Co-funded by the
Erasmus+ Programme
of the European Union

Corpus Linguistics & Digital Editions

Contributed by Ondrej Tichy

Contents

1. What is Corpus Linguistics

1a. What is a Corpus

A linguistic corpus represents a computer readable body of text. Usually, it also attempts to be:

a. **limited** to some variety (language, dialect, sociolect, genre, period etc.) depending on the subject matter under study. Very large, diffuse and linguistically “un-focused” corpora (like Google Ngrams) tend to give less linguistically reliable results.

b. **representative** of that variety (i.e. not skewed by subjective text selection criteria)

c. **large**, that is to say that it generally contain anywhere between 1 million to hundreds of billions of words/tokens.

d. enriched by **meta-data**.

1b. Linguistic Introspection

The traditional method of linguistic analysis is 'Linguistic introspection'. As theorized by Leonard Talmy, this is conscious attention directed by a language user to particular aspects of language manifest in their own cognition. Ultimately this methodology is quite precarious since it relies heavily on very small data-sets, i.e. the user's own report of linguistic introspection and the reports of others. There are limits to what cognitive phenomena can be accessed through introspection and, individual experience of meaning or grammaticality is ultimately not well suited to the analysis of historical language-use, where the language exists only as textual remnants independent of the minds that would have interpreted them.

1c. Corpus Methodology

A linguistic corpus enables linguists "to make more objective and confident descriptions of usage than would be possible through introspection. It allows them to make statements about frequency of usage in the language as a whole, as well as comparative statements about usage in different varieties. It permits them, in principle, to arrive at a total account of the linguistic features in any of the texts contained in the corpus. And it provides them with a source of hypotheses about the way the language works." (Crystal, 2003)

Google Ngrams, though not quite the typical corpus, provides an example of the methodology that is easily replicable online at https://books.google.com/ngrams.

Google Ngrams takes its data from the immense corpus of *Google Books*. This can be used to visualise diachronic word use, that is to say the incidence of particular words within the corpus over a period of time.

Comparing the frequency of the forms <colour> and <color> in British and American data, it may not surprise us that as of 2012 the spelling <colour> is more common in BrE and <color> in AmE. But using the corpus, we can also look back in time and pinpoint with some accuracy when the spellings <color> became more common in AmE (the mid 1840s) and note that this coincides with the publication and spread of the first editions of Webster's Dictionary.

While the importance of this particular finding may be more important for linguists than historians, it demonstrates the power of corpus linguistics to show how language has changed over time based on large data sets. Similarly, the methodology can be used to show changes in grammar (morphology or syntax), lexis (how new words and concepts are introduced into language) or shifts in discourse and genres.

Co-funded by the
Erasmus+ Programme
of the European Union

2 How to Create Editions for Corpus Linguistics

2a. Two examples of Digital Corpora

Google ngrams is a huge, diachronic corpus in a large range of languages and dialects. Its American-English data set consists of 155 billion words, its British data set of 34 billion, its Spanish data set contains 45 billion and so on and so forth.

Penn-Helsinki Parsed Corpus of Middle English, 2nd edition, (PPCME2), is another diachronic corpus, this time of 56 texts from the Middle English section of the diachronic Part of the Helsinki Corpus of English Texts (Currently available in XML at http://www.helsinki.fi/varieng/CoRD/corpora/PPCME2/), with a number of deletions and additions. (The full list of texts, arranged by date, is currently available at http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCME2-RELEASE-4/index.html) . The corpus consists of about 1.2 million words. These are arranged by date into periods ca. 70 years.

2b. Problems illustrated by these examples.

Though both corpora are impressive in scale, there are criticisms to be made of their representativeness and their reliability. In the case of *Google ngrams*, the scale of the project is itself is the cause of its most serious problems. Since the corpus contains only one of each book, it represents only what is written and not what is read. It's results are therefore only tenuously representative of language usage. Scholarly publishing output, for example, is overrepresented in the *Google ngrams* corpus. There are also reliability issues arising from incorrect dating and categorization. The 'optical character recognition' (OCR) technology employed in digitizing the corpus is imperfect. Some characters are misidentified and the effect of such small errors on reliability are magnified across the corpora.

