.

Row 1 shows that passive participles usually occur next to the verb *być* 'be' ('to be verbed'), but row 2 shows another parallel, with *zostać* 'become', which is used regularly as a copula with perfective passive participles (in the future or past 'will be/was verbed'). Although the use of *zostać* is a much younger development (*być* is the older form found in all Slavic languages), both forms were in use throughout the recorded history of Polish and evolved alongside the establishment of verb-stem aspect (Długosz-Kurczabowa and Dubisz, 2006: 316). Yet this construction does not occur in the old corpus at all, where the lemma *zostać* is used as the perfective form for the fully lexical verb 'to remain, stay', and *być* is used with both aspects of passive participles. In the modern language *zostać* is prevalent for perfective participles, though use of *być* continues both with imperfective participles to express imperfective actional passives ('is being/was being/is going to be verbed') and occasionally with perfective participles to express a statal or resultative sense of completion ('had/has/will have been verbed') (Swan, 2002: 312-314).

We may also wish to consider syntagms which satisfy certain internal constraints, such as congruence. To do this we require a function which receives the grammatical analyses of all tokens in a sequence, and outputs whether or not a pair of them may be congruent. Once we have the congruence information we may decide to discard lemma, number and gender information, which is less significant for certain questions once congruence is established. We may also choose to retain lemmas for certain classes of words, such as prepositions, or very common words such as *być* 'to be', the exact identities of which play significant grammatical roles. Figure 8 shows some examples of the output of such a function.

1. impfv pres 3 sg prep l o c s g F
   jest na pustyni > [VFin być impfv pres 3] [Prep na] [NN loc]
   he is in the desert
2. n o m p l M P pfv past 3 pl MP acc sg MP
   wieśniacy ujrzeli syna > [NN nom **agr**] [VFin pfv past 3 **agr**] [NN acc]
   the villagers saw the son

Fig. 8: Abstracting token sequences. The lemmas *być* and *na* are not stripped since they are very frequent, and in the case of *na* belong to the reserved class of prepositions. The tokens *wieśniacy* and *ujrzeli* are stripped, but receive the feature *agr*, since they are congruent.

As an example of how such sequences can be used to detect syntactic change, we examine possessive adjectives. These adjectives are derived from proper nouns with suffixes such as owy#, and were used in Old and Middle Polish, just as already in the oldest Slavic documents, to express possession (Pisarkowa, 1984: 128-9; Rospond, 2003: 195), e.g. *Syn Dawidowy* 'Son of David', literally: 'Davidian son'. Searching for consistent parallel bigrams with an old congruent noun and its possessive adjective yields the top five results in Table 6.

a (old corpus)	b (new corpus)	A	B	C	MI3
[NN gen agr] [AdjPos gen agr]	[NN gen agr] [AdjPos gen agr]	142	72	8	13.39
[NN voc agr] [AdjPos voc agr]	[NN voc] [NP gen]	97	77	5	11.81
[NN nom agr] [AdjPos nom agr]	[NN nom] [NP gen]	42	199	6	12.43
[NN dat agr] [AdjPos dat agr]	[NN dat agr] [AdjPos dat agr]	33	31	2	10.71
[NN acc agr] [AdjPos acc agr]	[NN acc] [NP gen]	63	194	3	8.88
[NN gen agr] [AdjPos gen agr]	[NN gen] [NP gen]	142	207	3	7.62
[NN gen agr] [AdjPos gen agr]	[NP gen] [NP gen]	142	71	2	7.41
Total: [N* agr] [AdjPos agr]	*Total:* [N* nom] [NP gen]	4779	2169	31	9.26
Total: [N* agr] [AdjPos agr]	*Total:* [N* agr] [AdjPos agr]	2169	696	14	8.60

Tab. 6: Parallel bigrams with an old possessive adjective.

As the query results show, the construction was often replaced by qualifying the noun (POS-tag NN) with a proper noun (NP) in the genitive (e.g. 'David's son'), a phenomenon which gradually reduced use of the old construction beginning as early as the 16th century (*ibid*.). The old construction has also been left intact relatively often e.g. in rows 1 and 4, though the next best match for the construction in row 1 would also be formed by the noun + genitive construction if it had not been split between two versions (see the gray rows in the table): one type of sequence qualifies a proper noun and the other a normal noun, and they are counted separately. An examination of all cases and noun types together (last two rows) shows that the association between the old possessive adjective and the newer genitive construction is in fact not much stronger than the archaic variant, meaning many cases (almost a third) were not replaced. This can probably be ascribed to the relative conservatism of the text. The accurate alignment of the corpus thus allows correct identification of the old construction and its competition with its younger contender, despite the relative infrequency of the phenomenon. A simple query on the proportion of the two constructions between corpora (disregarding parallelism) would show the genitive construction to have become much more dominant than it actually is in this use (Figure 9).

