Brendan D. Ferreri-Hanberry. Application of a POS Tagger to a Novel Chronological Division of Early Modern German Text. A Master's Paper for the M.S. in I.S. Degree. 106 pages. August, 2015. Advisor: Stephanie Haas.

This paper describes the application of a part-of-speech tagger to a particular configuration of historical German documents. Most natural language processing (NLP) is done on contemporary documents, and historical documents can present difficulties for these tools. I compared the performance of a single high-quality tagger on two stages of historical German (Early Modern German) materials. I used the TnT (Trigrams 'n' Tags) tagger, a probabilistic tagger developed by Thorsten Brants in a 2000 paper. I applied this tagger to two subcorpora which I derived from the University of Manchester's GerManC corpus, divided by date of creation of the original document, with each one used for both training and testing. I found that the earlier half, from a period with greater variability in the language, was significantly more difficult to tag correctly. The broader tag categories of punctuation and "other" were overrepresented in the errors.

Headings:

Natural language processing

Languages – German language

18th century materials

17th century materials

APPLICATION OF A POS TAGGER TO A NOVEL CHRONOLOGICAL DIVISION OF EARLY MODERN GERMAN TEXT

by
Brendan D. Ferreri-Hanberry

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

August 2015

Approved by

_____

Dr. Stephanie W. Haas

1. Introduction

Part-of-speech (POS) taggers are tools which assign part-of-speech tags to a variety of words, and are important for more than one reason. Brants (2000a, p. 224) cites two related categories of applications of part-of-speech (POS) taggers. The first is as a preprocessor, that is to send the tagger's output directly to another NLP tool. Along similar lines, the POS tagger can be used for corpus annotation projects. These are projects which involve both automatic annotation of corpora and human correction of these annotations, and can be used as resources for a variety of future projects by others.

More specific applications of the proper interpretation of parts of speech include being able to trace changes in syntax, morphology, spelling and semantics over time, which is of interest to researchers (Hauser et al., 2007, p. 3). "Basic language processing'' including POS tagging is necessary for most forms of natural language processing, "such as Machine Translation, Summarization, Dialogue systems, etc." (Carreras et al, 2004, p. 23). Another practical application of POS tagging is to assist in running search engines over historical documents. This is important in connection with recent European and international digitization projects, as these are dramatically increasing the amount of historical text available online in various languages. Recent and ongoing projects include the Open Content Alliance; Google Books, founded in 2002, which is the largest book digitization project ever (Piotrowski, 2012, p.26) and now offers a search interface in at least 35 languages; Project Gutenberg, founded in 1971, which offers materials in English, French, German and Portuguese; HathiTrust, a coalition of over 100 research libraries; the European digital library project; Wikisource, founded in 2003 and run by the Wikimedia foundation; and the Text Creation Partnership, begun in 1999, which is responsible for Early English Books Online and Eighteenth Century Collections Online (Hauser et al., 2007, p. 1; Rayson et al., 2007, p.

2; Piotrowski, 2012, p. 26). Project Gutenberg-DE, also known as the German Gutenberg Project (Hinrichs & Zastrow 2012), for example, has so far made over 5,500 German-language texts by over 1,100 authors available. Early English Books online contains over 125,000 titles. A single repository may well contain contemporary as well as historical documents, in which case the contemporary documents would not suffer the same access issues as the historical documents. However, users who are searching over a repository including historical documents will most likely use modern spellings in their queries and be unfamiliar with the historical spellings of those terms, where they differ from the contemporary ones. This will greatly limit their access to historical documents. To address this, there are three basic approaches cited by Hauser et al. (2007, p. 3-4).

One approach is to use special dictionaries, which connect a list of historical variants with each modern word. Another approach is rule-based generative matching, which does not use a dictionary but instead applies rules to either historical or contemporary terms (Ernst-Gerlach & Fuhr, 2006; Pilz et al., 2009). Some of these rules may change based on the time and place of the document's creation. The online variant applies rules to input contemporary terms in order to generate possible matching historical terms, while the offline variant attempts to normalize historical variants as part of indexing, i.e. to index them to modern terms. A third option is matching based on various means of calculation of word similarity or distance such as Levenstein distance, which can even be used to correct erroneous input (Schnell and Bachteler, 2004, Schulz and Mihov, 2002; Mihov and Schulz, 2004; Strunk, 2003).

POS taggers and other NLP tools designed to work on modern languages face challenges when applied to historical forms of these languages. This paper evaluates the performance of a tagger on German texts from two historical periods.

2. Literature Review

2.1. Modern and Historical German

German was of interest largely because it was the subject of my undergraduate degree and is thus a language with which I am familiar. It is also one of the dominant languages in the world, widely used in business, science, literature and philosophy, with approximately 90 million native speakers. Standard German is the current official language of Germany, Austria, and Lichtenstein as well as being an official language in Switzerland, Luxembourg, and Belgium. (Eisenberg, 1994, p. 349). It has a rich written history dating back many centuries. However, due to a relatively late official standardization of the language, many historical German documents are not as comprehensible to modern readers or researchers as contemporary German. Due to syntactic, vocabulary and spelling differences, they are also not as accessible to various natural language processing (NLP) tools for various purposes such as machine translation, running queries, and tracking linguistic change over time.

German is also of interest in that it is a moderately inflected language, meaning that many of its words are subject to changes in their form depending on grammatical conditions (Moeller et al., 2007). In contrast, most other members of the Germanic branch of the Indo-European language family, especially English, have very little inflection (Clark, 1957, p. 36). Beyond English, this language branch includes several official national languages, including Dutch (Willemyns, 2013, Donaldson, 2008) and three largely mutually intelligible Scandinavian languages, namely Danish, Norwegian and Swedish (Henrisken & van der Auwera, 1994 p.3). Lesser-known Germanic languages include Frisian and Afrikaans, both similar to Dutch; Faroese, spoken on the Faroe Islands in the North Atlantic and similar to Icelandic; and two descended from German, namely Yiddish, influenced by Hebrew and Aramaic as well as by Slavic

languages, and Pennsylvania German. The latter, also known as Pennsylvania Dutch, is spoken by about 300,000 descendants of German immigrants to the US including some Mennonites and Amish. It is officially considered a language rather than a German dialect due to having undergone "spontaneous uniformation (dialect levelling) and some standardization efforts," and the fact "that it marginally also functions for written communication" (ibid., p. 12-13). An exception to the general low level of inflection among Germanic languages today is Icelandic, "the most conservative of the Scandinavian languages," (ibid., p. 7) which is highly inflected and has remained very similar in vocabulary and grammar over centuries (Neijmann, 2001; Zoëga, 1926; Einarsson, 1945, p. 32-45; Thráinsson, 1994). Due to this, POS tagging on Icelandic is particularly complex, employing a standard tagset of about 660 tags (Pind et al., 1991), as opposed to the 54 used in the standard tagset for German (Brants 2000b; Loftsson, 2006, p. 176; Loftsson, 2008, p. 49). Accuracy figures are also considerably lower for taggers run on Icelandic corpora than some related languages (Loftsson, 2007; Loftsson, 2008), including not only English but also Dutch (Poel & Boschman, 2008). English also employed more inflection in its earlier periods, namely Middle English (c. 1100 – c. 1450 A.D.) and especially Old English (c. 700 - c. 1100 A.D.) (Clark, 1957, p. 26-27). Middle English had significantly less inflection than Old English, with case and gender "vestigial and absent, respectively" (McWhorter, 2008, p. 39). This simplification has been attributed to the influence of the invading Vikings, who were illiterate and attempted to learn Old English as adults, and thus would have been inclined to misinterpret or ignore some of the subtleties of their new language (McWhorter, 2008, pp. 92-94, 109-117). After the era of Middle English, Early Modern English (c. 1450-1600) inflected "exceptionally few words" and "in an exceptionally small number of ways" (Clark, 1957, p. 84). The English used today has "fewer inflexions than all but one or two of the languages related to it" (Clark, 1957, p. 26); it is the only Germanic language currently in use in which nouns do not require genders (McWhorter 95-97).

Other Germanic languages also experienced morphological simplification over time, including Danish which now has no cases for noun inflection (Haberland, 1994, p. 323), Norwegian for which the same is true with the possible exception of a genitive s (Askedal, 1994, p. 220), and Swedish which inflects nouns for only two cases

(Andersson, 1994, p. 279). In all of these standard languages, verbs are no longer inflected to express number, but all three of these are descended from the more highly inflected Old Scandinavian (Faarlund, 1994, p. 38). Similarly, while Middle Dutch had "a two-declension, three-case, four-gender system" for nouns (van der Wal & Quak, 1994, p. 75), the modern standard language retains very little of this. Some limited use of inflection, though limited to the written language, survived up to 1947. In this year, however, with the exception of the set expressions noted, these last vestiges were officially abolished with an official spelling reform (Donaldson, 2008, 23), with the exception of "a few set expressions" and limited use of a genitive s, somewhat like that used in English (DeSchutter, 1994, p. 459).

Contemporary German's inflection has also progressively simplified over the centuries (Hauser et al., 2007, p. 3). However, it not only retains three grammatical genders (masculine, feminine and neuter) of which every noun must have one, but also retains Old High German's four grammatical cases (nominative, accusative, dative and genitive) (Chambers & Wilkie, 1970, p. 119), although their endings have simplified somewhat since the OHG period (ibid., p. 121). These apply to some nouns and to all nouns' corresponding articles, adjectives, and pronouns (Moeller et al., 2007). In these respects it is somewhat similar to modern Icelandic and Faroese (Barnes & Weyhe, 1994, p. 198). It is also inflected for number, i.e. the difference between singular and plural, which is expressed in a greater variety of ways in German than in English (Moeller et al., 2007). Between German and English, such differences mean that NLP tools will need to recognize a greater variety of forms of a given word for German, and that tools developed for one will not be very reliable for the other.

German, both historically and today, is divided into a variety of dialects, meaning that there is variation between varieties of both the spoken and written language. The larger category of High German (*Hochdeutsch*), which will be the focus of this project, is distinguished, among other factors, by having undergone a consonant shift "probably from the fifth or sixth centuries onward" (Willemyns, 2013, p. 34). Low German (*Plattdeutsch*), by contrast, spoken in the Northern Lowland, did not undergo this consonant shift, and neither did other Germanic languages (Chambers & Wilkie, 1970, p. 112). Several consonants such as p, t and k were affected in various ways; for example,

the unaffected Old English's "open" was equivalent to Old High German's "offan" and "pund" in OE was "pfunt" in OHG. (p. 112-117). Upper German dialects, a subset of High German, "were (and are still) spoken in the Southern part of Germany, Switzerland, and Austria" (Dipper, 2010, p. 118). These include Alemannic, for example, which includes Swiss German and is spoken mainly in Switzerland. Swiss German is a significantly different language from Standard German, technically a dialect continuum due to its diversity and lack of official standards. Due to it not being a single standardized language, there is relatively little literature relevant to this dialect continuum, although in recent years efforts have been made at POS tagging (Hoellenstein & Aepli, 2014), dialect identification (Scherrer and Owen, 2010), dialect machine translation (Scherrer, 2012) and morphology generation. (Scherrer, 2011).

The term Middle German has more than one meaning in linguistics, but can be used (along with "Central German") to refer to dialects formerly spoken in the central regions of Germany. These are also classified under the larger category of High German. This variety was influential on Standard German (*Standarddeutsch* or *Hochdeutsch*), as explained in more detail below.

The historical progression of the German language is, briefly, as follows. The earliest known German language is known as Old High German (OHG), dating from approximately 700 or 770 to 1100 AD (Chambers & Wilkie, 1970, p. 31; Hauser et al., 2007, p. 2). The use of the written language was very limited during this time, and there are only approximately 70 extant examples of documents from this time period. During this period, the written language was the province of the clergy, and most OHG texts were based on Latin texts (Hauser et al., 2007, p. 2; Chambers & Wilkie, 1970, p. 31). The language of this period may seem difficult to modern German speakers, but was "much simpler than its Indo-European or Primitive Germanic ancestors" as there had been a "drastic reduction in inflexional forms" regarding declensions and conjugations (Chambers & Wilkie, 1970, p. 31). The same was true of the closely related Old Dutch, which was similarly "a language of monks and clerics" as far as the written language was concerned (Willemyns 2013, p. 45).

Subsequent to OHG came the period known as Middle High German (MHG), dating from 1100 to 1350 AD. During this period, use of the written language increased

somewhat in importance as the set of authors was expanded from the clergy to include the aristocracy as well. The variety of text genres was thus expanded from OHG's merely liturgical or theological works; this period displayed a "magnificent flowering of chilvalric and courtly poetry" reaching its peak at about 1200 (Chambers & Wilkie, 1970, p. 34). The first approximation of a standard German language occurred during this period; there developed "a literary language which was more widely used than any single dialect and which enjoyed a high cultural prestige," being used by the aristocracy of all German-speaking regions (Chambers & Wilkie, 1970, pp. 34-35). During this and the previous period, texts were produced which would seem quite unfamiliar to a modern reader for multiple reasons. For example, before the ENHG period, "the syntactical principle of punctuation" did not exist, so the syntactical structure was "not very pronounced" (Hauser et al., 2007, p. 3).

The period immediately following this is known as Early New (High) German (ENHG). ENHG covers a period which begins at a point which is not entirely agreed upon, but which ranges between 1350 and 1450. The end point of this period has been set at 1600, 1650 or even 1700 (Hauser et al., 2007, p. 2; Piotrowski, 2012, p. 112; Universität Duisburg-Essen). It was during this period that the printing press was invented, leading to a major increase in the number and variety of texts produced. Texts from this period, in contrast to modern High German, display "large dialectical variance" (Hauser et al., 2007, p. 2). In fact, like all the previous periods of High German, ENHG is "not a language in the modern sense" due to the lack of any standards which applied across all of the regions in which it was spoken (Piotrowski, 2012, p. 86). The language during the later part of this period, though, as well as the modern standard language, was strongly influenced by Martin Luther's translation of the Bible, completed in 1534. This work closely followed a Central German dialect known as East Middle German or Upper Saxon, and Luther stated that he intended to follow the speech of the common people rather than the aristocracy with this text, thus diverging from both written German of previous eras and earlier versions of the Bible from the same era (Chambers & Wilkie, 1970, p. 41). Due to the increasing use of words from foreign languages such as Latin, ENHG has been described as having "a vocabulary which is extraordinarily comprehensive as compared to other historical eras" (Hauser et al., 2007, p. 3). An

example of the contrast between ENHG and contemporary German, used by Piotrowski (2012) is as follows. This is an excerpt from a court record of a 16th-century defamation case.

> Erstlich hatt Wolfgang Lippuner, landtweibel zu Sarganß, vor einem ersamen gricht zu Sarganß herren Jacoben, caplonen zu Sarganß, anclagt, das er ime am palms abenndt sin meitli altem bruch nach geschick zu bychten. (Piotrowski 2012, pp. 20)

A modern German translation is as follows:

> Als erstes hat Wolfgang Lippuner, Landweibel in Sargans, vor dem ehrwürdigen Gericht in Sargans Herrn Jacob, Kaplan in Sargans, angeklagt [und berichtet], dass er ihm am Tag vor dem Palmsonntag sein Mädchen gemäß altem Brauch zum Beichten geschickt habe. (Piotrowski 2012, pp. 22)

The first efforts at deliberate standardization of a single German written language began in the late 16th century, in other words during the ENHG period. However, "the first influential figure belongs to the early seventeenth century" (Chambers & Wilkie, 1970, p. 47). Martin Opitz (1597-1639) was known partly for his work Buch von der deutschen Poeterey (1624), which "set the subject matter, metre, language, and style" for poetry in his native language for more than a century afterward (ibid.). During the next several decades various organizations known as Sprachgesellschaften (language societies) were founded which, among many other cultural and political aims, "advocated a unified language (with a unified pronunciation)" (ibid., p. 48). These included the Deutschgesinnte Genossenschaft of Hamburg (1642), the Hirten – und Blumenorden an der Pegnitz of Nuremberg (1644) and the Elbschwanenorden of Lübeck (1648).

The timeline after ENHG next includes a period from approximately 1650 to 1800 known as Early Modern German (EMG) (Scheible et al. 2011a, 19) while other sources describes the period from 1600-1800 as the first period of a larger period stretching until the present day and entitled New High German (NHG) (Hauser et al, 2007, p. 2). The beginning of New High German may also be placed at 1650 (Piotrowski, 2012, p. 19).

Texts from this period differ notably from modern German, although often less dramatically so than in the case of ENHG. Whether EMG texts are easily comprehensible to a modern reader or how effective modern NLP tools are for use on them depends on the source. Gotscharek et al. (2009) note that "the movement from old and arbitrary

spelling to modern and normalized spelling did not have the same speed for all text genres" (p. 196). To use an example given by Piotrowski (2012), a modern German spellchecker was run over two documents from late in this period, one from 1784 and one from 1786 (p. 18-19). The earlier document was the famous essay by German Enlightenment philosopher Immanuel Kant, *Was ist Aufklärung?* while the later text was a draft for a wine tax by the city council of Rapperswil, Switzerland, transcribed from a manuscript. While only 6% of tokens were interpreted as errors in Kant's essay, the corresponding figure for the Swiss manuscript was a full 45%. The latter contains numerous oddities of spelling not present in the former, such as "ey" in place of the modern "ei" or sß where in modern language ß or ss would be used, as well as obsolete words such as *verumbgeltet* and *pottmäßigkeit* (ibid., p. 18).