PPCME2 has a different set of problems. The key issue lies with text selection and the overriding practical constraints under which the corpus was compiled. In the first place, the corpus is composed of long texts in small samples. To make the task of producing such a corpus feasible, the number of texts incorporated was limited and their representativeness determined subjectively. Ideally such a corpus would aim for a greater number of texts in each of its periods. In addition, the editions used as to compile this corpus were created by non-linguists, whose interests were quite different.

Co-funded by the
Erasmus+ Programme
of the European Union

2c. The Problems of using Critical Editions as the basis for digital Corpora

Representativeness in corpus composition is of absolute importance for compilers of linguistic corpora. Yet, for the editors of critical editions it is often either less significant or interpreted differently. Linguists, unlike editors, are not necessarily interested in literary, culturally, historically or politically significant texts. Such texts as these tend to be overrepresented in the editions upon which corpora rely.

Moreover, critical editions often represent more the editorial choices than the language of the manuscript. In an effort to uncover the original text/authors voice, critical editions often reconstruct an ‘authorial text’ in a way not supported by the manuscript evidence. This is done by the compilation of the best readings from number of manuscripts. Such a method can lead (and has lead) to circular logic. In these cases, the emended language (*parole*) of one text is used to re-construct a grammar/standard (*langue*) that is later used as a model to emend other texts.

Linguists, by contrast, are not interested in the voice of the author (since that is conjectural), but in the voice of the manuscript, i.e. the scribe. This is problematic in itself, yet it is more useful from the standpoint of trying to establish historical word-use.

2d. How to Work with Multiple versions of one Text

Furthermore, many linguistic corpora (such as PPCME2) use only one version of the text. Yet, historical linguists are primarily interested in diachronic change, which is usually the result of synchronic variation. For this reason, multiple versions of a given text are interesting but critical compilations are not. Historical linguists are looking for a representative reading and not the ‘best reading’.

Multiple copies may skew the results, but they should to be edited and included in the corpus, if they are properly tagged, so that the technology can account for the text’s multiplicity.

Co-funded by the Erasmus+ Programme of the European Union

2g The Basic Principles of a Good Corpus

- Maximal information preservation
- No irreversible editorial changes
- Explicit indication of responsibility
- Reliable documentation
- Maximal flexibility/accessibility

2f. Editorial Features of a Good Corpus

Editions produced for corpus linguistics should go beyond what is typically the norm for traditional diplomatic editions. There should be no irreversible emendation or normalisation of the text. Accurate *graphemic* representation should be maintained at all times (i.e. the lettering and wording should be encoded as they appear in the source manuscript).

However, even in diplomatic editions, some features are often omitted or silently normalised. Allographs are often silently normalised so as not to cause problems with search/retrieval function of a corpus. So, for example, a corpus might normalise the non-standard graphemes found in insular manuscripts thus: s=ʃ=ɼ?; ƿ=w?; ð=þ=th?; a=ɑ?. If normalisation occurs, it should always be annotated.

When creating the corpus, the editor has to decide what to do with the original punctuation, some of which may not correspond to the modern punctuation schemes. They must also choose whether to annotate the abbreviations, which are to be found in practically all Latin manuscripts. Are paratexts to be differentiated? Paratexts such as headings, paragraphs, columns, rubrics etc. are frequently omitted in corpora but at other times (for example when encoding poetry), they are normalised. Non linear text is also a cause for concern.

Ultimately there are lots of approaches to these questions, rather than one standard to be followed by all. Though there are better and worse ways to organise a corpus, the key thing is to make sure that, whatever you do, the edition includes documentation of all the editorial principles and decisions and of the source material.

Co-funded by the
Erasmus+ Programme
of the European Union

2g. Technical Features of a Good Corpus

A good corpus needs to be compatible and comparable with other corpora. This means that the depth of detail encoded needs to be similar, i.e. it needs to be based on more or less diplomatic transcriptions. The mark-up should, ideally, conform to a standard, e.g. TEI.

There should be little or preferably no use of proprietary software or interfaces. The use of such software often limits the availability of the original data, which should instead be accessible to all and compatible with freely available software.