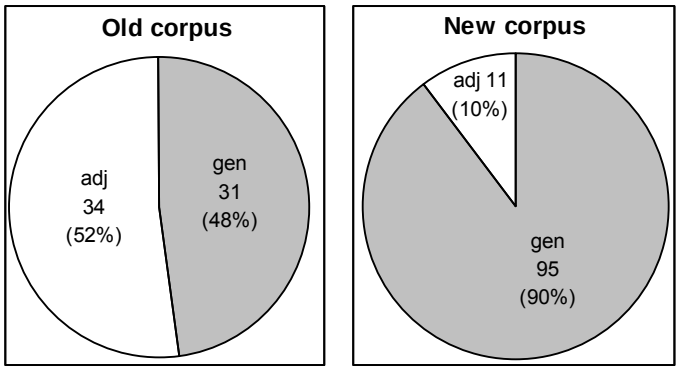


Fig. 9: Proportions of the proper noun-genitive and possessive adjective constructions.

This is because we cannot guarantee that all occurrences of the genitive construction in the new corpus are in fact translating old possessive adjectives; they may simply represent a coincidental appearance of a genitive proper noun next to another noun, and indeed, the construction appears almost three times as often as there are possessive adjectives in the old corpus, and about 150 percent more often than the adjectives and the genitive construction in the old corpus put together. Some matches are therefore clearly unrelated to this development, and this can only be discerned by taking advantage of the parallel alignment as in Table 6.

6 Conclusion

In this article I have discussed some of the uses of diachronic parallel corpora for historical linguistics using the example of developments in the language of the Polish Bible. As we have seen, a parallel corpus can be used to directly extract diachronic developments. It can be used to find and point out phenomena automatically by extracting differences between similarly tagged items (the suffix change examples in section 3). Comparison of the distributions of retrieved items can be more illuminating thanks to the virtually identical content of the texts – differences related to subject matter, register, genre *etc*. can be neutralized. Quantitative studies on change between language stages, and especially in the lexicon, where a parallel corpus can output a subset of a historical dictionary, can easily be carried out; this would be much more effort intensive manually, and less informative if done only using distributions in non-parallel corpora. Using MT techniques, we can directly address questions of what replaces what and to what extent (sections 4-5), rather than be limited to general statements on relative frequency, which might not be due to consistent correspondence of particular items but to other factors or items. We can also target particular items or constructions with specific queries and get answers directly from the data. These have, in this study, generally been in line with traditional historical grammars, bearing in mind that Bible text is more conservative in many ways than the general language, though this perhaps makes the changes that are found more meaningful, and draws attention to discrepancies as features of this text.

On the other hand, parallel corpora bring with them their own methodological problems. They are typically smaller, and historical ones are often limited to religious texts. In some cases this is all we have of a language stage (e.g. Old Church Slavonic), but in cases like Middle Polish, we could have considered many more texts if we did not limit ourselves to a parallel corpus. This reliance on a homogeneous text can provide very accurate results on the relationship between two texts from two stages, but is consequently limited to a small sub-language, and is incapable of separating different factors such as stylistic variation, diatopic influences and in the case of many documents, the peculiarities of translated text. A partial improvement in this situation might be to use multiple contemporary versions of a text (in the case of the Bible and other canonical works multiple translations may be available), filtering out the differences between texts from the same period as synchronic variation, and focusing on commonalities as diachronic evidence. But most importantly, parallel corpora can be used alongside larger, more heterogeneous historical corpora, in the light of which the peculiarities of a particular parallel corpus can be assessed, and the role of its sub-language in larger subsets of the general language understood. This would have major applications in supporting the creation of fine grained historical lexica, or the enrichment of general lexica with historical information, especially using larger corpora. Finally, like any corpus, a parallel corpus can only reveal information about what is annotated in it. Much work can still be done using e.g. parsed or sense annotated corpora, which may also facilitate implementation of more complex MT techniques, such as example-based machine translation (see Somers, 1999 and to appear), for more advanced parallel construction extraction.

References

Cicekli, I. and H. A. Güvenir (2001) ‘Learning translation templates from bilingual translation examples’. *Applied Intelligence* 15, 57-76.

Curzan, A. (to appear) Historical Corpus Linguistics and Evidence of Language Change, in A. Lüdeling and M. Kytö (eds) Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter.