Standardization efforts continued during this period, although as the above evidence shows, they were not universally accepted throughout German-speaking Europe. Justus Georg Schottel "encouraged the acceptance of unified rules of spelling" in his *Teutsche Sprachkunst* (1641) and *Ausführliche Arbeit von der Teutschen Haubt-Sprache* (1663) (ibid.). The most prominent figure in the further establishment of the standard language in the following 18th century was Johann Christoph Gottsched (1700-66) with his *Deutsche Sprachkunst* (1748), which among his other works "laid down rules for grammatical and stylistic usage in [literary] German" (Chambers & Wilkie, 1970, p.49). What has been characterized as "the first great dictionary of standard German" was also published during this period; Johann Christoph Adelung published the *Verusch eines vollständigen grammatischkritischen Wörterbuchs der hochdeutschen Mudart* from 1774 to 1886 and follwed it with a text on German spelling, the *Vollständige Anweisung zur deutschen Orthographie* (1788) (ibid.). Following Luther's example, these works were unofficially based on East Middle German, and major poets and writers of the late 18th century such as Gotthold Ephraim Lessing, Johann Wolfang von Goethe and Friedrich Schiller followed this model.

More recently, German achieved something closer to its contemporary standard form largely due to the efforts of a school headmaster named Konrad Duden. Partly as a result of government directives on the subject, Duden published a very influential twelve-volume work in 1880, which is now in its 26th edition and generally known as "the

Duden." The first volume of this is still respected as the official source of proper German spelling, while the others deal with grammar and other aspects (Chambers & Wilkie, 1970, p. 50; Eisenberg et al., 1998; Dudenredaktion, 2000) This had a significant influence in decreasing the synchronic variance in the language (Hauser et al., 2007, p. 3). German spelling today is characterized as "highly systematic" and corresponds very closely with the pronunciation of the language, unlike English (Eisenberg, 1994, p. 358). The year 1996 saw a set of much-contested spelling reforms imposed by a federal government committee. However, these were relatively minor and affected only a small minority of words. Further, despite being official standards, they were rejected by many, including some major newspapers. Their effect on synchronic or diachronic spelling variation is thus not of great consequence (Johnson, 2005).

2.2. Issues of NLP on historical texts

Various types of natural language processing (NLP), including POS tagging, are more difficult on historical texts for several reasons. Firstly, historical texts in any language differ from contemporary texts in the matter of spelling; throughout most of history any given language lacked an official standard orthography. "German orthography was formally regulated as late as 1901" (Piotrowski 2012, p. 12; Ernst-Gerlach & Fuhr, 2006) although there were various attempts to standardize this and other aspects of the language in previous centuries, as described in the introduction section. Before this effort there existed "different "schools" of spelling or written dialects and a much wider range of acceptable spellings, often reflecting regional pronunciations" (Piotrowski, 2012, p. 12). English spelling, by contrast, was "more or less fixed around 1800," implying that German has a greater mass of more recent pre-standardization historical texts (Ernst-Gerlach & Fuhr, 2006, p. 49). This contrast is even greater for Spanish or French, as the nations from which these languages originated have had "institutions defining spelling standards" for centuries (ibid.). Dealing with such texts, then, may be more pressing for German than for several other major languages, providing one more reason to focus on German.

Both diachronic and synchronic variation in spelling are apparent. The term

diachronic variation in this context refers to change over time; a particular word in a 16th or 17th century document will in many cases be spelled differently than the same word in a modern one. These differences tend to become more pronounced as one moves further into the past. A graph (Figure 1) reproduced by Piotrowski (2012, p. 13) gives the example of what in modern orthography is the word "wird," meaning "he/she/it will, will be, or becomes." In 1680 the modern spelling was universal, but in 1620, about 40% of the time, this word was instead spelled "wirdt," and in 1500 the modern spelling was practically nonexistent, accounting for only 0.6% of occurrences. Along similar lines, as shown in Figure 2, one paper estimated that for one corpus of Early Modern German (1650-1800), 35-40% of all tokens in material published near the beginning of the period represented nonstandard spellings, while the proportion of such tokens was only 11.3% for the material from 1790 and 5.4% in 1798 (Scheible et al., 2011a). "Nonstandard" in this context refers to tokens which do not fit the standard German spellings of the present day. Another estimate by Baron et al. (2009) on historical English corpora, shown in Figure 3, suggests a dramatic decrease in the percentage of such "variants" from about 1500 to between 1700 and 1800, to about 10% of their peak rates (Piotrowski, 2012, p. 16).

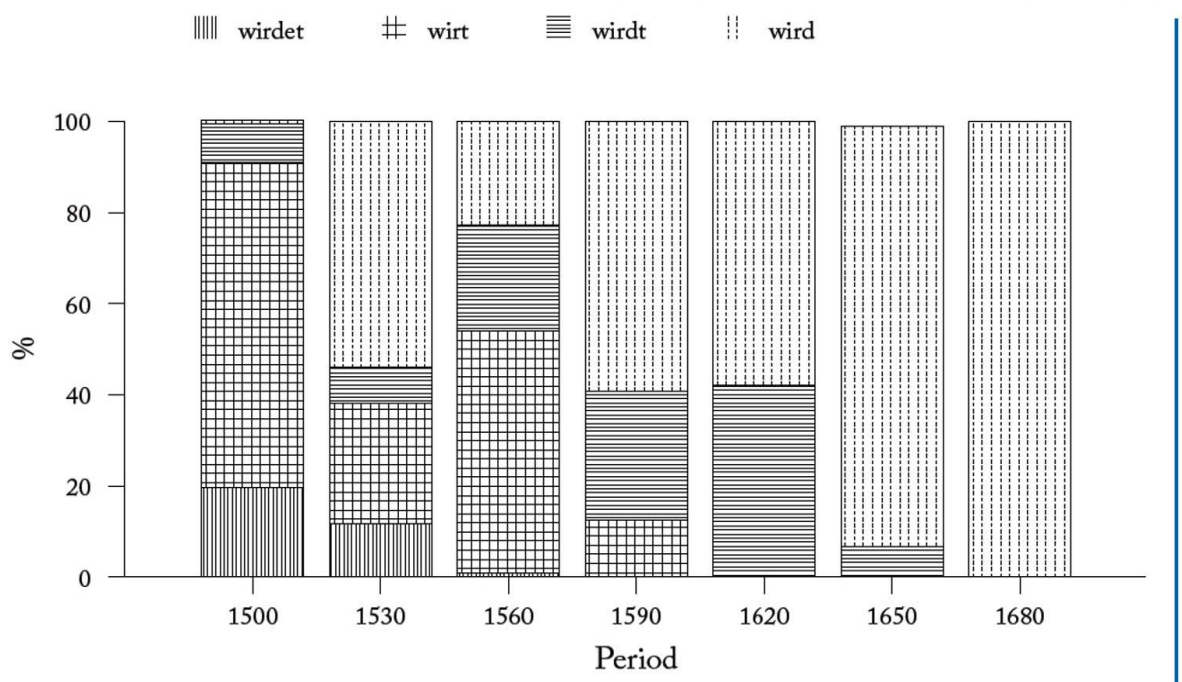


Figure 1: Percentage distribution of spelling variants of the German word *wird*, meaning "will be/becomes," plotted against year. Data from Ruge, 2005, p. 69. Piotrowski, 2012, p. 13.

Synchronic variation refers here to variation between spellings of the same word during the same time period, and as such is demonstrated by Figure 1 as well. As another example of both types of spelling variation, the GerManC corpus (discussed in more detail in the Corpora and Methodology sections) includes two files named SCIE_P3_OOD_1788_Chimie and SCIE_P3_OMD_1781_Chymie. In both of these the last word means "chemistry," but the modern spelling is "Chemie." In historical documents, even the same author in the same text may use different spellings for a word (Piotrowski, 2012, pp. 3, 17-18; Dipper, 2010, p. 117; Willemyns, 2013, p. 71). Piotrowski (2012) cites the 16th-century defamation case record already mentioned, in which a word for "girl" was spelled both "*meitli*" and "*meittli*" within the same sentence:

> Inn derselbigen bycht habe er, herr Jacob, zum meitli gredt, es steli ime das höw, die eyer und das kruth unnd derglychen ungschickte frag stuck mit ime, meittli, triben. Weliches aber sich, ob gottwil, nit erfinden werde unnd vermeine derhalben er, herr Jacob, sölle söliches darbringen oder in des meittlis fußstapffen gestelt werden.
> (Piotrowski, 2012, p. 20)

As another example of historical spelling variation, the vowel sound “i” has been represented by all of the following: j, y, ÿ, ie, iee, i^{e}, ij, ye, ih, jh, ieh, yh, with o and ä each having another six variants of their own, although it is not specified in the source how many of these were used synchronically (Hauser et al., 2007, p. 3).

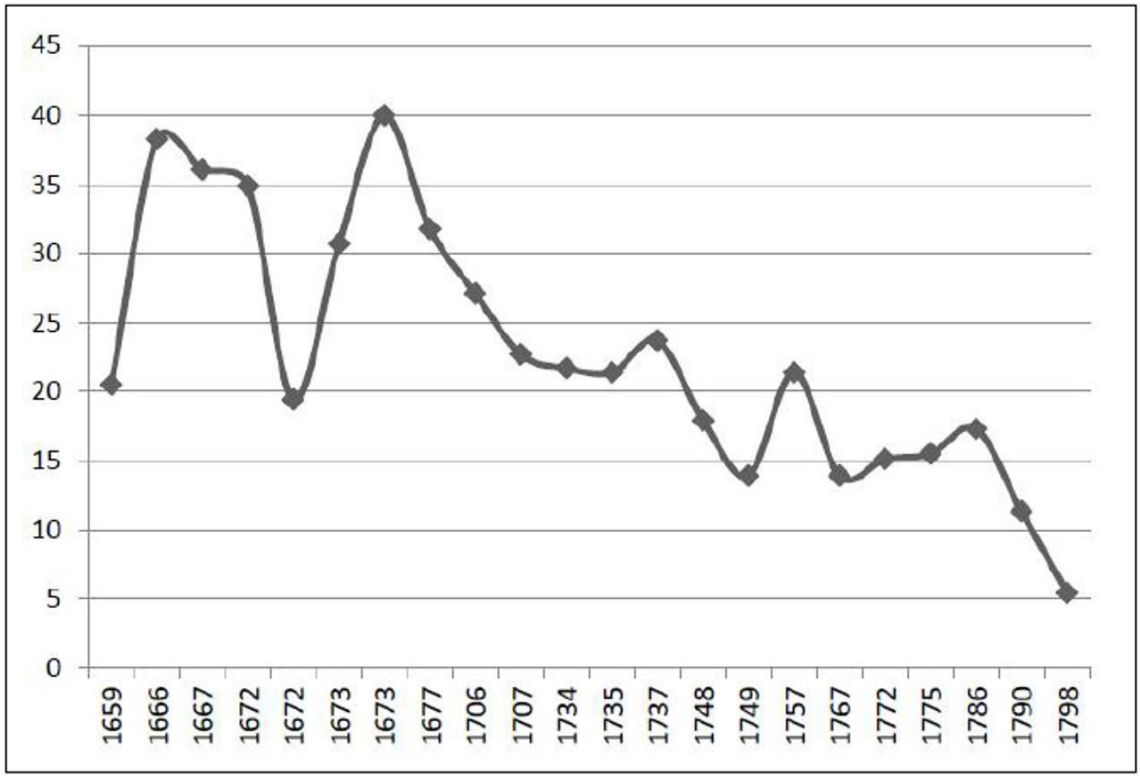


Figure 2: Percentage of normalized tokens, i.e. those which did not correspond to the present-day standard forms before normalization, plotted against year of document creation. TreeTagger evaluated on GerManC. Scheible et al., 2011a, p. 21.

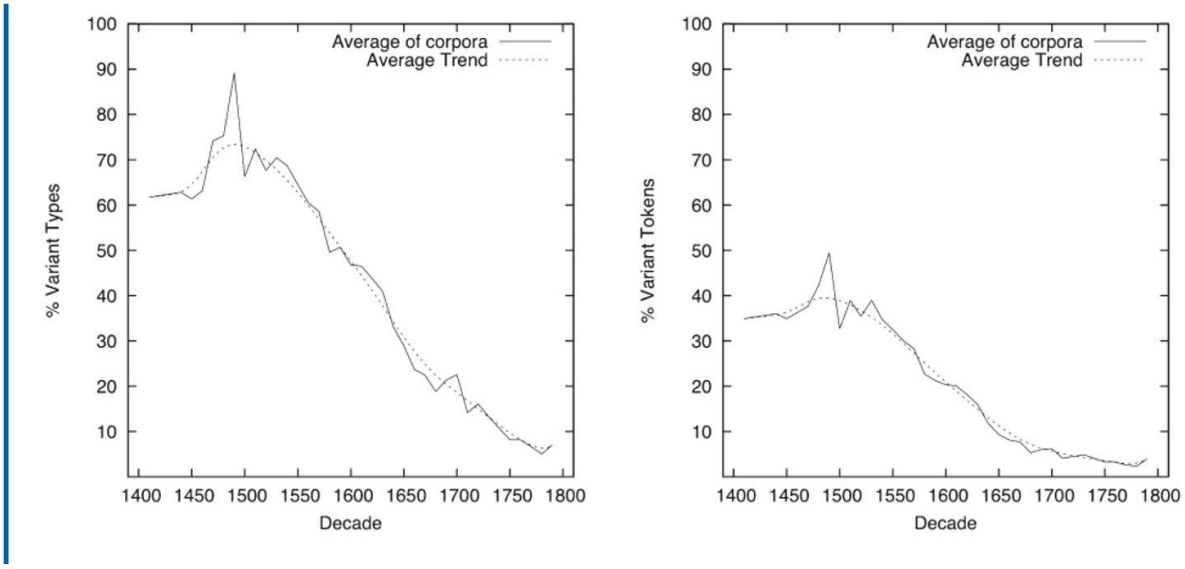


Figure 3: Extent of spelling variation in Early Modern English, based on six corpora; plotted figures are averages. "Variant types" and "variant tokens" refer to types and tokens which do not correspond to the present-day standard. Both are plotted against time, with figures for each decade. Diagrams originally from Baron et al. 2009, p. 52. Piotrowski, 2012, p. 16.

This lack of standardized orthography is problematic in more than one way. Constructing a dictionary to use in POS tagging (or any other NLP analysis) is made more complex by many single words each having a variety of spellings. To again cite the graph used by Piotrowski (2012, p. 13), in documents from 1530 there were four spellings of a single word extant; in order of increasing frequency, these were wirdt, wirdet, wirt, and wird. There can also be confusion based on overlap between words with different meanings. To use an English example, the modern term "bee" refers to an animal, but is also an archaic spelling of the verb "be." An example in German is the word "statt," which is a modern word for "instead of" but is also a 15th and 16th century spelling of "Stadt," meaning "town" (Gotscharek et al., 2009, p. 196). This type of confusion, when encountered by a spell-checker, is known as a real-word error or a "false friend," and also arises with pairs such as "affect" and "effect," as well as with "pray" and "prey" (Piotrowski, 2012, p. 16). The modern German term Urlaub, whose contemporary meaning is "vacation," means "permission" in Old High German and

Middle High German and “farewell” in Early New High German (Hauser et al., 2007, p. 3).

One word having multiple spellings also means that each of these spellings will be less frequent than a single universal spelling would be. This leads to “problems of data sparseness, for example, due to the size of the training corpus, only spelling A may occur in some context, even though spellings B and C are in fact equally likely” (Piotrowski, 2012, p. 3). This may mean that historical documents require larger corpora for use as training data, so that more of the possible spellings of any particular word will be encountered during the training.

Before performing POS annotation or tagging, a text must first be tokenized, which involves identifying every word as an individual unit. This does not mean simply recognizing each bit of whitespace, as words are often not only adjacent to whitespace but adjacent to punctuation marks which generally do not count as part of the same token. That is, in the phrase “German, the greatest language in the world,” the first token is “German” rather than “German,”. Tokenization can present other problems, for example with respect to what are known as clitics, like the English “don’t” or the German “hastu” (from “hast du”). These must be split into two separate tokens before they can be tagged with parts of speech. Clitics are not limited to historical documents, but they “can occur in various non-standard forms” in older texts (Scheible et al., 2011a, p. 20). Durrell et al. (2012) customized the software they employed for tokenization to deal with features of Early Modern German “such as certain ligatures (where two or more graphemes are joined as a single glyph, as in *Æ*) or combining letters such as a superscripted e in place of an umlaut” (p. 7).