Legally, the copyright status of the edition should be explicit and consistent. If there are any parts of the edition that carry a separate license, these should be made separate so as to protect both the editors and users.

2h. *WebAnno*

An excellent tool for collaborative tagging editions is *WebAnno* (https://webanno.github.io). This is a multi-user annotation tool that is designed to support large projects that involve numerous people in different roles. It is especially useful therefore for crowdsourcing. It is free and fully web based, so there is no need to buy or install anything. The results are therefore accessible to anyone with a computer. Furthermore *WebAnno* is TEI compatible so an edition produced using it will conform to a standard.

For a full set of Tutorials, please see:

(https://www.youtube.com/watch?v=sQ5pFoFzxIk&list=PLvYKmi8P7TYdC-7A_VT4td95629aZIwDb&index=2).

Co-funded by the
Erasmus+ Programme
of the European Union

3 Examples of the use of corpus linguistics

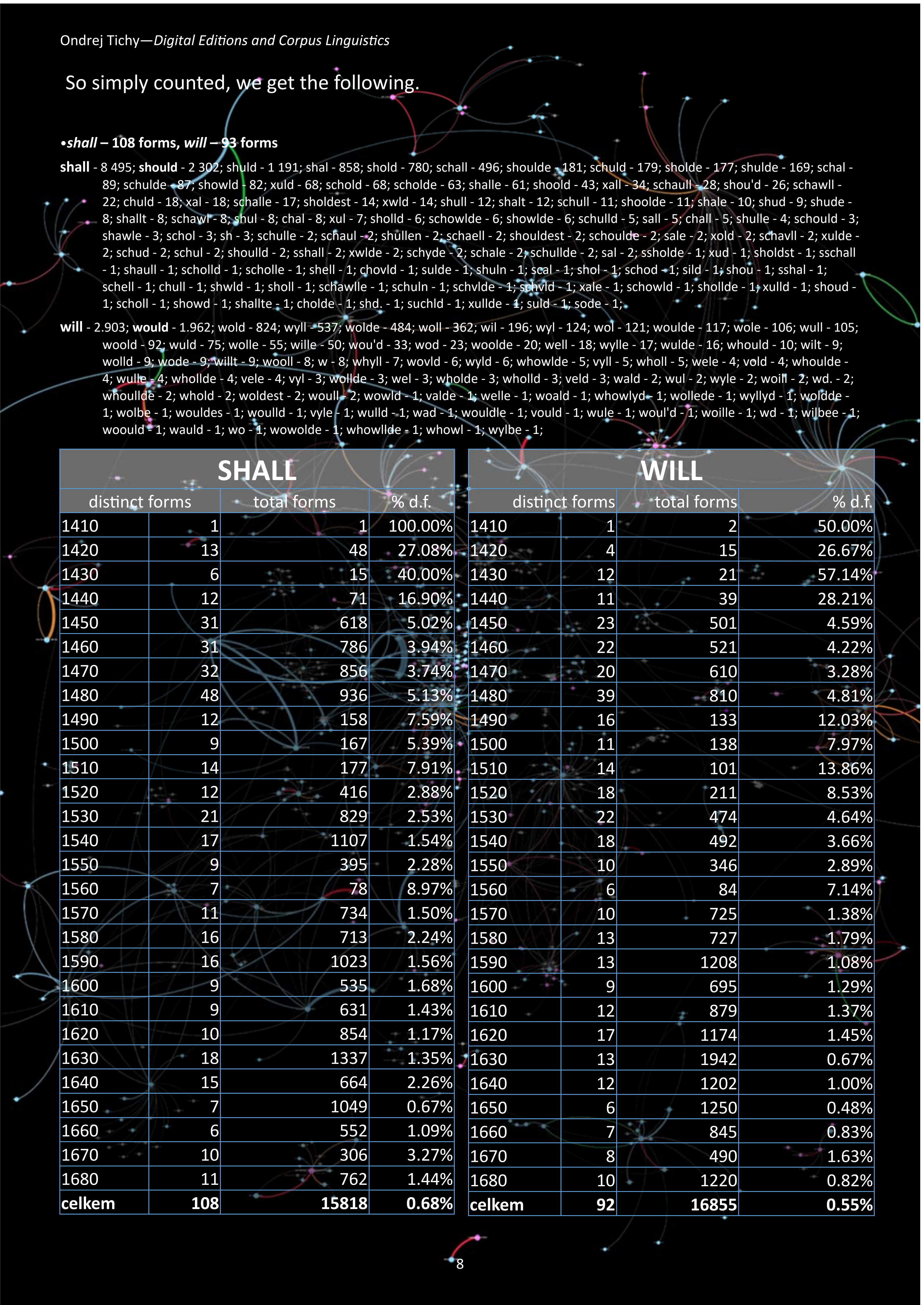
3a. CEEC / PCEEC—An example of Diachronic Corpus Research