Cysouw, M. and B. Wälchli (to appear) 'Parallel texts: Using translational equivalents in linguistic typology'. Special issue of *STUF*.

Daille, B. (1995) Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering. Unit for Computer Research on the English Language Technical Papers 5, Lancaster University.

Długosz-Kurczabowa, K. and S. Dubisz (2006) Gramatyka historyczna języka polskiego. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.

Dunning, T. (1993) 'Accurate methods for the statistics of surprise and coincidence'. *Computational Linguistics* 19(1), 61-74.

Evert, S. (2004) The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD dissertation, University of Stuttgart.

Fung, P. (1998) 'A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora'. *Lecture Notes in Artificial Intelligence* 1529, 1-17.

Jespersen, O. (1924) The Philosophy of Grammar. London: George Allen & Unwin.

Klemensiewicz, Z. (1999) Historia języka polskiego (wydanie siódme, uzupełnione). Warsaw: Wydawnictwo Naukowe PWN.

Klemensiewicz, Z., T. Lehr-Spławiński and S. Urbańczyk (1955) Gramatyka historyczna języka polskiego. Warszawa: Państwowe Wydawnictwo Naukowe.

Koziara, S. (2001) Frazeologia biblijna w języku polskim. Kraków: Wydawnictwo Naukowe Akademii Pedagogicznej.

Labov, W. (1994) Principles of Linguistic Change. Volume 1: Internal Factors. Oxford, UK and Cambridge, MA: Blackwell.

Nurmi, A. (2002) Does Size Matter? The Corpus of Early English Correspondence and its Sampler, in H. Raumolin-Brunberg, M. Nevala, A. Nurmi and M. Rissanen, (eds), Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen. Mémoires de la Société Néophilologique de Helsinki LXI, pp. 173-184. Helsinki: Société Néophilologique.

Oakes, M. P. (1998) Statistics for Corpus Linguistics. Edinburgh: Edinburgh University Press.

Pearce, D. (2002) A comparative evaluation of collocation extraction techniques. *Third International Conference on Language Resources and Evaluation, May, 2002, Las Palmas, Canary Islands, Spain*.

Pisarkowa, K. (1984) Historia składni języka polskiego. Wrocław *et al*.: Polska Akademia Nauk.

Resnik, P., M. Broman Olsen and M. Diab (1999) 'The Bible as a parallel corpus: Annotating the "book of 2000 tongues"'. *Computers and the Humanities* 33, 129-153.

Rissanen, M. (1989) 'Three problems connected with the use of diachronic corpora'. *ICAME Journal* 13, 16-19.

Rissanen, M. (to appear) Corpus Linguistics and Historical Linguistics, in A. Lüdeling and M. Kytö (eds) Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter.

Rospond, S. (2003) Gramatyka historyczna języka polskiego, z ćwiczeniami. Warsaw: Wydawnictwo Naukowe PWN.

Simard, M., G. Foster, M-L. Hannan, E. Macklovitch and P. Plamondon (2000) Bilingual Text Alignment: Where do we Draw the Line?, in S. P. Botley, A. M. McEnery and A. Wilson (eds) Multilingual Corpora in Teaching and Research. Amsterdam - Atlanta, GA: Rodopi.

Śmiech, W. (1986) Derywacja prefiksalna czasowników polskich. Prace wydziału I – językoznawstwa, nauki o literaturze i filozofii 87. Wrocław et al.: Łódzkie Towarzystwo Naukowe.

Somers, H. (1999) 'Review article: Example-based machine translation'. *Machine Translation* 14, 113-157.

Somers, H. (to appear) Corpora and Machine Translation, in A. Lüdeling and M. Kytö (eds) Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter.

Swan, O. E. (2002) A Grammar of Contemporary Polish. Bloomington, IN: Slavica Publishers.

Wiśniewska, H. (1994) Kulturalna polszczyzna XVII wieku (na przykładzie Zamościa). Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.

Włodarczyk, A. and H. Włodarczyk (2001) 'La préfixation verbale en polonais. I. Le statut grammatical des préfixes, II. L'Aspect perfectif comme hyper-catégorie'. *Études cognitives / Studia kognitywne* 4, 93-120.

Wolf, H. (1996) Einführung, in H. Wolf (ed.) Luthers Deutsch. Sprachliche Leistung und Wirkung, pp. 9-29. Frankfurt am Main: Peter Lang.

Zeldes, A. (2006) Abstracting suffixes: A morphophonemic approach to Polish morphological analysis. *Proceedings of Konvens'06, Konstanz, 4-7 October, 2006*, 151-158. An extended version of this article is set to appear in a special issue of *Zeitschrift für Sprachwissenschaft*.