The nature of historical texts may also lead to mistakes in OCR (optical character recognition), by which physical documents such as manuscripts are converted into text files on a computer. Not only is all OCR subject to some degree of error, but historical texts may often be either handwritten or printed in unusual fonts such as blackletter fonts, or even in a combination of fonts (Piotrowski 2012, p. 32-33). Both of these conditions can make OCR much less accurate, leading in turn to more mistakes in tokenization, as well as in tagging (Lin, 2003). In some such texts, there may be a varying amount of whitespace between words. “For instance, sometimes [the author attempts] to squeeze in

an extra word at the end of a line, and as a result it is not straightforward to determine if one or two words were intended" (Bennet et al., 2009, p.6). OCR on 16th century text in particular has been noted as often producing "a complete disaster," involving not only unlikely sequences of characters such as "bcftgclrcn" but also "merges of two or three words" (Gotscharek et al. 2009, p. 198). Also, particular letters may be difficult to discern. If particular characters within a word are recognized incorrectly, the word may be wrongly judged as not being in the lexicon, making it less likely that the tagger will assign it the correct tag. It may also be mistaken for another word which is in the lexicon. Any of these errors may even affect surrounding words, since POS taggers tend to take the context of each word into consideration (Lin, 2003).

It is also customary that before further annotation, the sentences must be marked with beginning and end points, which is known as sentence splitting. This is not as simple as simply recognizing periods, as a period is sometimes used for an abbreviation which does not end a sentence (Wilcock, 2009, p. 22). Sentence splitting can also be difficult with historical German, including Early Modern German (EMG), in which "punctuation is far from standardised and may vary not only over different genres but also over time and even within a single text" (Bennett et al., 2009, p. 5). Sometimes sentences appear to end at some points which in modern German would delimit sentence fragments; Demske (2007) provides an example of the following sentence from a 17th- century newspaper called the Mercurius:

> Es sind viel/ welche vermeynen/ daß man allhier nicht die beste Opinion gegen Franckreich habe/ aus Ursachen/ weil sie in der Levante die unsrige nicht zum besten meynen sollen Mons. d'Almeras soll damahls/ als die Veränderung in Portugall vorgegangen/ mit vielen Frantzösischen Schiffen auff der Revier von Lisabona gewesen seyn. Worüber auch seltzame Gedancken entstehen. (Demske, 2007, p. 95)

Most of these words are actually identical or very similar to contemporary German terms, but the period before the word "Worüber" is not the end of a sentence; "Worüber" marks the beginning of a relative clause, like the English "about which," which certainly would not start a sentence. Further, texts from this period sometimes include no instances of the periods, question marks, and exclamation points which are used as sentence boundaries in contemporary texts. They may instead display ambiguous

use of colons, semicolons and slashes (also known as virgules) to mark "both clause and sentence boundaries, and it is often difficult to decide which function was intended by the author" (Durrell et al., 2012, p. 7).

Another common aspect of NLP is known as lemmatization. This involves assigning each word a lemma, that is, a basic root form. The lemma for "runs," "running," and "ran," for example, is simply "run." Like any NLP process, this can be hindered in historical texts by the OCR errors detailed above. It may also be complicated by the spelling variation already discussed, as this implies a larger set of forms for each lemma or even ambiguity as to which lemma a particular word belongs to.

Due to these issues, although "[m]ost existing NLP tools are tuned to perform well on modern language data," they "perform considerably worse on historical, non-standardised data (Rayson et al., 2007)" (Scheible et al., 2011b, p. 124). Research (Hinrichs & Zastrow, 2012) has suggested a general trend of tagging accuracy decreasing progressively as the year of document origin gets farther from the present day, although the trend does not follow a straight line. Such changes are observable not only between periods but even within the EMG period (Scheible et al. 2011a). Inter-annotator agreement for POS tagging has also been reported as lower for the first 50 years of this period than for the last 50, at 89.3% compared to 93.1% respectively (Scheible et al., 2011b, p. 127). Scheible et al. (2011a) compared the accuracy of a POS tagger on original and normalized documents from various years between 1659 and 1789 (Figure 4). Normalization in this case refers to modifying the words to fit present-day orthography standards. They note that the usefulness of normalization for increasing the accuracy of a tagger gradually decreases as one moves to more recent documents within this period, as described in more detail in the corpora section. In other words, more recent documents need less modification to fit the modern standard as they are already more similar to it. This is consistent with material cited in the introduction section suggesting that the written German language gradually became more standardized over several centuries. The progressively lower standardization in older documents makes any type of automatic annotation difficult on historical material, including POS tagging.

2.3. POS Tagging

"POS tagging in general "is essentially considered a "solved task", with state-of-the-art taggers achieving per-word accuracies of 97%–98% (Schmid, 1995; Toutanova et al., 2003; Shen et al., 2007)" (Giesbrecht & Evert 2009, 27). A more recent paper from Neunerdt et al. (2013, p. 139) repeats these figures. This does not refer to all texts in all languages, but only to major standardized languages; most of the world's approximately 6,000 spoken languages are not official languages of any nation and thus have no official standard, written or otherwise (Ernst-Gerlach & Fuhr 2006). The given accuracies are clearly not perfect, but "results of language processing techniques rarely hit a 100% success rate" (Nugues 3). Another reason to interpret these results as good enough to demonstrate a solved task is that this is often higher than the rate of inter-annotator agreement in existing research. In other words human beings will not agree about the proper tags 100% of the time, and often agree at an equal or lower rate than state-of-the-art taggers. Scheible et al. (2011a), for instance, cited a 96.9% inter-annotator agreement rate for spelling variation and characterized this as evidence that the task was fairly easy, in comparison with 91.6% for POS tags (p. 21). If professionally trained human beings on a fairly easy task cannot agree at a higher rate, then a tagger achieving an accuracy of this rate should be regarded as quite successful. Results in the high 90s are also acceptable in that they have proven sufficient in pipelines to further applications. However, "the reported tagging accuracies of 97%–98%" represent ideal conditions which are not always present, namely "(i) the taggers are applied to edited, highly standardized text with a low rate of errors and unusual patterns; and (ii) training and test data are very similar (usually from the same volume of the same newspaper), so that overfitting of the training data is rewarded to a certain degree" (Giesbrecht & Evert 2009, 27). In other words, there may still be significant room for improvement in POS tagging of less standardized text and with out-of-domain data. Previous research has suggested that using out-of-domain data causes a significant decrease in accuracy compared to alternatives. "Out-of-domain" in this context has referred to data which differs in dialect region (Dipper, 2010) or in text genre (Neunerdt et al., 2013; Giesbrecht & Evert, 2009; Kübler & Baucom, 2011) or even in both text genre and date of origin, as with the use of modern newspaper text to train a tagger for use on historical materials of a

variety of genres created over a very broad range of years (Hinrichs & Zastrow, 2012). Differences between training and testing data have also included the use of a gold standard for training (Scheible et al., 2012), comparison of original and normalized forms of a text (Scheible et al., 2012; Dipper, 2010) as well as different interpretations of the characters in the text, known as word forms and described in more detail in the taggers section (Dipper, 2010). Neunerdt (2013) demonstrated the significantly lower accuracy achieved with four state-of-the-art taggers on several genres of web-based text, material which was less in line with prevailing standards for several reasons. Hinrichs & Zastrow (2012) warned that using out-of-domain material for training "leads to a significant drop in accuracy already for synchronic data from different domains (Bikel, 2004, Kübler & Baucom, 2011), and is even more problematic when such models are applied to heterogeneous diachronic materials (Dipper, 2010)" (Hinrichs & Zastrow, 2012, 6-7).

As explained above, another type of text which is less standardized and thus tends to yield lower accuracy results is historical text. Dipper (2010), for example, even with the configuration in which training and test data were most similar, had results ranging between 86.6% and 92.9% for overall accuracy on Middle High German (MHG) using in-domain training data. Far lower results were achieved using out-of-domain training data, with out-of-domain here referring to data from a different dialect region. Dipper's (2010) experiments are described in more detail in the taggers section. Scheible et al. (2012) obtained results of between 80.25% and 89.58% on Early Modern German (EMG) text, and their experiments are described further in the corpora section. I used a historical period (EMG) which seemed to be less fully studied, divided in a manner I had not seen done in existing literature, and used this to test the significance of both out-of-domain training data and differences between time periods in a new way. This is described in more detail in the methodology section.

2.4. Corpora

The School of Arts, Languages and Cultures at the University of Manchester completed an Early Modern German corpus in August of 2011 known as GerManC, having begun in September of 2008 (Bennet et al., 2012, p. 2). The project was inspired

by Anita Auer, who complated her doctorate at Manchester University in 2005. Her work "drew attention to the lack of corpus-based data for German during this period compared to English;" she recommended the creation of such a corpus and "completed some preparatory work on it" (Durell et al., 2012, p. 1). Each of the 50-year periods from 1650 to 1800 is represented by 15 texts for each genre in this corpus (Piotrowski, 2012, p. 112-113; Scheible et al., 2012, p. 3611). This resource covers eight genres overall, chosen in an attempt to equally convey both spoken and written language of the period: drama, newspapers, sermons, personal letters, narrative prose (both fiction and non-fiction), scholarly writing in the humanities, scientific and legal texts (Manchester; Scheible et al., 2011a, p. 20).

The genre divisions, use of 50-year-periods, and the use within each period of equal numbers of equally-sized text samples from each genre were based on those aspects of a similar English historical corpus known as ARCHER (Manchester; Piotrowski, 2012, p. 112). The project was intended to serve as "a basis for comparative studies of the development of the grammar and vocabulary of English and German," including their standardization, as well as for various other purposes, thus the deliberate similarities with ARCHER (Manchester; Piotrowski, 2012, p. 112). Unlike the ARCHER corpus, however, which only covered American and British English, this corpus covered multiple regions, namely North German, West Central German, East Central German, South-West German (including Switzerland) and South-East German (including Austria). "[A]n equal number [three] of [2,000-word] texts for each genre and sub-period" was included for each of these major regions of the German Empire (Manchester; Scheible et al., 2012, Bennet et al., 2009). This resource comprises about 800,000 words in total (Piotrowski, 2012, p. 112-113).

GerManC's files are available in four formats and versions, known as RAW, TEI, LING-GATE, and LING-COL. The RAW files are very raw, i.e. they do not even have apparent line breaks. Also, the lack of apparent tokenization most likely makes them useless to for my purposes. Another version of each text, known as the TEI form, is an XML file annotated with structural annotations according to the TEI standard. These include, but are not limited to, header information such as title, author, and publication

information, as well as paragraph breaks and indications of particular types of tokens such as dates, abbreviations, and foreign language material. The linguistically annotated corpora include seven types of annotation including tokens, sentence boundaries, normalized (modern) spellings, lemmas, POS tags, morphological tags such as dat|sg|fem|3 for dative feminine 3rd-person singular, the position of each token in its own sentence (these are simply numbers, beginning at 1), and two categories of parser output, namely syntactic category and dependency relation (Durrell et al., 2012, p. 405). Although the linguistic annotations were done automatically, the project participants developed the automatic tools in question "specifically for early Modern German in order to improve the annotation quality" for the first 5 categories listed above (ibid., p. 5). The LING-GATE format, also known as GATE XML, not only expresses all of this linguistic information but adds structural annotations according to the TEI standard (Durell et al., 2012). The LING-COL format, however, is more appropriate to what is required for my chosen tagger, described in more detail in the taggers section. This format provides tokens in text format, "where tokens (plus annotations) are printed in tab-separated rows" with empty lines to indicate sentence breaks (ibid., p. 7).

A manually annotated gold standard subset of this corpus, consisting of only 57,845 tokens, was developed as well and used for GerManC. This was intended to be representative of the main corpus in terms of genre and time period, but not in the case of region (Scheible et al., 2011b). This latter variable was ignored, i.e. only one region (namely North German) was used; the authors explained that the region variable "exhibited considerably less significant relevant variation" than the other variables (Durrell et al., 2012, p. 5).

These gold-standard tags were assigned to tokens using an involved process. The first step consisted of tokenization and determining the boundaries of sentences using the German Language pluging of a text engineering tool known as GATE (General Architecture for Text Engineering) (Durell et al., 2012, p. 5). The output of this automatic annotation was then submitted to one annotator, who inspected it and "further added a layer of normalised spelling variants," a concept which will be discussed more below. The tagger was then applied to these new annotations, yielding POS tags and lemmas. All of these annotations "were subsequently corrected by two annotators, and disagreements

were reconciled to produce the gold standard" (Scheible et al., 2011b, p. 127). There was a 91.6% rate of inter-annotator agreement for POS tags, while the figure was 95.5% for lemmas.

In the normalization process, all spelling variants "were normalised to a modern standard;" each token was "labelled with a normalised head variant" (p. 20). The inter-annotator agreement for this task was 96.9%. The formatting of the files in the Gold Standard corpus was similar to that in the larger corpus. The gold standard corpus, containing 57,845 tokens (Manchester), was also discussed in their paper from the same year, A Gold Standard Corpus of Early Modern German (Scheible et al., 2011b).

The larger corpus should comprise 360 texts, but the available section of it has only 335 and seven genres. The genre "letters" is not yet publicly available due to "outstanding access problems and potential copyright issues" (Manchester University). Most of the letters are currently in the form of unannotated text and there were difficulties in acquiring sufficient numbers of letters, particularly for the South-West (South Upper) area (Durrell, personal communciation, 5 June 2015). Three letters files, including versions with linguistic annotations, were included in the gold standard, however. Otherwise, the corpus is freely available for download by any user from multiple sources including not only the project's website but also the Oxford Text Archive and the Economic and Social Data Service archive (Durrell et al., 2012, p. 2).

Scheible et al. also used GerManC as a corpus for POS tagging in their 2011 paper *Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text* (2011a). Their tagset, STTS-EMG, was based on Schiller et al.'s (1995, 1999) STTS (Stuttgart-Tübingen Tag Set), which is the standard for German, recently used for example by Hollenstein and Aepli (2014) in a modified form for Swiss German (p. 89). To the original 54 categories (Brants 2000b) the authors have made the addition of several new categories to make the set more relevant to Early Modern German (EMG). One of these was the result of the merging of two indefinite determiner categories from the original STTS "as the criteria for distinguishing them are not applicable in EMG" (Scheible et al., 2011a, p. 21), while other tags were relevant to categories such as interrogative terms and adjectives used as nouns. All of the specialized tags only accounted for about 2.0% of all tokens in the corpus.

Scheible et al. (2011a) used TreeTagger, developed by Schmid (1994) and refined by Schmid (1995), a probabilistic tagger which uses decision trees and can be trained on any language. As the title of their paper indicates, it was not trained on data specific to the intended testing, but employed in its given "off-the-shelf" state, which was based on newspaper text. Their results included an accuracy on normalized input of 79.7%, a dramatic improvement over the 69.6% yielded from data without this normalization applied. They did however note that the improvement granted by using normalized input gradually decreased as they moved from older to more recent texts, as expressed in Figure 4. As another measure of the value of normalization, around half (47%) of the normalized tokens in the corpus are only tagged correctly in the event that they have been normalized; their original forms produce an incorrect POS tag (p. 22). Of the remaining 53% of nonstandard-spelling tokens, almost all were tagged correctly regardless of whether than had been normalized; normalization had a negative effect on accuracy in only 3% of cases.

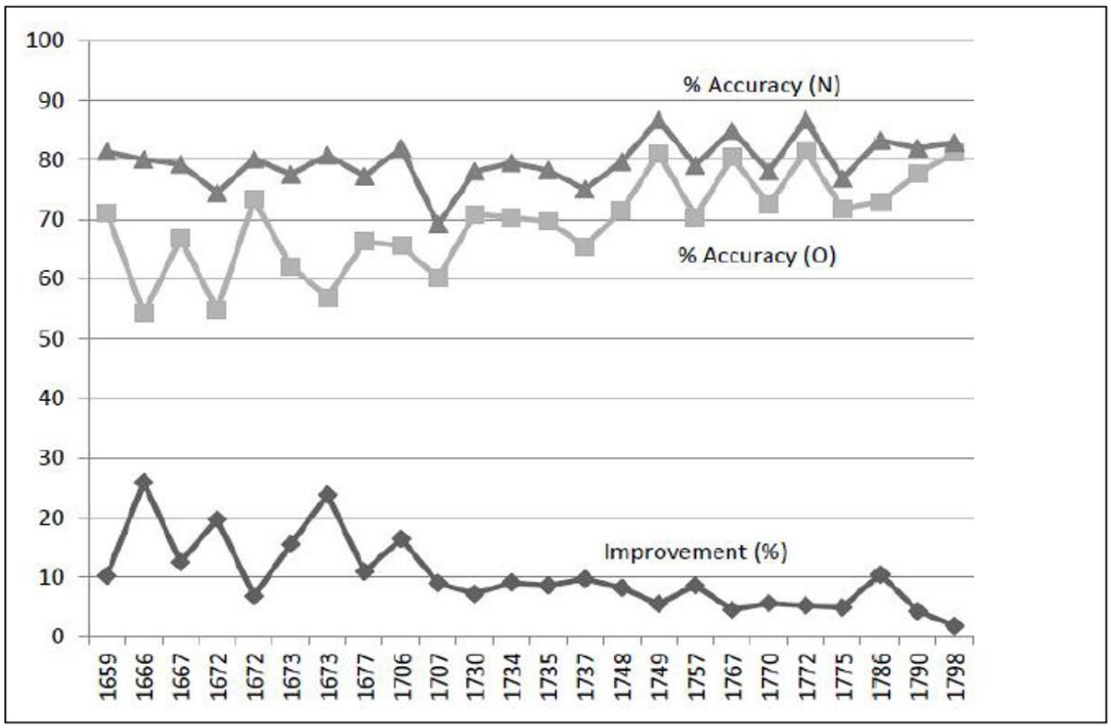


Figure 4: PoS tagging accuracy for original (light gray) and normalized (dark gray) corpora, along with difference between the two (darkest gray),

plotted against year of document creation. TreeTagger applied to GerManC. Scheible, 2011a, p. 22.