The Corpus of Early English Correspondence is a corpus that was created at the University of Helsinki. The corpus includes letters written between 1410 and 1680. This means that it has 4970 letters from 84 collections, by 666 authors and fully 2,159,132 [2.7 million] words or 'tokens' within the corpus. This has been marked up with meta-data including: year, authenticity, author, addressee, with their gender, age and relationship (social class and region/dialect remain as yet unpublished). The words have not been lemmatised.

The PCEEC was an evolution of the original database. This version contains the bulk of the collections [2.2 million words], has part of speech tagging and is parsed in dependency syntax. The corpus itself is available through the Oxford text archive.

This corpus , for example, may be used to show the degree to which the spelling of a particular word is standardised. In our example, we will use the words 'shall' and 'will', because they are tagged separately and it it is easy to find all their forms. Simply counting the number of wordforms for 'shall' and 'will' at a given time would provide a very crude measure of language use. However, we can use the information contained within the corpus to produce more precise statistics. We may, for example, calculate how predictable a particular spelling was at a given time. That is to say, that rather than simply illustrating the number of variants, we may extrapolate from the corpus an idea of the regularisation of letters and words within the language. By attempting to produce 'objective' methods for historical linguistics, scholar will often find themselves forced to ask more specific questions of the data and therefore to re-evaluate and reformulate their aims.

Co-funded by the Erasmus+ Programme of the European Union



So simply counted, we get the following.

- ***shall* – 108 forms, *will* – 93 forms**

shall - 8 495; **should** - 2 302; shuld - 1 191; shal - 858; shold - 780; schall - 496; shoulde - 181; schuld - 179; sholde - 177; shulde - 169; schal - 89; schulde - 87; showld - 82; xuld - 68; schold - 68; scholde - 63; shalle - 61; shoold - 43; xall - 34; schaull - 28; shou'd - 26; schawll - 22; chuld - 18; xal - 18; schalle - 17; sholdest - 14; xwld - 14; shull - 12; shalt - 12; schull - 11; shoolde - 11; shale - 10; shud - 9; shude - 8; shallt - 8; schawl - 8; shul - 8; chal - 8; xul - 7; sholld - 6; schowlde - 6; showlde - 6; schulld - 5; sall - 5; chall - 5; shulle - 4; schould - 3; shawle - 3; schol - 3; sh - 3; schulle - 2; schaul - 2; shullen - 2; schaell - 2; shouldest - 2; schoulde - 2; sale - 2; xold - 2; schavll - 2; xulde - 2; schud - 2; schul - 2; shoulld - 2; sshall - 2; xwlde - 2; schyde - 2; schale - 2; schullde - 2; sal - 2; ssholde - 1; xud - 1; sholdst - 1; sschall - 1; shaull - 1; scholld - 1; scholle - 1; shell - 1; chovld - 1; sulde - 1; shuln - 1; scal - 1; shol - 1; schod - 1; sild - 1; shou - 1; sshal - 1; schell - 1; chull - 1; shwld - 1; sholl - 1; schawlle - 1; schuln - 1; schvlde - 1; schvld - 1; xale - 1; schowld - 1; shollde - 1; xulld - 1; shoud - 1; scholl - 1; showd - 1; shallte - 1; cholde - 1; shd. - 1; suchld - 1; xullde - 1; suld - 1; sode - 1;

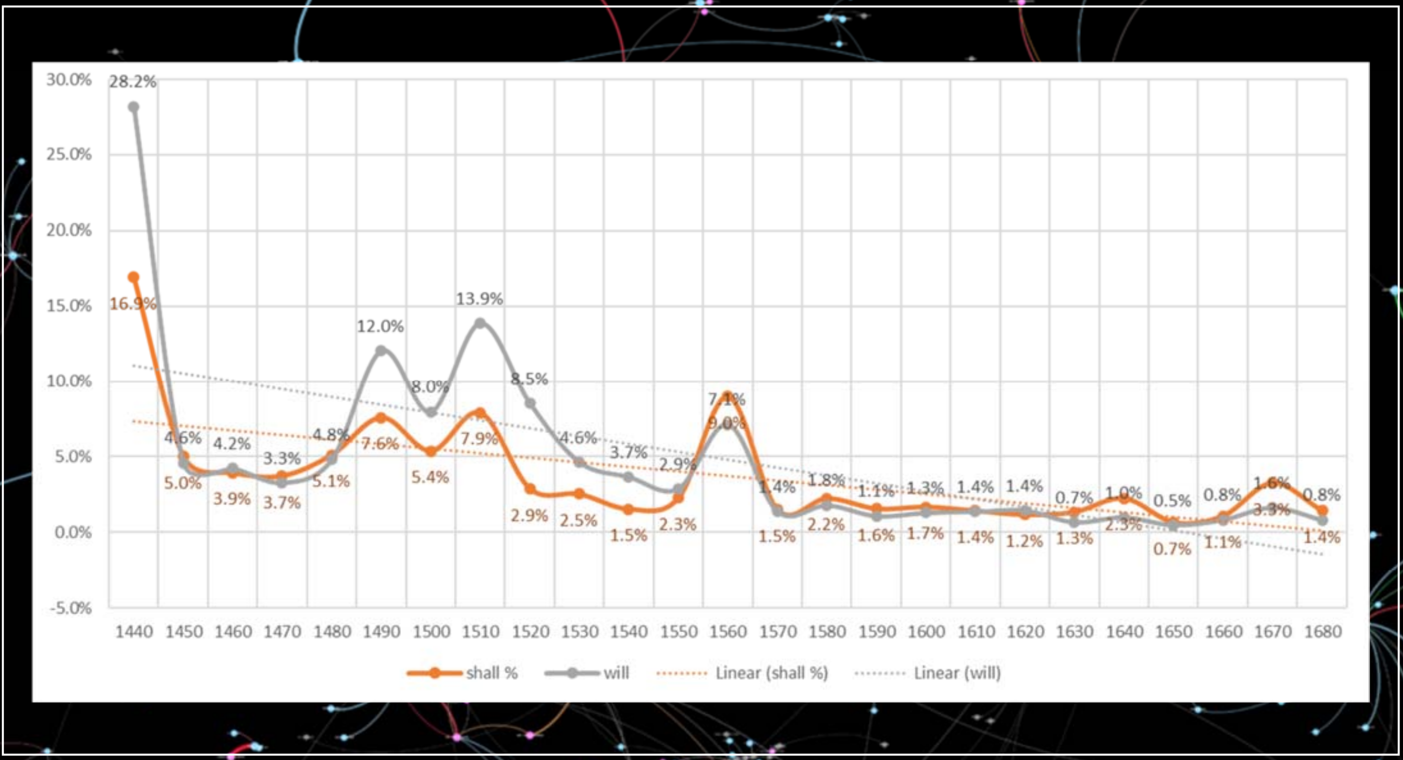
will - 2.903; **would** - 1.962; wold - 824; wyll - 537; wolde - 484; woll - 362; wil - 196; wyl - 124; wol - 121; woulde - 117; wole - 106; wull - 105; woold - 92; wuld - 75; wolle - 55; wille - 50; wou'd - 33; wod - 23; woolde - 20; well - 18; wylle - 17; wulde - 16; whould - 10; wilt - 9; wolld - 9; wode - 9; willt - 9; wooll - 8; w - 8; whyll - 7; wovld - 6; wyld - 6; whowlde - 5; vyll - 5; wholl - 5; wele - 4; vold - 4; whoulde - 4; wulle - 4; whollde - 4; vele - 4; vyl - 3; wollde - 3; wel - 3; wholde - 3; wholld - 3; veld - 3; wald - 2; wul - 2; wyle - 2; woill - 2; wd. - 2; whoullde - 2; whold - 2; woldest - 2; woull - 2; wowld - 1; valde - 1; welle - 1; woald - 1; whowlyd - 1; wollede - 1; wyllyd - 1; woldde - 1; wolbe - 1; wouldes - 1; woulld - 1; vyle - 1; wulld - 1; wad - 1; wouldle - 1; vould - 1; wule - 1; woul'd - 1; woille - 1; wd - 1; wilbee - 1; woould - 1; wauld - 1; wo - 1; wowolde - 1; whowllde - 1; whowl - 1; wylbe - 1;