Scheible et al. also dealt with GermanC in their 2012 paper GATEtoGerManC: A GATE-based Annotation Pipeline for Historical German. Here they constructed an annotation pipeline using TreeTagger, which uses decision trees (Scheible et al., 2012). They trained their tagger on three copora, namely the already discussed gold standard corpus, GerManC-GS, as well as the original unaltered corpus and the automatic normalization. Scheible et al. compared the results, testing with each of the resulting models on both the gold standard corpus and an alternative normalization produced by "an automatic tool developed by Jurish (2010)" (ibid., p. 3615). As in their previous work, the TreeTagger was employed. Rather than 10-fold cross-validation as is often employed (Brants, 2000a; Dipper, 2010; Giesbrecht & Evert, 2009), they employed leave-one-out cross-validation. In this method, "the 24 gold standard corpus files were used to carry out 24 train-and-test cycles, in which 23 files were used as training material, and the remaining one file for testing" (ibid., p. 3615). However, this description does not entirely capture their methods; the 24 files involved were not always taken from the same subcorpus. As a table they include explains, they tested six combinations of training and testing, and apparently did 24 train-and-test cycles for each of these. These combinations consisted of training on each of the following three: the original unaltered corpus, the gold standard corpus, and the automatically normalized corpus, while testing for each of these on both the original corpus and the automatically normalized corpus. The authors explain that this ameliorates the problem of overfitting somewhat, as the training and test data are somewhat less similar. The authors' reported results included accuracy rates for their POS tagger. These results included original, manually normalized, and automatically normalized data. Their tagging scheme was again the STTS-EMG, a 54-tag set modified from the STTS "to account for differences between modern and Early Modern German (EMG), and to facilitate more accurate searches." The set again included categories in response to peculiarities of EMG, "such as various kinds of non-standard relative markers (Scheible et al., 2011b)" (ibid., p. 3615). For the POS tagger with all trained versions, they found an overall average accuracy of 89.44%, and normalization of

the test data further improved these results (ibid., p. 3611). This was a major improvement over the off-the-shelf version, which "only achieved 69.6% accuracy on the same data" (ibid., p. 3615).

Hinrichs & Zastrow (2012) used selected materials from Project Gutenberg-DE, a resource which unlike the main Project Gutenberg is under copyright (Zastrow, personal communication, April 17, 2015), to produce their own corpus which they called the Tübinger Baumbank des Deutschen/Diachrones Korpus (TüBa-D/DC), or Tübinger Treebank of German/ Diachronic Corpus. The term "treebank" here refers to a corpus to which syntactical annotations have been applied. The time period covered ranges from 1210 to 1930 (ibid., p. 3). The numerous genres covered include short stories, novellas, novels, plays, poetry, letters, fairy tales, autobiography and essays (ibid., p. 3). This corpus contains 525,529,365 tokens in 19,377 texts by 875 authors. Annotations, which were generated automatically, include tokenization, POS tags, lemmas, sentence boundaries, named entity information (persons, locations and organizations), and constituent parse trees. Three pre-existing tools were used to produce these, namely the OpenNLP tokenizer, the TreeTagger (Schmid 1994; Schmid 1995) for the POS tags and lemmas, and the Berkeley Parser (Petrov et al., 2006) for the constitutent parse trees, along with an in-house tool for sentence boundaries and named entity information. As the authors admit, "the materials are not always guaranteed to fulfill the same exacting standards of data quality" given the unknown origin of the data (Hinrichs & Zastrow, 2012, p. 2). They contrast their materials and metadata with the manually checked annotations of Deutsch Diachron Digital (DDD) (p. 3-4). They note metadata which are "often incomplete or misleading (for example by not documenting the actual textual sources that served as the input for the digital edition in [Project Gutenberg-DE])" (ibid., p. 4). After making some improvements themselves on the metadata, they still consider this aspect of the corpus imperfect. They employed the tagger TnT, described in more detail below in the Taggers section, and yielded accuracy rates ranging from 68.9% for Tristan (1210) by Gottfried von Strassburg to 98.7% for Die Leiden des jungen Werther (1774) by Johann Wolfgang von Goethe (ibid., p. 8-9).

Another decision by Hinrichs & Zastrow (2012) which is of interest is their focus on a single tag in the ground truth, NN (common noun), for which they counted the

mistaken tags produced by the tagger. They presumably chose NN due to the fact that it was the most common tag in every one of their texts. NE (proper noun) was by far the most common of these errors, followed by ADJA (attributive adjective, such as "large" in "the large house") and VVFIN (finite full verb) or VVINF (infinitive full verb). ADV (adverb) and even CARD (cardinal number) were also notable, though mainly in one particular text of the five involved in this comparison (p. 12-13). Hollestein & Aepli (2014) confirm one aspect of Hinrichs & Zastrow's (2012) results. They note that in their experiments on Swiss German corpora, the most common tagging errors were "the confusion of nouns (NN) and proper names (NE), which represent ca. 15% of all errors." They add that this is "a common problem for Standard German due to the capitalisation of nouns" (p. 92).

Hinrichs and Zastrow (2012) also employed another method of analyzing the prominence of particular tags in their results. Namely, using two particularly error-filled texts from their experiments, they arrive at a rough estimate of the difference in frequency of occurrence of particular tags between statistically tagged and manually corrected versions. Their graph indicates particularly high differences for the tags NE (proper noun), ADJA (attributive adjective), ADJD (predicate adjective, or adjective used adverbially), VVFIN (finite full verb), and FM (foreign language material). They note that there findings here agree with the "received wisdom that nominal, verbal, and adjectival categories are hard to tag for German" (p. 12).

Another relevant corpus in existence is the Bonn Early New High German Corpus, or Bonner Frühneuhochdeutschkorpus. This corpus was developed between 1972 and 1985 by Werner Besch, Winfried Lenders, Hugo Moser und Hugo Stopp. We can see here that there is uncertainty regarding the official boundaries of Early New High German; this corpus extends its diachronic range from 1350 to 1700 rather than 1650. These texts are divided by language region (Gotscharek et al. 2009). However, it is quite small, estimated at only 16,000 words (Piotrowski 2012, 112, Universität Duisburg-Essen). It consists of excerpts from 40 sources of varying length although the approximate typical length is estimated at 30 pages (Gotscharek et al. 2009). This resource is freely available online for noncommercial use at www.korpora.org/fnhd, both in unaltered form and with annotations indicating parts of speech and morphological

information. There is also apparently no gold-standard corpus available for it (M. Durrell, personal communication, May 26, 2015).

An ongoing project focused on essentially the same period of German (1350-1600) and extending the same corpus is the Referenzkorpus Frühneuhochdeutsch at Martin Luther University Halle-Wittenberg (M. Durrell, personal communication, May 26, 2015). This project intends to produce a representative multi-genre corpus of Early New High German as a resource for linguistic research (Martin Luther University; Deutsch Diachron Digital). The larger project, formerly called Deutsch Diachron Digital (DDD) but now referred to by several names including "Referenzkorpus", is the work of three universities, namely Friedrich-Schiller-Universität Jena, Goethe-Universität am Main, and Humboldt-Universität zu Berlin. Currently, there are five connected projects, as follows. A corpus of Old German (750-1050) was completed in 2013, contains about 650,000 words, and is accessible for querying through an online system called ANNIS (Piotrowski 2012, p. 112; LAUDATIO). The ANNIS system is a project of the Sonderforschungsbereich 632, affiliated with the Deutsch Forschungsgemeinschaft, Universität Potsdam, Humboldt-Universität zu Berlin and Freie Universität Berlin. Another corpus of Middle High German (1050-1350) was completed in 2014 and should be available through ANNIS "very soon" (S. Kwekkeboom, personal communication, June 2, 2015). The remaining three ongoing projects within the same framework include the already mentioned Early New High German (1350-1650) project, another related to Middle Low German which has been going on since 2012, and finally German inscriptions up to 1600, in progress since 2014. The ENHG project deals with manuscripts and prints, which will be "transcribed, lemmatised and annotated with grammatical information (PoS and morphology)" and will ultimately be "usable as a working instrument for medievalist and Early Modern research in ANNIS" (S. Kwekkeboom, personal communication, June 2, 2015). However, the project has not yet made any materials generally available for download. This project has been supported since 2011 by the Deutsche Forschungsgemeinschaft (German Research Foundation).

The Deutsches Textarchiv is a project of the Berlin-Brandenburg Academy of Sciences and Humanities. This project is only partially complete but will ultimately be comprised of works from circa 1600 to 1900. The distribution of documents varies, but

decades from the mid-18th century onward are more highly represented than previous years. These include literary works, scientific and scholarly texts, and "texts from everyday life," but the focus is intended to be on the social sciences and humanities forms (Berlin-Brandenburg Academy, Deutsches Textarchiv). Documents are remarkably varied in subject. They are categorized into three larger categories, namely literature, science, and "Gebrauchsliteratur," i.e. non-fiction which was designed for and is used for a particular practical purpose (such as textbooks). This is further divided into 27 subcategories including politics, law, economics, and theology. Literature, a much larger category than others outside of the top category (science), is divided into 19 subcategories while science (with a very broad definition of the term) is comprised of 48. It currently makes about 1,600 works available, mainly first editions in order to provide a more authentic picture of the language at various stages in history, and comprises approximately 11,496,659 words. This corpus is also notable in that its "search facilities tolerate a range of spelling variants," accounting for the synchronic variation issue discussed above (Deutsches Textarchiv). The text has been subject to annotations including part-of-speech labeling (Berlin-Brandenburg Academy). Many (about 16,000) texts are currently freely available to download under a Creative Commons license, while another 1,000 are currently undergoing a type of quality control.

The Historisches Korpus, a project of the Institut für Deutsche Sprache (Institute for the German Language), is another corpus of historical German, which ultimately intends to cover the period from 1700 until about 1918 (Institut für Deutsche Sprache). Seven subcorpora, some composed of further subcorpora, make up this corpus, including the works of Karl Marx and Friedrich Engels, Johann Wolfgang von Goethe, and the Brothers Grimm, among others. Other elements of the corpus, divided by genre, include legal documents, newspapers, magazines, scientific and legal texts, and philosophy. One subcorpus even purports to be a representative corpus of written German for the 1650-1800 period, based on GerManC. The overall resource is very large, currently containing about 70 million words, and is advertised as a basis for comparison between contemporary and historical German (Historisches Korpus). Gotscharek et al. (2009) estimated that this resource contained 408 ground truth texts and 3,044,255 tokens. However, for reasons of copyright, the materials are not all freely available for general

use. Only a limited subset of the texts can be used for research outside of the IDS. There is a retrieval tool known as COSMAS II, however, which can be used to search through this subset of texts online. (Piotrowski, 2012, p. 112)

The Historisches Korpus described above was used in Gotscharek et al. (2009) to create another historical German corpus called the Main Corpus. Other sources used by Gotscharek et al. (2009) included the Bonner Frühneuhochdeutschkorpus, the GerManC Corpus, and a manually selected "sample of 53 twice proofread German texts from 1504 to 1904 found on http://www.wikisource.org." Their complete corpus covered texts from the 15th century to 1932, with the years between 1500 and 1950 divided into nine 50-year periods and one additional category for all years before 1500. This resource contained 2,693,966 tokens and 288,709 unique words. However, POS tagging was not performed on this corpus; instead the resource was used for "analyzing the vocabulary of distinct periods, for dictionary construction," and for OCR and word recognition experiments (p. 194). While this resource was used for training, a disjoint resource was constructed for tests. The authors constructed separate test sets for each of the 16th, 18th and 19th centuries, using books from the Bavarian State Library (BSB). All pages were "manually groundtruthed" (p. 194) and the resulting corpus came to 25,745 tokens. They tested the effect of specialized historical dictionaries they had constructed in reducing word recognition errors from a baseline empty dictionary, finding that these improvements increased significantly when moving from earlier to later centuries. Their results include several references to 1750 as a turning point, including that before this point, "the percentage of spelling variants becomes very large, both in terms of types and tokens" (p. 196). To the best of my knowledge other researchers have not employed this corpus for POS tagging.

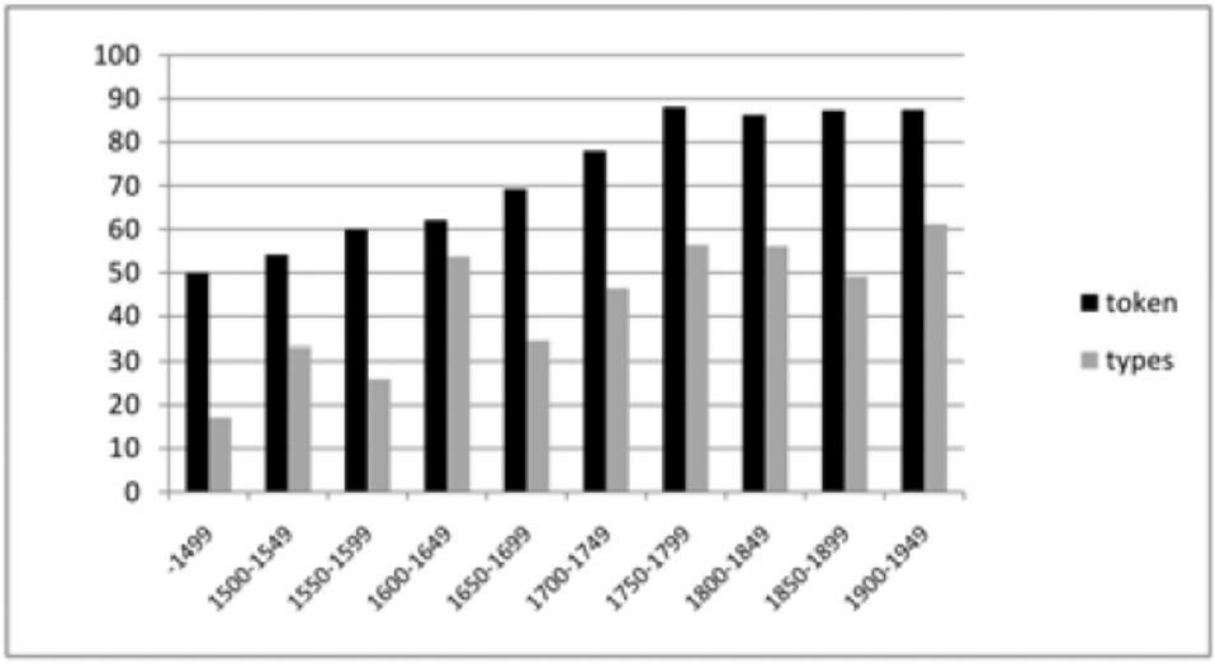


Figure 5: Percentage of modern words in texts from different time periods, both as a proportion of tokens and of types (unique words). This excludes compound words, Gotscharek et al. 2009, p. 196.

TextGrid is an open-source German-language project of a coalition of ten institutional and university partners, supported by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) (BMBF).[1] The project deals with the humanities and has produced tools and resources to which these tools can be applied. One of these resources is the Digitale Bibliothek (Digital Library), which bills itself as a collection of texts ranging from the advent of the printing press to the first decades of the twentieth century. This resource is particularly relevant for the study of German literature with 693 texts, as well as philosophy and cultural studies. The categories listed for the Digital Library include not only these three but also art, fairy tales, music, natural sciences, sociology and reference works. However, the fully prepared and available texts are at the moment limited to the category of literature (Text Grid). An estimated 1,267 texts are included overall, making it somewhat smaller than the Deutsches Textarchiv and TüBaD/DC in terms of number of texts but larger than GerManC or the even smaller Bonner Frühneuhochdeutschkorpus. Document counts are given for those Digital Library materials which are publicly available through zeno.org, a large online multimedia library in German for which the German publishing house Zenodot Verlagsgesellschaft mbH is responsible. XML and the data format TEI are supported.

Another corpus of historical German is known as the Mercurius Treebank of Early New High German, compiled from 2003 to 2005 at Saarland University (Linguist List). This is the first treebank ever created for historical German (Demske, 2007, pp. 92, 98), and represents two years, one in the later period of the Early New High German period and one arguably in the next period, Early Modern German. However, this is entirely limited to newspaper text. The newspapers in question are [*Nordischer*] *Mercurius*, published in 1667, and *Annus Christi*, published in 1597, meaning that it also does not offer a wide diachronic range (Linguist List). The limitation to newspaper text means that this resource is surpassed in diversity of material by most of the other noted corpora. This may put it at a disadvantage in term of overfitting, as explained by Scheible et al. (2012), and a representative corpus is preferable for studying linguistic trends at a particular point in time (Demske 2007, pp. 91, 92). The project currently consists of 8,387 sentences and 158,259 words (INESS), for about 170,000 tokens overall (LAUDATIO). Annotation was done semi-automatically, with syntactical annotation

according to the TIGER-SCHEMA. This project is discussed in Demske (2007).