SHALL			
distinct forms		total forms	% d.f.
1410	1	1	100.00%
1420	13	48	27.08%
1430	6	15	40.00%
1440	12	71	16.90%
1450	31	618	5.02%
1460	31	786	3.94%
1470	32	856	3.74%
1480	48	936	5.13%
1490	12	158	7.59%
1500	9	167	5.39%
1510	14	177	7.91%
1520	12	416	2.88%
1530	21	829	2.53%
1540	17	1107	1.54%
1550	9	395	2.28%
1560	7	78	8.97%
1570	11	734	1.50%
1580	16	713	2.24%
1590	16	1023	1.56%
1600	9	535	1.68%
1610	9	631	1.43%
1620	10	854	1.17%
1630	18	1337	1.35%
1640	15	664	2.26%
1650	7	1049	0.67%
1660	6	552	1.09%
1670	10	306	3.27%
1680	11	762	1.44%
celkem	**108**	**15818**	**0.68%**

WILL			
distinct forms		total forms	% d.f.
1410	1	2	50.00%
1420	4	15	26.67%
1430	12	21	57.14%
1440	11	39	28.21%
1450	23	501	4.59%
1460	22	521	4.22%
1470	20	610	3.28%
1480	39	810	4.81%
1490	16	133	12.03%
1500	11	138	7.97%
1510	14	101	13.86%
1520	18	211	8.53%
1530	22	474	4.64%
1540	18	492	3.66%
1550	10	346	2.89%
1560	6	84	7.14%
1570	10	725	1.38%
1580	13	727	1.79%
1590	13	1208	1.08%
1600	9	695	1.29%
1610	12	879	1.37%
1620	17	1174	1.45%
1630	13	1942	0.67%
1640	12	1202	1.00%
1650	6	1250	0.48%
1660	7	845	0.83%
1670	8	490	1.63%
1680	10	1220	0.82%
celkem	**92**	**16855**	**0.55%**

Co-funded by the
Erasmus+ Programme
of the European Union

The Percentage of distinct forms can be plotted thus

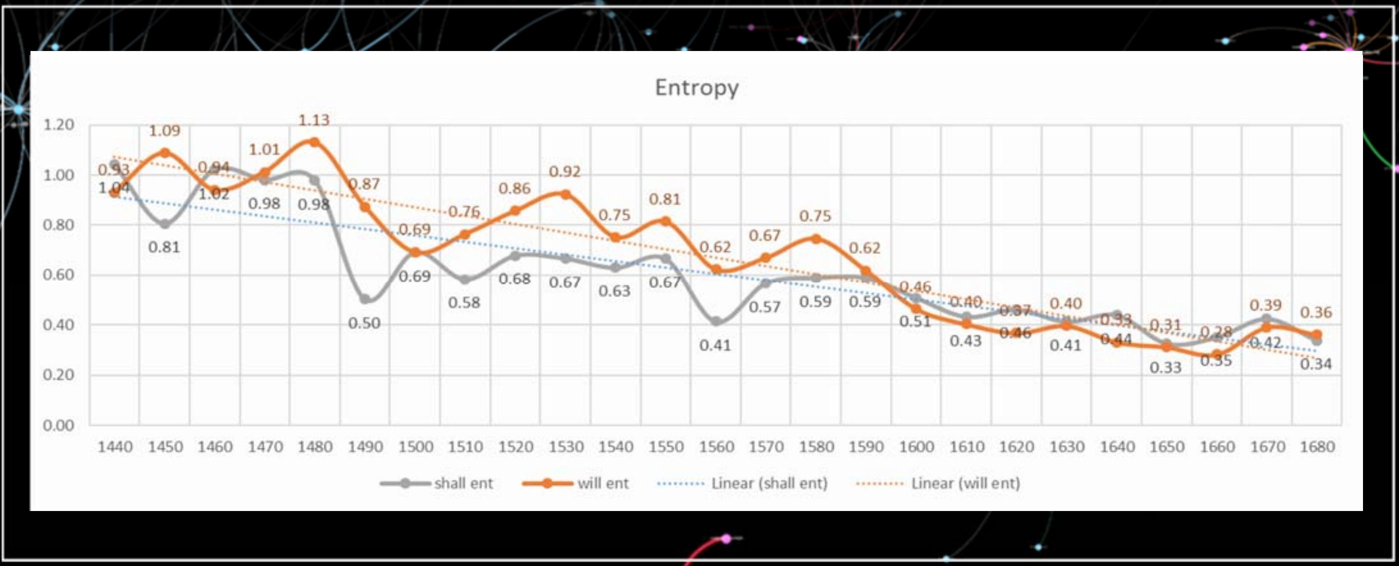


Using Shannon's Entropy, equation, however, we can measure predictability of different spellings based on this corpus. In so doing, the corpus can be used to draw specific conclusions about the standardization of language use at these times. Shannon's entropy is a logarithm of 'probability distribution'. That is to say that it can be used to calculate the probability of various possible outcomes in an experiment. In this case it predicts the likelihood of any given spelling at any given time.

$$- \sum (d/f \times \ln (d/f))$$

Here, d = the number of distinctive forms and f = the total number of forms for each period/section.

This gives us the graph below, which presents us with the surprising conclusion that the predictability of different spellings of 'shall' and 'will' actually declined between 1440 and 1680.



Co-funded by the
Erasmus+ Programme
of the European Union

Bibliography

•Auer, Anita et al. (2016) *English Urban Vernaculars, 1400–1700: Digitizing Text from Manuscript in Corpus linguistics on the move*. ed. López-Couso et al. Brill

•Bastian M., Heymann S., Jacomy M. (2009) Gephi*: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media.

•Benskin, Michael, Margaret Laing et al. (2013). *An Electronic Version of A Linguistic Atlas of Late Mediaeval English*. University of Edinburgh. Available at: http://www.lel.ed.ac.uk/ihd/elalme/elalme.html.

•Crystal, David. The Cambridge Encyclopaedia of the English language. 2nd ed. New York: Cambridge University Press, 2003.

•Kroch, Anthony, Beatrice Santorini, and Lauren Delfs (2004) The *Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3).

•Kroch, Anthony, and Ann Taylor (2000). The *Penn-Helsinki Parsed Corpus of Middle English (PPCME2).* Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4 (http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4).

•Laing, Margaretand Roger Lass. (2013). LAEME: *A Linguistic Atlas of Early Middle English*. http://www.lel.ed.ac.uk/ihd/laeme2/laeme2_framesZ.html

•Lass, Roger. (2004). 'Ut custodiant litteras: Editions, corpora and witnesshood'. In Marina Dossena and Roger Lass (eds.), *Methods and Data in English Historical Dialectology*. Bern: Peter Lang, 21–48.

•Marttila, Ville (2014) *Creating Digital Editions for Corpus Linguistics, The case of Potage Dyvers, a family of six Middle English recipe collections*. University of Helsinki

•*Middle English Dictionary* (2002), ed. Francis McSparren. Michigan: University of Michigan.

•Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative *Analysis of Culture Using Millions of Digitized Books*. Science (Published online ahead of print: 12/16/2010)

•Nevalainen, Terttu; Helena Raumolin-Brunberg (1996) *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence. In Language and computers : studies in practical linguistics*, Rodopi

•*OED Online*. (2016) Oxford University Press. Web. 7 December 2016.

•*Parsed Corpus of Early English Correspondence, tagged version*. (2006) Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki.

•Rissanen, Matti; Merja Kytö and Minna Palander (1993), Early *English in the computer age: explorations through the Helsinki Corpus*. Berlin - New York: Mouton de Gruyter