LAUDATIO (Long-term Access and Usage of Deeply Annotated Information), a repository for multiple historical German corpora, is a project of the Berlin School of Library and Information Science (BSLIS) in coordination with several project partners. These include the department of Corpus Linguistics, that of Historical Linguistics, and the Computer and Media Service (CMS) at Humboldt-Universität zu Berlin and The National Institute for Research in Computer Science and Control (INRIA France). The entire project is funded by the Deutsche Forschungsgesellschaft (DFG). The German corpora hosted here include the Tatian Corpus of Deviating Examples; the Deutsche Diachrone Baumbank; the already discussed GerManC; the Kasseler Junktionskorpus; the Fürstinnenkorrespondenz; the already mentioned Referenzkorpus Altdeutsch; the Historishces Predigtenkorpus zum Nachfeld; the Mannheim Corpus of Historical Newspapers and Magazines; the Maerchenkorpus Version 1.0; the already mentioned Mercurius-Baumbank; and the RIDGES (Register in Diachronic German Science) Herbology Version 4.1.

RIDGES Herbology, produced by Humboldt University Berlin, covers the development of the German scientific language from the mid-16th to late 19th century, and contains 154,267 tokens in 29 documents. Annotation includes lexical, graphical, morphological, syntactical and other informaiton, including in a format for ANNIS and for MS Excel 2010 as well as PAULA (Potsdamer Austauschformat Linguisticscher Annotationen, or Potsdam Exchange Format for Linguistic Annotations).

The Tatian Corpus of Deviating Examples was designed, also at Humboldt University Berlin, as a function of Project B4 of the Collaborative Research Center on Information Structure 632. The corpus is drawn from the Old High German period and is confined to a single genre as it is comprised of only a single text. This text, stemming from the mid-9th century, is a compilation of Christian canonical gospels which was translated from Latin. Specifically, it is comprised of "parts of the Old High German translation attested in the MS St. Gallen Cod. 56, traditionally called the OHG Tatian, one of the largest prose texts from the classical OHG period." Only 11,295 tokens make up this work. Annotation was done by the same group which designed the corpus and includes the EXMARaLDA and relANNIS formats, including lexical, syntactical, and

other information.

The Deutsche Diachrone Baumbank (DDB, German Diachronic Treebank) is a product of a project sponsored by the Berlin Senate, namely the "Interdisziplinärer Forschungsverbund Linguistik - Bioinformatik zur Berechnung von Verwandtschaft und Abstammung." Overall size of the corpus comes to 8,580 tokens in 29 documents. Annotation includes EXMARaLDA format and reIANNIS, and includes a variety of syntactical elements along with POS tags allows identification of particular syntactic elements and distinction between grammatical [classes]. The objective was to find ways to apply methods from bioinformatics to automatically measure the relationships between language data. The data selected were as similar as possible with the exception of the fact that they were derived from different time periods, namely Old High German, Middle High German and Early New High German. Annotation includes lexical, morphological, syntactical and other information, including in the Negra format (version 4) and in PAULA-XML (Potsdamer Austauschformat Linguisticher Annotation or Potsdam Exchange Format for Linguistic Annotation).

The Kasseler Junktionskorpus developed from the DFG-funded project "Explizite und elliptische Junktion in der Syntax des Neuhochdeutschen. Pilotprojekt zu einer Sprachstufengrammatik des Neuhochdeutschen" from 2007 to 2009. Texts covered are taken from the 17th and 19th centuries. The intention here was to investigate juncture in New High German. This is a type of audible "boundary between two [adjacent] phonemes" such as a pause which helps to distinguish between phrases such as "a part" and "apart" (Rajimwale, 2006). The size of this resource comes to 119,420 tokens in seven documents. Annotation includes XML and numerous types of graphical, lexical and syntactical information.

The Fürstinnenkorrespenz 1.1 corpus stems from the DFG-funded project "Frühneuzeitliche Fürstinnenkorrespondenzen im mitteldeutschen Raum," and covers the period from 1546 to 1756. The contents are letters sent between women of the aristocracy and various men, mostly their relatives but also court officials and clergy. This resource comes to 262,465 tokens over 600 documents. The Historisches Predigtenkorpus zum Nachfeld lacks a single overall description but appears to be divided into 10 documents which are marked by language area and language period.

Dates of original documents seem to range from the 12th to the 17th century, and sizes from about 3,00 to 21,000, with a total of 93,536. Annotation includes EXMARaLDA and lexical, graphical, morphological, syntactical and other information.

The Mannheim Corpus of Historical Newspapers and Magazines represents the 18th and 19th century, namely newspapers and magazines, as the title implies. Over 4.1 million work tokens (4,739109 tokens overall) on 4,678 pages in 651 volumes and 1,287 documents comprise this corpus. It was assembled and digitized between 2009 and 2011, and in 2013 TEI P5 annotation was added. Annotation also includes TUSTEP. The Märchenkorpus Version 1.0 was assembled for the purpose of a seminar which took place at the University of Tübingen in 2013. The contents include 201 fairy tales and 10 further children's stories, all by the Brothers Grimm. The size of the corpus is 295,880 tokens. The annotation was done at least partly with the TreeTagger and includes POS tags and lemmas.

Another historical German corpus is the Parsed Corpus of Early New High German, created and maintained by Caitlin Light of the University of Pennsylvania using text from Wikisource. The corpus and the annotation manual were developed in collaboration with another much larger project, namely the Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al. 2011). The ENHG text all represents the year 1522, as it comes from a single source. Namely, it includes the books of Matthew, Mark, and John from the *Septembertestament*, Martin Luther's influential 1522 translation of the New Testament. The IcePaHC also contains a portion of a New Testament translation, albeit in Icelandic, which was done by Oddur Gottskálksson and printed in 1540. The creator of the ENHG corpus hoped to "initiate a set of parallel New Testament corpora for comparative syntactic and information structural ressearch" (Light 2015) which would also include the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) (Kroch et al. 2004). The PPCEME includes part of the 1534 Tyndale New Testament. The Parsed Corpus of ENHG currently consists of over 100,000 words, and the researcher is planning to expand this with the book of Acts. The corpus is available under a free open source license known as LGPL. Annotation, including part-of-speech tags and syntactic annotation, is based on the annotation guidelines for the Penn Historical Corpora of English (Santorini, 2010), only differing based on differences

between English and German. The annotation is described in greater detail on the website, including a list of tags and the broader categories into which they fit.

2.5. Taggers

The Stanford tagger (Toutanova et al., 2003), a maximum-entropy-based tool, was originally developed at Stanford University and tested on Wall Street Journal text the Penn Treebank. Its notable features included taking into account both preceding and following tags in calculating a particular tag. This was unusual at the time; "most of the best known and most successful approaches of recent years have been unidirectional," i.e. the current tag was predicted based on the current word and the previous tag, or in unusual projects the current word and the subsequent tag (ibid., p. 173). The result was 97.24% accuracy, which constituted "an error reduction of 4.4% on the best previous single automatically leanred tagging result" (ibid., p. 173).

The Stanford tagger was also used in a multi-tagger comparison of four "state-of-the-art POS taggers," chosen for "their tagging accuracy aplied to German standardized texts" (Neunerdt et al., 2013, p. 143) along with the TreeTagger (Schmid 1994, 1995), TnT (Brants 2000a), and SVMTool (Giménez & Màrquez, 2004). This dealt with social media text, a less standardized type of text and an area in which no POS tagging experiments had yet been done for German. Previous POS tagging experiments on social media text in other languages had dealt with models trained on newspaper text, and this use of out-of-domain training/test data, for example in Giesbrecht & Evert (2009), resulted "in a significant performance loss" (Neunerdt et al., 2013, p. 139). This had been the only option due to the lack of a a "social media text reference corpus, which is sufficiently large to train a tagger" (Neunerdt et al., 2013, p. 139).

Neunerdt et al. (2013) applied the four taggers listed above to social media text data. As training data, they drew from a new corpus of Web comments known as WebCom. This corpus consists of comments taken from a popular German news site known as Heise.de which deals with "different technological topics" (ibid., p. 141). They made selections from Webcom "randomly over different users according to their posting frequenices" in an attempt to obtain a corpus "where many kinds of social media characteristics are represented" (ibid.). All selections were annotated with "manually validated POS tags and lemmas" (ibid.). The result was dubbed WebTrain, making its purpose clear, and included 36,000 annotated tokens in contrast to the original source's 15 million. WebTrain was considered rather small to serve as a training set on its own, and so a joint-domain treebank was created by combining it with the existing TIGER corpus of contemporary newspaper text (ibid., Brants et al., 2004). The latter is much larger, containing about 900,000 tokens. This treebank includes manually annotated POS tags as well as lemmas, morphosyntactic features and parse trees, although for Neunerdt et al.'s (2013) purposes only the POS tags according to the STTS (Stuttgart/Tübinger TagSet) (Schiller, 1999) standard were employed.

The project in question also produced a corpus known as WebTypes, composed of approximately 5,000 tokens annotated in the same way as WebTrain, to use as a test corpus for all taggers. This corpus contained four different types of German social media texts, namely blog comments, chat messages, YouTube comments, and comments taken from the website of of [Rheinischer] Merkur, a German weekly newspaper. This corpus also included "a corpus extract from the Dortmunder chat corpus BalaCK 1-b (Beisswenger, 2007)" (Neunerdt et al., 2013, p. 142).

For their evaluation, the authors firstly employed 10-fold cross-validation. This was performed with all taggers, firstly on the TIGER corpus and secondly on a combination of the TIGER and WebTrain corpora. The folds employed here were "randomly selected sentences," presumably not contiguous (Neunerdt et al., 2013, p. 143). The models produced during cross validation were later applied for testing on the WebType corpus (Neunerdt et al., 2013, p. 144).

After being trained on either the TIGER corpus or a combination of this and the WebTrain corpus, each tagger was tested separately on both TIGER and the WebTrain

corpus alone. The results are shown in Tables 1 and 2. As is clear in these tables, the overall accuracies are always far higher when testing on the TIGER corpus than when testing on the smaller and less homogeneous WebTrain. Also, although both testing and training on what is partly the same corpus appears to give an accuracy boost of about 6 percentage points for WebTrain, it has no such effect for TIGER, a far larger corpus. The particular taggers mostly give very similar results, with narrow wins for SVM (on TIGER both as in-domain and out-of-domain data), TNT (on TIGER-WebTrain), and TreeTagger (on WebTrain as in-domain data). Also, the TreeTagger, trained on the WebTrain-TIGER combination, outperforms every other tagger when tested on WebTrain.

Text type	Tree-SPF	TreeTagger	TnT	Stanford	SVM
TIGER test	95.54 ± 0.06	97.18 ± 0.04	97.29 ± 0.05	97.42 ± 0.03	97.45 ± 0.03
WebTrain test	87.08 ± 0.87	88.51 ± 0.99	88.57 ± 1.14	87.74 ± 1.02	87.65 ± 1.13
Merkur comments	94.95	93.11 ± 0.34	90.78 ± 0.73	89.96 ± 0.42	91.64 ± 0.31
Chat messages	81.89	85.63 ± 0.39	84.34 ± 0.24	83.78 ± 0.26	82.80 ± 0.27
YouTube comments	78.88	77.53 ± 0.59	74.85 ± 0.39	74.44 ± 0.55	74.27 ± 0.42
Blog comments	87.98	88.14 ± 0.53	86.93 ± 0.68	86.53 ± 0.51	85.13 ± 0.67

Table 1: Tagger evaluation for different text types trained on TIGER, including a version of TreeTagger trained with the standard parameter file (SPF), including division by text type. All accuracy figures are averages over 10 iterations with standard deviations. Neunerdt et al., 2013, p. 146.

Text type	#Tokens	TreeTagger	TnT	Stanford	SVM
TIGER test	5,306	97.18 ± 0.03	97.31 ± 0.01	97.44 ± 0.01	97.47 ± 0.01
WebTrain test	3,628	93.72 ± 0.49	93.63 ± 0.37	93.18 ± 0.32	93.30 ± 0.56
Merkur comments	990	94.89 ± 0.38	93.49 ± 0.36	92.46 ± 0.38	93.72 ± 0.41
Chat messages	1,728	89.12 ± 0.18	87.96 ± 0.11	87.81 ± 0.16	86.57 ± 0.13
YouTube comments	1,463	84.03 ± 0.24	81.18 ± 0.19	81.23 ± 0.16	80.56 ± 0.19
Blog comments	815	91.35 ± 0.18	90.46 ± 0.12	90.29 ± 0.17	88.04 ± 0.13

Table 2: Tagger evaluation for different text types trained on WebTrain, including division by text type. All accuracy figures are averages over 10 iterations with standard deviations. ibid., p. 146.

	TreeTagger	TnT	Stanford	SVM
Total	93.72 ± 0.49	93.63 ± 0.37	93.18 ± 0.32	93.30 ± 0.56
Known	95.83 ± 0.43	95.81 ± 0.51	95.61 ± 0.40	95.58 ± 0.45
Unknown	67.98 ± 3.14	70.58 ± 2.08	68.14 ± 1.97	69.33 ± 2.54
Percentage unknowns	7.58 ± 0.75	8.65 ± 0.62	8.81 ± 0.62	8.65 ± 0.62

Table 3: Results for taggers trained on the joint-domain data. All accuracy figures are averages over 10 iterations with standard deviations. Neunerdt et al., 2013, p. 147.

As shown in Table 4, they also performed experiments on further training possibilities with the TreeTagger in particular, as it had provided the highest accuracy on the social media texts. They also trained a model on the TIGER corpus and tested it on text of each of the four types.

	Tree-SPF	*TIGER*	*TIGER* + Web Lex.	*WebTrain*
Total	87.08 ± 0.87	88.51 ± 0.99	92.63 ± 0.68	93.72 ± 0.49
Known	92.05 ± 0.53	94.47 ± 0.57	94.74 ± 0.53	95.83 ± 0.43
Unknown	44.77 ± 2.46	54.13 ± 3.15	66.47 ± 3.04	67.98 ± 3.14
Percentage unknown	10.50 ± 0.76	14.71 ± 0.96	7.58 ± 0.75	7.58 ± 0.75

Table 4: Results for the TreeTagger trained on several sets of training data, including the standard parameter files (SPF), TIGER, an expanded version of TIGER, and the WebTrain corpus.

The authors (Neunerdt et al., 2013) demonstrated that the addition of in-domain to the out-of-domain training data improved overall tagging accuracy significantly. For the TreeTagger, this improvement consisted of about 5 percentage points for the addition of in-domain data to the TIGER treebank, while in-domain data alone provided a further improvement of about 1.1%, as shown in Figure 4 (Neunerdt et al., 2013, p. 140). The individual taggers tended to give similar results to each other, within about one percentage point. Comparing models trained using this new corpus with those trained on the TIGER corpus alone (Brants et al., 2004), they found an improvement of more than five percentage points for the in-domain models

The TreeTagger, already discussed in connection with the GerManC corpus

(Scheible et al. 2011a, 2012) and in comparison with the Stanford Tagger (Neunerdt et al., 2013), is a probabilistic tagger which is based on a Markov model and employs decision trees. First developed by Schmid (1994) for English, it was later subject to improvements by Schmid (1995) to reduce its error rates on German by more than a third. It is cited in numerous papers including Volk & Schneider (1998) and more recently Giesbrecht and Evert (2009) and Dipper (2010). In Volk & Schneider (1998) it was compared to the Brill-Tagger, a rule-based tagger for German, and the results for the two were similar, but with a slight advantage for the TreeTagger in overall accuracy. The authors used a manually tagged corpus consisting of about 70,000 tokens from the Frankfurter Rundschau, a contemporary daily newspaper. They obtained this corpus from the University of Stuttgart and split it into 8 non-contiguous sets of sentences using a tool supplied by Eric Brill, using 7 of them for training and the remaining one for testing. The standard tag set for German, STTS (the Stuttgart-Tübingen Tag Set), which consist of 54 tags, was employed with one small modification (ibid., 126). In this experiment, the TreeTagger yielded overall accuracy of 95.27% while this figure for the Brill-Tagger was 94.75%. However, the TreeTagger was arguably preferable due to its much greater speed. The Brill-Tagger took 30 hours to train while the TreeTagger, training on the same corpus, took less than 2 minutes (ibid., 126-127).

The authors also listed the frequency of particular error types, i.e. which tags were incorrectly assigned to tokens and which tags should have been assigned in these cases. For both the Brill-Tagger and the TreeTagger the most common category by far was mistaking NE for NN; this is a type of confusion between common nouns and proper names which is common in German due a particular German grammar rule, namely the capitalization of all nouns. Confusion in the opposite direction, i.e. mistaking NN tokens for NE, came in second place for the Brill-Tagger and third place for the TreeTagger (Volk & Schneider 1998, 129). Giesbrecht and Evert (2009) reported similar results for error types for the TreeTagger applied to the TIGER Treebank (Brants et al., 2004) with the standard parameter files, with NE being mistaken for NN listed as the most common confusion, and NN mistaken for NE the third most common (p. 32). When applied to the DeWaC (Baroni et al., 2009), “a German Web corpus containing approx. 1.6 billion tokens of text collected in the year 2005” (ibid., p. 30), again with standard parameter

files, the most common confusion was identical. The third-place and fifth-place items were another confusion involving NN, and in fourth place again was NN mistaken for NE.

Dipper (2010) employed the TreeTagger on text from a historical period, namely Middle High German (1050-1350). The corpus used "is created and annotated in the context of the projects "Mittelhochdeutsche Grammatik" and "Referenzkorpus Mittelhochdeutsch" (ibid., p. 117) and contains texts from the 12th-14th centuries, "including religious as well as profane texts, prose and verse" (ibid., p. 118). The total size of the corpus is about 211,000 tokens. The texts contain a variety of both Middle German (MG) and Upper German (UG) dialects, with 27 MG texts, 20 UG and 4 mixed (ibid., 119). The corpus was semi-automatically annotated and these annotations included not only POS tags but also morphology, lemmas, and a normalized word form which represented a "virtual historical standard form" for each word, rather than normalizing to any modern standard (ibid., 118). The author employed an alternative form of 10-fold cross-validation in which the material was split into blocks of ten sentences, "or "units" of a fixed number of words, if no punctuation marks were available" (ibid., 119). For each of these blocks, nine sentences were used as training material and the tenth was used for testing.

Two parameters were modified during training, namely word forms and dialects. The options for dialects consisted of either the MG subcorpus, the UG subcorpus, or the entire corpus. The word forms parameter consisted of three different options. These are based on different interpretations of the original transcription, the latter having been produced by human transcribers, as well as a normalized version which was generated semi-automatically using a tool developed by the Mittelhochdeutsche Grammatik project. Although there was no official single standard at the time to which to normalize, the normalization was an attempt to "level out dialectical differences" by using an artificial standard (ibid., 120).

For testing, a third variable was introduced, namely the tagger (i.e. trained model). A tagger could be trained and tested (using the 10-fold cross-validation mentioned above) on a version of the corpus which was the same in terms of both training variables; this setting was referred to as "specific." If the tagger was trained on

one subcorpus and tested on another, for example trained on UG material and tested using MG, this setting was called "incorrect." For the setting "general," the tagger was trained on the larger corpus and then tested on a particular subcorpus.

Means and standard deviations across all 10 runs in the 10-fold cross-validation were given. Dipper (2010) found that "taggers achieve higher scores with UG data than with MG data, in all scenarios" (p. 120). The advantage of UG over MG was sometimes quite dramatic. For example, for the "incorrect" tagger used on the normalized form in MG, accuracy was 81.59 ± 0.44, while the corresponding figure for UG was 89.43 ± 0.49. The author expected results along these lines, as "MG data is more diverse and has a higher type/token ratio" (p. 120). Accuracy was always considerably lower for the "incorrect" tagger than for the specific or the general model, and was never above 93%. Also as the author expected, tagging with normalized forms was the most accurate overall, and displayed the least drop in accuracy when comparing "incorrect" tagger result to the specific or the general. This latter result was to be expected as the normalization had been intended to reduce the differences between dialects and would thus make a normalized corpus more internally consistent. Dipper (2010) provided no analysis of particular tag confusions, but only referred to this as an item for future research. Date of origin was also referred to in the same context.

The TreeTagger was also compared to Brants' TnT by Giesbrecht & Evert (2009), both narrowly losing in competition with it and narrowly defeating it depending on the context, discussed in more detail below. Neunerdt et al. (2013), already discussed in connection with the Stanford tagger, also employed it and found that with a particular training model, it had the highest accuracy for all of four types of German social media texts.

TnT (Trigrams 'n' Tags), first developed by Brants (2000a), is a widely used statistical tagger which uses Markov models (Giesbrecht & Evert, 2009, p. 29). This tool is unidirectional and utilizes trigrams. In other words the probability of a particular tag for a particular token is based not only on an examination of that token but on the previous two tokens as well. It includes a predetermined model trained on the NEGRA corpus (Skut et al., 1997; Skut et al., 1998; Brants, 2000b), a corpus of contemporary German newspaper text produced by Saarland University in Saarbrücken, Germany as

part of a project begun in 1996 (Brants et al., 2003). It also includes pre-trained models based on the Penn Treebank (Marcus et al., 1993), which consists of Wall Street Journal text, and another English corpus known as the Susanne Corpus which has its own tagset (Langendoen, 1997; Sampson, 1993; Brants, 2000c). Both of these are derived from the well-known Brown Corpus of American English, "which has been the cornerstone of language-related research across disciplines in the United States" (Ide & McLeod, 2001, p. 4). To avoid problems caused by sparse data, TnT employs a smoothing method, namely linear interpolation of unigrams, bigrams and trigrams (Brants, 2000a, p. 225). This avoids problems which would be caused by including a probability of zero in any of the relevant calculations.

The information TnT deals with as input and output is surprisingly simple. Untagged files used as input are simply text files with a single column, each column containing a single token, with blank rows generally indicating sentence breaks. Tagged files, which the tagger outputs, have one additional column consisting of the tags themselves. These are given the extensions .t and .tt respectively. Both of these formats are significantly simpler than the LING-COL versions of the GerManC files, so some preprocessing was required before the main functions of the tagger could be applied to them, described in more detail in the methodology section.

Brants (2000a) tested his TnT tagger on two corpora, namely the German-language NEGRA corpus described above, consisting of about 20,000 senteces (355,000 tokens) from a newspaper known as the Frankfurter Rundschau and secondly the English-language Penn Treebank (Marcus et al., 1993), particularly its Wall Street Journal text which consists of about 50,000 sentences (1.2 million tokens) (Brants, 2000a, p. 227). The author produced a learning curve by testing on test sets which gradually increased in size to demonstrate the effect of this change on average accuracy. The average here was taken at each particular size through 10-fold cross-validation; each such figure is the average of 10 iterations. Training and test sets were disjoint, to ensure that the accuracy given was applicable to data which was unseen from the perspective of the model. Brants preferred contiguous test sets, as employed by Giesbrecht & Evert (2009), as opposed to the "round-robin procedure that puts every 10^{th} sentence into the test set," (ibid., p. 227), as employed by Dipper (2010, p. 119). His justification was that using the latter would

mean that “parts of an article are already seen, which significantly reduces the percentage of unknown words” and would unrealistically inflate accuracy results (Brants, 2000, p. 227).

Statistical taggers in general have shown high performance; comparisons “of approaches that can be trained on corpora (van Halteren et al., 1998; Volk and Schneider, 1998) have shown that in most cases statistical aproaches (Cutting et al., 1992; Schmid, 1995; Ratnaparkhi, 1996) yield better results than finite-state, rule-based, or memory-based taggers (Brill, 1993; Daelemans et al., 1996). They are only surpassed by combinations of different systems, forming a “voting tagger”” (Brants, 2000a, p. 224).

However, in an independent comparison of 7 taggers by Zavrel and Daelemans (1999) the TnT tagger “not only yielded the highest accuracy, it also was the fastest both in training and tagging” (Brants, 2000a, p. 224). This indicated even higher performance than the Maximum Entropy framework, another statistical tagging approach which has a “very strong position” and was the only other approach which yielded “comparable results” to those of TnT (Brants, 2000a, pp. 224, 230). Further, Brants’ own experiments with the TnT yielded good results on the Penn Treebank (Marcus et al., 1993), which consists of Wall Street Journal text with a 36-tag tagset (Boehm 2005, p. 2), and the NEGRA corpus, which consists of contemporary German newspaper text. Brants’ results for the Penn Treebank can be seen in Figure 6 and for the NEGRA corpus in Figure 7. For the Penn Treebank, Ratnaparkhi’s (1996) accuracy with a Maximum Entropy approach was 96.6% while the authors’ accuracy was 96.7%, but Brants’ simpler model meant that his tagger was faster and therefore preferable.

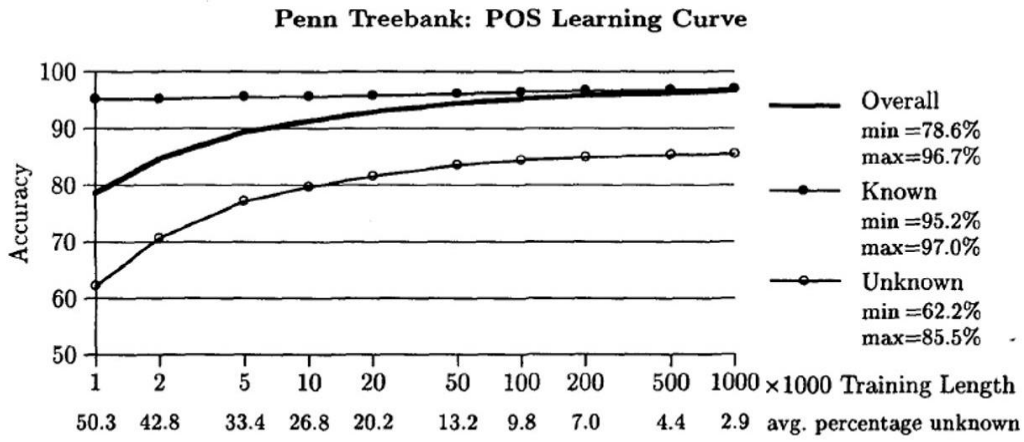


Figure 6: Learning curve for tagging the Penn Treebank. The training sets of variable sizes as well as test sets of 100,000 tokens were randomly chosen. Training and test sets were disjoint, the procedure was repeated 10 times and results were averaged.

Figure 6: PoS tagging accuracy vs. size of training set for TnT evaluated on Penn Treebank. All test sets consisted of 100,000 randomly chosen tokens, and were disjoint with training sets. Procedure was repeated 10 times; plotted figures are averages. Brants, 2000a, p. 229.

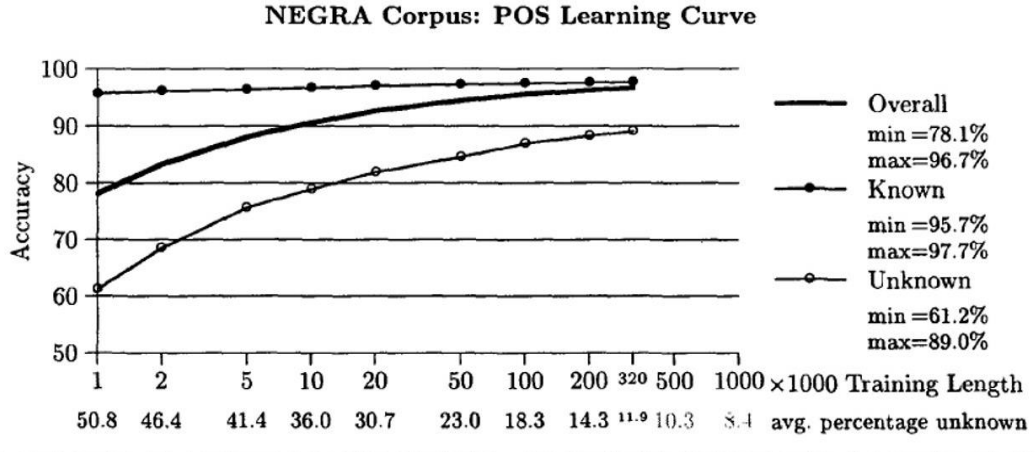


Figure 7: PoS tagging accuracy vs. size of training set for TnT evaluated on NEGRA corpus. All test sets consisted of 30,000 randomly chosen tokens, and were disjoint with training sets. Procedure was repeated 10 times; plotted figures are averages. Brants, 2000a, p. 228.

In a 2005 paper, Boehm compared TnT, which uses trigrams, with a simpler approach, namely the backoff tagger from NLTK (Natural Language Toolkit, a free online resource at www.nltk.org) (Loper & Bird, 2002)(Bird, 2006). The latter uses only unigrams, which makes it less sophisticated than state-of-the art taggers. The corpora in question were the Penn Treebank (Marcus et al., 1993), consisting of approximately one million tokens of Wall Street Journal text, and the Boston University Radio News corpus, consisting of "over seven hours of speech recorded from seven radio announcers taken from actual broadcasts" (Boehm, 2005, p. 2). In all configurations, the taggers were trained on the Wall Street Journal text, but testing took place on both corpora. The test set for the Boston Radio corpus contained 5,872 tokens while that the Wall Street Journal text was almost ten times this size, at 56,760. The effect of using out-of-domain training data was investigated here, and it was discovered that this option decreased the NLTK tagger's accuracy by 11.3% but that of TnT by only 9.5%. TnT outperformed the backoff tagger by about 10% in overall accuracy with in-domain data, and outdid the same tagger 84.04% to 72.1% in the out-of-domain configuration.

Hinrichs & Zastrow (2012) applied this tagger to their own large diachronic corpus, TüBa-D/DC, and yielded accuracy rates (percentage of tokens correctly tagged) ranging from 68.9% for *Tristan* (1210) by Gottfried von Strassburg to 98.7% for *Die Leiden des jungen Werther* (1774) by Johann Wolfgang von Goethe (pp. 8-9). Their accuracy rates did not vary entirely linearly based on time; accuracy was much higher for the just cited 1774 work (98.7%) than for a 19th century work (93.87%), and higher for a 1530 work (88.6%) than for one from 1703 (80.1%) (pp. 8-9). They speculate that wide variation in average sentence length between works may have been relevant; in both of the just-cited comparisons, the work with the higher accuracy rate had a significantly lower average sentence length, although their own testing of this theory is not conclusive. In any case, their results demonstrate the potential for a very high accuracy on EMG material; 98.7% in particular is above the 95.42%-98.25% figures cited as "state-of-the art" by Giesbrecht and Evert (2009). This is particularly notable in that the latter's results were on "easy genres....types of expository prose—all quite similar to typical newspaper text" (ibid., p. 32).

Giesbrecht and Evert (2009) also make several references to the quality of the TnT. They refer to it as one of the two "best-performing German taggers" (p. 29) along with TreeTagger, and conduct their own experiments whose results support a high opinion of it. These authors first employed the TIGER treebank (Brants et al., 2004), which they cited as "currently the largest manually annotated German corpus," consisting of "about 900,000 tokens (50,000 sentences)" of text taken from a contemporary newspaper known as the Frankfurter Rundschau (Giesbrecht & Evert 2009, 29). This treebank has been annotated with a variety of annotations by two independent annotators, followed by a consistency check. However, out of this annotation, the only type used for Giesbrecht and Evert's purposes was the POS tags according to the STTS tagset (Schiller et al., 1999). The taggers they used included not only the TnT but also the TreeTagger (Schmid, 1995), the SVMTagger, which uses support vector machines (Giménez and Màrquez, 2004), the Stanford tagger, a bidirectional Maximum Entropy tagger which had the best published tagging accuracy for English (Toutanova et al., 2003), and the Apache UIMA Tagger, an "open-source HMM tagger written in Java, implemented by one of the authors" (Giesbrecht & Evert, 2009, p. 29). First, using the standard parameter files which were included with TnT and TreeTagger for training, they evaluated these two taggers on the entire TIGER treebank. The results can be seen in Table 4; the TreeTagger narrowly defeated the TnT overall, although the TnT was dramatically more accurate on unknown words, which it came up with a much higher proportion of.

	overall	KW	UW	% unk.
TreeTagger	95.82	96.27	79.88	2.7
TnT	95.71	96.97	86.94	12.6

Table 4: PoS tagging accuracy on the TIGER treebank, using the standard parameter files (SPF) distributed with TreeTagger and TnT. Giesbrecht & Evert, 2009, p. 30.

Next, using five different taggers, they employed 10-fold cross-validation with contiguous folds as already noted, and calculated means and standard deviations for total accuracy rates, as well as for known words and unknown words, across all 10 folds. In this stage of the experiment, described in Table 5, the Stanford tagger achieved the

highest accuracy, while TnT achieved the third-highest, narrowly losing to the SVMTagger and narrowly beating the TreeTagger. The Stanford tagger achieved the highest rate of any tagger on unknown words as well, and again the TnT came in third after the SVMTagger, beating the TreeTagger.

	TreeTagger	Stanford	UIMA	TnT	SVM
total accuracy (%)	96.89±0.34	**97.63±0.24**	96.04±0.38	96.92±0.31	97.12±0.20
known words (%)	97.62±0.21	–	97.50±0.18	97.59±0.20	97.71±0.17
unknown words (%)	87.89±0.99	**91.66±0.83**	79.59±1.30	89.16±0.85	90.16±0.84
% of unknown words	7.44±0.78	7.52±0.46	8.10±0.71	7.85±0.88	7.82±0.82

Table 5: PoS tagging accuracy for 10-fold cross-validation on the TIGER treebank. Mean and standard deviation across all 10 folds are reported. Giesbrecht & Evert, 2009, p. 31.

In another experiment as part of the same paper, though, the TnT actually narrowly defeated the Stanford tagger. For this experiment, all taggers were trained on the entire TIGER treebank (Brants et al., 2004), while The TreeTagger was also trained on the standard parameter file (SPF). The performance of all taggers was then evaluated on a gold standard the authors had produced. This gold standard had been compiled by first selecting "a random sample of Web pages from DeWaC (Baroni et al., 2009), a German Web corpus containing approx. 1.6 billion tokens of text collected in the year 2005" (Giesbrecht & Evert, 2009, p. 30). The results are shown in Table 6. In this context, TnT outperformed all of the other taggers with an overall accuracy of 92.69%, and for known words again came in second, losing by only 0.01% to the SVM. For unkown words, the Stanford tagger was by far the most succesful here. Using the same training data as the rest of this experiment, though, the TreeTagger only attained 90.78 for overall accuracy, putting it in last place.

	TT-SPF[a)]	TT[b)]	Stanford	UIMA	TnT	SVM
total accuracy (%)	**93.71**	90.78	92.61	91.68	**92.69**	92.36
known words (%)	95.42	93.59	—	95.59	95.90	95.91
unknown words (%)	54.30	69.12	**75.35**	66.49	71.99	69.45
% of unknown words	4.15	11.48	13.00	13.43	13.44	13.43

[a]TreeTagger with standard parameter file included in distribution
[b]TreeTagger with parameter file trained on the TIGER treebank

Table 6: PoS tagging accuracy on the DEWaC gold standard, with all taggers having been trained on the TIGER treebank as a whole except TT-SPF. Giesbrecht & Evert, 2009, p. 32.

The Stanford tagger's overall accuracy in the 10-fold cross-validation experiment is tied for the highest published accuracy for the TreeTagger, namely 97.63%, making it "the best known POS tagger for German text" (Giesbrecht & Evert, 2009, p. 30). The high accuracy of the Stanford tagger, though, came at the price of far longer computation time. The Stanford tagger took 45 minutes to perform the same tagging task that the TreeTagger performed in less 10 seconds. The TreeTagger was again able to perform its task in less than 10 seconds for the training on the entire corpus, which took the Stanford Tagger 5.5 hours.

The TnT tagger has also done well in more recent comparison with other taggers. Hollenstein and Aepli (2014) compared five open-source statistical taggers, namely TreeTagger (Schmid, 1995), a reimplementation of TnT known as hunpos tagger (Halácsy et al., 2007), RFTagger (Schmid and Laws, 2008), Wapiti CRF Tagger (Lavergne et al., 2010), TnT (Trigrams'n'Tags) tagger (Brants, 2000a) and BTagger, a bidirectional tagger employing guided learning (Shen et al. 2007; Gesmundo and Samardžić, 2012). This was on a corpus they constructed of five written genres of contemporary Swiss German, using a version of the STTS tagset which they had modified to accomodate peculiarities of Swiss German. They found that the TnT tagger only narrowly lost to the BTagger, at 90.14% versus 90.62% overall accuracy, while the initial results of the other taggers were apparently not good enough to be worth mentioning in the paper, in other words the TnT outdid them (Hollenstein & Aepli, 2014, p. 90). Although this may seem like low accuracy, no other studies have been done on

POS tagging on Swiss German, and its high variability may account for the low accuracy; it is a dialect continuum with no official spelling rules, rather than a single standardized language (ibid., p. 85). In this, it is similar to historical German, hence the relevance of this study.

Another comparison of taggers involved TnT was done by Kübler and Baucom (2011) for the purpose of studying domain adaptation. This work used three corpora including the Penn Treebank (Marcus et al., 1993) and two other corpora. These two corpora consisted of "dialogues collected in a collaborative task," an unusual type of text which diverges in some ways from biomedical texts, a more commonly used domain for domain adapation. This type of text includes, for example, more imperatives and questions than formal written text, and also includes many contractions and instances of hesitation or self-correction (Kübler and Baucom, 2011, p. 41). While the Penn Treebank was used as the source domain, the target domain used for testing was the HCRC Map TASK Corpus, composed of "18 hours of digital audio and 150 000 words of transcription, representing 128 two-person conversations" (Kübler and Baucom, 2011, p. 43). This latter corpus was contained annotations including POS tags, although the annotations were ignored for the purposes of this experiment. Another corpus, the CReST corpus, consists of "7 dialogues, comprising 11,317 words in 1,977 sentences" (ibid.). This corpus, used as the target domain, is annotated for "POS, syntactic dependency and constituency, disfluency, and dialogue structure. The POS tagset is a superset of the tagset for the Penn Treebank, with the additional tags representing features unique to natural dialogue" (ibid.). Examples of additions to the training data included adding a set of all full target domain sentences on which all three taggers agreed, the same for all sentences on which two agreed, and adding lexical information to the TnT model. These authors compared the TnT tagger with the Maximum Entropy Lexicon-Enriched Tagger (MElt) (Denis and Sagot, 2009), and SVMTool (Support Vector Machine Tool) (Giménez and Màrquez, 2004), explaining that each of the three uses a different approach and has different biases. TnT provided superior performance overall in comparison with the other taggers, varying between a baseline rate of 85.77% and 89.15% accuracy depending on various changes in experimental design.

The above comparisons form the basis of my preference for this tagger over other

POS taggers cited in the literature, such as the TreeTagger used by many of the above. Further, unlike the TreeTagger, TnT has not yet been tested on the GerManC corpus to my knowledge. Despite its higher performance in some categories as measured by Giesbrecht and Evert (2009), the Stanford tagger is not my preference due to its much greater calculation time mentioned in that work.

2.6. Significance of the Study

There is a limited amount of literature dealing with the application of POS taggers to historical German and to Early Modern German in particular. Giesbrecht and Evert (2009) noted that most POS taggers "have been developed for English" (28). There is "no specialized tagger available" to deal with Early Modern German, 1650-1800 (Scheible et al. 2011a, 19), and in a paper for the following year, still "no specialised tools available for processing this particular stage of the language..." (Scheible et al. 2012, 3611), although the authors did develop a tagset specifically tailored to the language. This is significant given the vast amount of EMG text in existence, which remains relevant for various types of research. I applied a tagger, which has not to my knowledge yet been tested on EMG, to such materials.

There is also relatively limited research concerning the use of POS taggers on corpora which are not entirely newspaper text; Giesbrecht and Evert noted in 2009 that most POS taggers had been trained on a single corpus, namely the Penn Treebank which is composed of Wall Street Journal text (28). Most natural language processing as of 2012 "is currently done for newspaper texts" (Piotrowski 2012, 2). Giesbrecht and Evert estimated in 2009 that "Virtually all taggers have been trained and evaluated on newspaper text.... and it is not clear whether they would achieve equally high accuracy on other genres..." (27). Hinrichs & Zastrow (2012), who as noted believe that annotated corpora should each contain a variety of genres, note this focus on newspapers: "In the past, corpus collection efforts which have focused on the criterion of sufficient size and quality have concentrated on synchronic newspaper data" (2). Two major corpora already mentioned in connection with Brants' POS tagging, the NEGRA corpus and the Penn Treebank, are entirely composed of newspaper articles.

There have of course been examples of POS taggers applied to other genres. Tsuruoka et al. (2005) developed a POS tagger specifically for biomedical text, and other examples cited earlier in this paper also include other genres, including working with a variety of genres in combination. However, I have improved somewhat on the methodology of the existing work on historical German multi-genre text. This included implementing a tagger/corpus combination which has not yet been seen and dividing the corpus in a new way as already discussed. Previous work has divided subcorpora by dialect region (Dipper, 2010, Hollenstein & Aepli, 2014), by text type (i.e. genre) (Neunerdt, 2013), by individual year (Scheible et al., 2011a), and even tested the effectiveness of a tagger trained on contemporary text when tested on historical data (ibid.), but not divided the corpora into halves of a historical period of German. My results should contribute to understanding the strengths and weaknesses of this tagger on these subcorpora and the nature of these two sub-periods, and provide a basis for further research. This might indicate the relative strength or weakness of this tagger on these sub-periods overall and provide a basis for comparison with others, particularly the TreeTagger, which have already been tested on this language stage.

3. Methodology

I chose to test the effect on the TnT tagger's POS tagging acuracy of a division of historical German which has not to the best of my knowledge appeared in the literature to date. Namely, using the GerManC corpus, I divided one previously defined historical period of German (Early Modern German or EMG, 1650-1800) into two approximately equal halves. More precisely, I split a corpus whose documents range in date of origin from to 1654 to 1799 at a midpoint of 1729/1730 for reasons which will be explained in more detail below. My purpose was to investigate the significance of the differences between these periods for purposes of tagging accuracy with configurations based on three variables. These variables were training set period, test set period, and training set size, as explained below after the discussion of the corpus and tagger.

3.1. Tools/Resources Used

Hinrichs and Zastrow (2012) provide a convenient checklist for the criteria which corpora should ideally satisfy in order to be useful for research, as follows.

> 1. They should be freely available – and thus sharable – for academic use.
> 2. They should include materials of different kinds and text genres.
> 3. They should be of sufficient size.
> 4. They should be of sufficient quality, i.e. contain few typos and ideally only textual material proper. [i.e. not tables or hyperlinks] (Hinrichs & Zastrow 2012, p. 4)

Ideally, I would have liked to make selections from the TüBa-D/DC (Hinrichs & Zastrow, 2012). This corpus fulfills some of the above requirements in that it is large, covers all of the desired time periods, and consists of material from a variety of genres. The last point helps to avoid overfitting. It has also undergone some additional annotation work beyond that of the source materials in the German Gutenberg Project, which ameliorates but does not eliminate some of the problems caused by the difficulty of

automatic annotation on historical text.

Unfortunately, the TüBa-D/DC does not seem to be available publicly, and I am not able to gain access to these materials from the original source (Zastrow, personal communication, April 17, 2015). For this reason, I used an already existing corpus called GerManC, already described above. GerManC is desirable in that it has neither the size limitations of the Bonner Frühneuhochdeutschkorpus nor the access limitations of the Historisches Korpus, nor does it appear to be unfinished like the Deutsches Textarchiv. It is also not limited to a single genre, distinguishing it from the Mercurius Treebank, and although it is unfortunately not as wide in diachronic range as the TüBa-D/DC, it does cover the years 1650-1800, which still represent significant diachronic change. Finally, I see no mention for GerManC of the flaws in the data/metadata which Hinrichs and Zastrow (2012) describe for their own corpus.

The tagset I selected is the STTS-EMG, a modified version of the standard tagset for German. The modifications made by Scheible et al. (2011a, 2011b, 2012), described in the corpora section of the literature review in the present paper, make this set more attuned to the pecularities of Early Modern German. These modifications are relatively small; the nine new categories apply to only 2.0% of all tokens found in the gold standard subcorpus of GerManC. The total tagset consists of 54 tags, and is described in Appendix A with the tags specific to EMG marked in red. I analyzed the results using a simpler set of tags expressing broader categories such as "noun," included in Appendix B.

The tagger I used in this research is known as TnT (for "Trigrams and Tags"), described in more detail in the taggers section. This tool was developed and described by Thorsten Brants in a 2000(a) paper, which I chose due to the accuracy and speed in comparisons with other taggers cited in the taggers section, along with its free availability. This tagger is "freely available to universities and related organizations for research purposes" (Brants, 2000a, p. 230). This tagger has not to my knowledge been previously used on this corpus. The TnT package includes two predefined models, one based on the Susanne corpus of English text and another on the NEGRA corpus of German newspaper text. The latter uses the Stuttgart-Tübingen Tagset (STTS) (Schiller, 1995). The tagger can also produce further models by training on tagged files. This produces both a lexicon and an n-gram file. The former consists of a list of words

encountered in the tagged file, along with their frequencies and the frequency of each tag each word is given. The latter consists only of a list of unigrams, bigrams and trigrams, each followed by the number of occurrences in the tagged file.

3.2. Preprocessing

Before being accessible to the tagger, the files needed to be subject to preprocessing. In dealing with corpus documents, TnT can only train on documents in a simple two-column text format. This consists of a single column representing the tokens and a second column representing the tag given to each token. The format required for test sets is even simpler, consisting only of a single column representing the tokens. In either case, sentence breaks are indicated by blank lines. In order to make the relevant files conform to these standards I created a set of scripts using version 2.7.6 of Python. The files chosen were those adhering to the LING-COL format described in the corpora section. These files contained more than the necessary one or two columns, so the scripts I wrote stripped these columns in two ways to create separate training and test versions for each file. The same scripts included code to ensure that every sentence's end was marked by a blank line. Further, although the vast majority of the given files adhered to a single coherent standard in terms of the placement of the columns, there were exceptions. Eight files in the first half and 11 files in the second half produced unusual results when subjected to the main script as a result of the placement of their token columns. These then needed to be dealt with through a special version of the main script which took this into account. Python script was also used to concatenate files in various ways in connection with the training and test sets described below.

3.3. Variables and Configurations

Firstly, there was the question of the accuracy of a model trained on one period but tested on the other, in other words using out-of-domain data. The variables here would be training set time period and test set time period, and either one could be set to

either the early or the late period. Although four configurations are listed in Table 7, this did not mean four training runs; rather, the same model trained using one period was then tested using each of the two periods. To use the terminology of Dipper (2010), I first tested the "specific" configuration, in which both the training data and the test data were drawn from the same corpus. The "specific" configuration was then contrasted with what Dipper (2010) called the "incorrect" configuration, that is, the out-of-domain configuration described above. For the incorrect setting, I expected somewhat lower performance based on the evidence regarding out-of-domain training data, discussed in the literature review section. Dipper (2010) also referred to a third configuration called "general," in which "the tagger trained on the entire corpus is applied to some subcorpus" (p. 139). However, her results indicated that there was little difference in accuracy between specific and general settings, while the difference between specific and incorrect was dramatic. For this reason I chose to ignore the "general" setting for my own work.

Configuration	Training set	Test set
Specific	Early EMG	Early EMG
Incorrect	Early EMG	Late EMG
Specific	Late EMG	Late EMG
Incorrect	Late EMG	Early EMG

Table 7: Training and testing configurations for the TnT tagger on the GerManC corpus.

With the "specific" configuration I also compared the earlier and later periods in a way which might indicate to what degree the later period was significantly more standardized than the earlier period. In accordance with the historical changes in the language discussed in the introduction, including the "Issues of NLP on historical texts" section, I expected performance in the "specific" case to generally be superior with material from the more recent period.

As noted above, working with GerManC, I divided the EMG period into two sub-periods, one up to 1729 and one from 1730 onward. The overall GerManC corpus begins at 1654 and extends to 1799, which would put 75 years in the first half and 69 in the second.

This also meant 177 files in the first half and 159 files in the second. The document counts for particular genres are displayed in Table 8, while the token counts are displayed in Table 9.

I chose the division based on an already existing graph by Scheible et al. (2011a) of the change in performance of another tagger, the TreeTagger, on documents in the same corpus from different years from 1659 to 1798 (see Figure 4). In their graph, approximately the same accuracy is given for original documents in 1659 as for 1775. However, there is dramatic variation in accuracy between documents in these years, which seems to level out somewhat at 1730; the figures for 1730, 1734 and 1735 are close to identical, as are those for 1748, with a slightly lower figure for 1737. Subsequent to this, there is again dramatic variation, but with the average apparently higher than that of the years before 1730.

First half (1654-1729)			Second half (1730-1799)		Overall	
Genre	# of Documents	Proportion of Corpus	# of Documents	Proportion of Corpus	# of Documents	Proportion of Corpus
Drama	19	11%	26	16%	45	13%
Humanities	21	12%	23	14%	44	13%
Legal	28	16%	17	11%	45	13%
Narrative	23	13%	23	14%	46	14%
News	39	22%	27	17%	66	20%
Science	22	12%	23	14%	45	13%
Sermons	25	14%	20	13%	45	13%
Total	177	100%	159	100%	336	100%

Table 8: GerManC document counts by genre, divided by corpus half and overall.

First half (1654-1729)				Second half (1730-1799)			Overall		
Genre	Count	Proportion	Average	Count	Proportion	Average	Count	Proportion	Average
Drama	45458	12%	2525	68095	18%	2619	113553	15%	2581
Humanities	51041	13%	2431	58610	16%	2442	109651	14%	2437
Legal	69072	18%	2467	39981	11%	2352	109053	14%	2423
Narrative	52346	13%	2379	54285	14%	2468	106631	14%	2423
News	60375	15%	1548	53224	14%	1971	113599	15%	1721
Science	53787	14%	2445	55032	15%	2393	108819	14%	2418
Sermons	60213	15%	2409	47631	13%	2382	107844	14%	2397
Total	392292	100%	2242	376858	100%	2370	769150	100%	2303

Table 9: Total token counts over each genre, including raw counts and their proportion of the whole, and average token counts per document within each genre. All of these are calculated for each corpus half separately and also for the overall corpus. Counts obtained with TnT tagger on GerManC.

First half (1654-1729)			Second half (1730-1799)	
Training set	Documents	Tokens	Documents	Tokens
Fold 1	21	51040	18	43345
Fold 2	42	91029	36	86534
Fold 3	63	144605	54	125777
Fold 4	84	188054	72	169639
Fold 5	105	233646	90	211460
Fold 6	126	280755	108	255400
Fold 7	147	328232	126	295688
Fold 8	167	373868	145	340850
Test set	10	23734	14	34111
Total	177	1714963	159	1562804

Table 10: Document counts and token counts per fold in each time period, also including test sets (gold standard) and total documents in each fold.

This seemed to be a reasonably even distribution of genres, with the one exception of newspaper samples being more numerous than other genres in the first half. The average length in tokens was also generally similar across genres, with the exception of the news genre which contained considerably shorter pieces, especially in the first half. Finally, I calculated a sum of 376,858 tokens for the second half of the corpus and 392,292 for the first half. I hoped none of these differences was enough to cause a difference in the validity of either half as a training set or test set. However, given the relatively small size of the corpora, this is a possibility.

The third variable tested here was the size of training sets. I tested the influence of increasing training set sizes on accuracy, hoping to find an ideal size and/or a point of diminishing returns. In accordance with Brants' (2000) methodology, I kept the size of the test set constant while the training set size increased. However, since the only viable ground truth was the gold standard which covered only a small minority of documents, testing could be done only on those documents which appeared in the gold standard. I first selected the gold standard for each half as the test set, which consisted of 10 documents for the first half and 14 for the second, and concatenated it into a single file for each. Next I created training sets for each time period. Using a pseudo-random number generator in Python, I made "random" selections at the document level, rather than at the level of words or sentences, for the sake of simplicity and contiguity, and concatenated them together. My chosen sizes started at one-eighth the size of the particular subcorpus (minus the test set), then ¼, increasing by the same interval of 1/8 up to the size of the entire subcorpus, minus the test set. These sets were all sampled with replacement, so some overlap is expected. The training set sizes can be seen in Table 10, which also includes the numbers of tokens.

Plotting accuracy against the increasing training set sizes provides an interesting graph known as a learning curve. Given the size of the overall corpus, dividing it into nine parts (including one test set) should be sufficient to show the shape of the learning curve and the possible point of diminishing returns.

In this respect I partly followed the recommendations of Brants (2000a, p. 228), which are also followed in Hollenstein and Aepli (2014, p. 91). Figures 5 and 6 portray Brants' (2000a) use of this method, while Hollenstein and Aepli's (2014) implementation

of this technique is shown in Figure 7. Brants (2000a) used 9 and 10 sizes in his learning curves, while Hollenstein & Aepli (2014) use eight.

As explained above, the only feasible source of test sets was the gold standard. Because of this, the typical method of cross-validation, in which each of n folds of a larger set serve at some point as a test set, was not an option. Therefore, I attempted to get a somewhat similar result by simply repeating the random assignment of files to training sets multiple times for each of the configurations listed in Table 7 and averaging the results. To save time and effort I used eight iterations rather than ten for each combination of corpus half, configuration, and training set. The numbers displayed in the results section are therefore all averages of eight runs.

3.4. Evaluation

Accuracy here simply refers to the percentage of all tagged tokens which were tagged correctly. This is judged according to the relevant ground truth, namely the gold standard subcorpus. Reporting accuracy not only for all tokens overall but also for known tokens and unknown tokens seems to be standard across languages, so I have followed that example (Loftsson, 2007; Loftsson, 2008; Neunerdt et al., 2013, p. 147; Giesbrecht & Evert, 2009). In existing literature, accuracy is consistently lower for unknown tokens than for known tokens. I have also reported the proportion of unknown tokens, as like the accuracy, this reflects on the quality of the trained models; the fewer unknown tokens, the better.

I had some initial difficulties with the comparison function of the tagger which would normally provide the accuracy rates, due to differences in format between the tagger's output and the gold standard files. I instead judged them by importing both tagger results and gold standard files into Excel and making comparisons between them there using various formulas there.

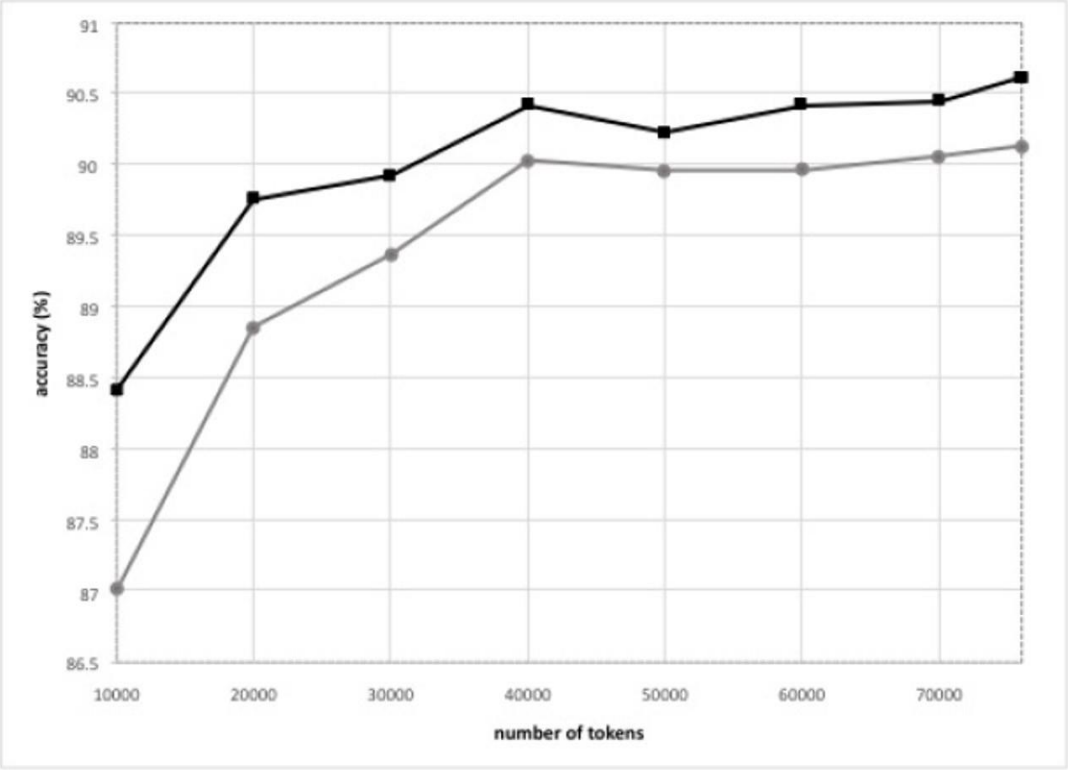
In addition to plain accuracy rates, I also included measures of Δaccuracy as the training set sizes increased, in other words the change between the accuracy for a particular training set and the previous, smaller training set. This is displayed in Tables 14 and 15 and was relevant for the purpose of giving an impression of a point of diminishing returns,

i.e. a point beyond which further increase in the training set size is not worthwhile because the Δaccuracy is so small. I have however not attempted to precisely calculate such a point. For similar purposes, the same tables also display a Δunknown, that is, the change in the percentage of unknown tokens in the same context.

For each testing configuration, I also used Excel formulas to evaluate the rates of particular errors, both confusions between specific tags (such as NE and NN) and confusions between broader categories of tags, such as noun and verb. This produced confusion matrices, large and small. The former were quite large and are not included in the present paper.[2] One of the latter, by contrast, is included as Table 13 in this work and represents the errors for the first half of the corpus with the "incorrect" configuration. This is the result of the averaging of eight iterations, although the differences in results between iterations for a particular corpus half and configuration were practically zero. I chose this half as it contained more errors proportionally than the second, and due to this is generally due more attention than the second. Along the same lines, the "incorrect" configuration also contained slightly more errors than the "specific" configuration, so I include the former in the body of this paper. For both first and second half individually, the differences in the content of the confusion matrix between "specific" and "incorrect" configurations was also negligible, with many figures being identical between the two. The remaining three confusion matrices, i.e. those for the "incorrect" configuration for the first corpus half and for both configurations of the second, are included in Appendix D.

In accordance with a principle of graphical integrity advocated by Edward Tufte that "the representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented" (Tufte, 1983, p.77), I felt that Hollestein & Aepli (Figure 7) also presented a more accurate visual representation of the effect of increasing sample sizes by using equal intervals between each of their successive sizes, and thus equal intervals for each tick mark on the ensuing graph. This contrasted with Brants' (2000a) use of intervals which varied from 1,000 tokens to hundreds of thousands. I chose to follow Hollenstein & Aepli's (2014) example in this.

Figure 7: PoS tagging accuracy vs. training corpus size for the TnT tagger (gray line) and the BTagger (black line), using NOAH's Corpus of Swiss German Dialects and sets of non-contiguous sentences. The corpus is divided into eight sizes for training, counting the largest size which consists of the full 73,616 tokens. Hollenstein & Aepli, 2014, p. 91.



4. Results and Discussion

For both configurations, I had higher overall accuracy for the second half than for the first half, as seen in Tables 11 and 12 and Figures 8 and 9. This was expected; as was discussed earlier, the second half represents a period during which language standardization was more advanced and text should thus be more consistent and easier to tag. Along the same lines, the first half had a higher percentage of unknown tokens, and accuracy was as expected dramaticaly lower for unknown than for known tokens. The general difference between corpus halves held true across both the "specific" and "incorrect" configurations, although the difference in unknown tokens between the first and second halves was somewhat larger for the "incorrect" than for the "specific."

Another notable result is that with regard to accuracy, the differences between the first and second half are much greater than the differences between the configurations. The greater standardization mentioned above is probably partly responsible for this. Another notable difference is the size of the test sets. While both halves have roughly equal training set sizes (with about a 10% difference in favor of the second half), the gold standard for the second half, consisting of 14 files, produces 34,111 lines after removing blanks, while for the first half, with 10 files, the comparable figure is only 23,370. A line here refers to a single token along with the tag assigned to it. Whether this is partly responsible for the difference in performance is unclear.

My results do not equal those of Brants (2000) with the same tagger, even for the second half of the corpus. Namely, for an already discussed modern German newspaper corpus, Brants (2000) achieved a maximum overall accuracy of 96.7% as opposed to my 93.4%, along with a maximum unknown token accuracy of 89.0%, providing a much greater contrast to my 75.94%. Brants' (2000) results for known tokens ranged between 95.7% and 97.7%. By contrast, my own ranged from 90.89% to 95.41%, across both halves, but similarly to Brants' results had a variation of less than 2% within a particular half. Across both corpus halves, my unknown accuracies exhibited greater variation across the range of training set sizes than my known accuracies, as did my overall accuracies. This contrast was again in line with Brants' results. The increase in overall accuracy as training sizes increase is thus more due to the increase in accuracy on

unknown tokens, as well as the decrease in the number of unknown tokens, considering their lower accuracy.

	First half (1654-1729)				Second half (1730-1799)			
Training set	Overall Accuracy	Known Acc.	% Unk. Tokens	Unk. Acc	Overall Accuracy	Known Acc.	% Unk. Tokens	Unk. Acc
Fold 1	82.67%	87.69%	22.85%	66.59%	89.81%	94.86%	21.01%	71.26%
Fold 2	84.38%	88.30%	18.97%	67.66%	91.44%	95.07%	16.68%	73.76%
Fold 3	85.00%	88.58%	17.21%	67.78%	92.13%	95.22%	14.49%	74.47%
Fold 4	85.56%	88.83%	16.01%	67.85%	92.55%	95.29%	13.10%	74.73%
Fold 5	85.69%	88.83%	15.20%	67.74%	92.88%	95.34%	12.08%	75.62%
Fold 6	86.24%	89.01%	14.35%	68.59%	93.14%	95.34%	11.30%	76.46%
Fold 7	86.39%	89.09%	13.85%	68.50%	93.28%	95.39%	10.75%	76.32%
Fold 8	86.54%	89.10%	13.32%	68.81%	93.43%	95.41%	10.20%	76.55%

Table 11: “Specific” configuration; overall accuracy, accuray for known tokens, and percentage unknown tokens and their accuracy.

	First half (1654-1729)				Second half (1730-1799)			
Training set	Overall Accuracy	Known Acc.	% Unk. Tokens	Unk. Acc	Overall Accuracy	Known Acc.	% Unk. Tokens	Unk. Acc
Fold 1	80.96%	87.39%	26.52%	64.03%	89.19%	94.78%	22.91%	70.96%
Fold 2	83.15%	88.08%	22.10%	66.14%	90.79%	94.95%	18.67%	72.92%
Fold 3	84.03%	88.37%	20.13%	67.16%	91.66%	95.13%	16.27%	74.01%
Fold 4	84.54%	88.51%	18.61%	67.40%	92.01%	95.24%	15.20%	74.50%
Fold 5	84.81%	88.65%	17.79%	67.50%	92.23%	95.30%	14.23%	73.89%
Fold 6	85.24%	88.79%	16.72%	68.05%	92.43%	95.25%	13.47%	74.62%
Fold 7	85.42%	88.86%	16.20%	68.11%	92.64%	95.29%	12.81%	74.97%
Fold 8	85.63%	88.92%	15.56%	68.27%	92.74%	95.30%	12.39%	75.05%

Table 12: “Incorrect” configuration; overall accuracy, accuracy for known tokens, and percentage of unknown tokens and their accuracy.

	Gold Standard											
Tnt Tag	article	pronoun	number	verb	noun	adj.	adverb	conj.	adpos.	part.	punct.	other
article	1543	48	1	0	11	3	4	8	0	0	0	3
pronoun	10	2871	0	5	12	10	28	16	0	0	0	158
number	1	1	257	0	16	34	7	1	0	0	0	8
verb	0	9	0	3214	41	50	30	3	11	0	0	7
noun	1	17	5	21	6716	79	10	1	4	0	0	190
adj.	0	46	6	38	79	1526	32	1	3	0	1	30
adverb	0	49	2	1	5	46	1568	48	11	21	0	127
conj.	0	3	0	2	4	0	8	1625	13	0	0	1653
adpos.	1	1	1	1	4	2	17	8	1555	18	0	10
part.	0	1	0	0	3	0	29	1	14	427	0	1
punct.	0	1	0	2	0	0	1	1	0	0	4767	3963
other	0	0	2	1	58	4	8	3	2	3	0	416

Table 13: Confusion matrix comparing the number of broad-category tags assigned by TnT against those assigned by the gold standard: “incorrect” configuration, first half of the corpus. Results are averaged over eight iterations.

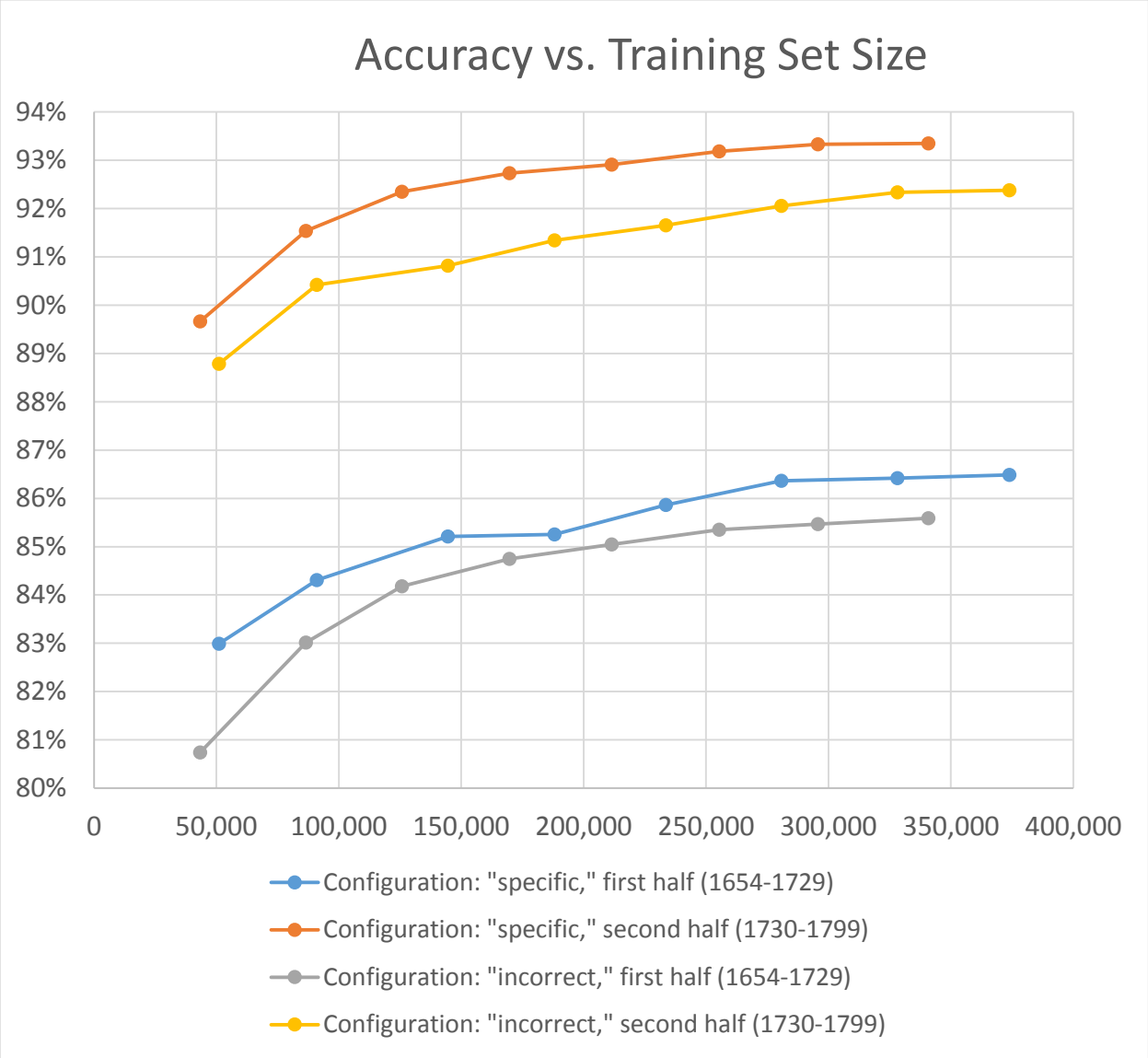


Figure 8: Accuracy plotted against training set size for each corpus half, including both "specific" and "incorrect" configurations.

	First half (1654-1729)				Second half (1730-1799)			
Training set	Δ Overall Accuracy	Δ Known Acc.	Δ % Unk. Tokens	Δ Unk. Acc.	Δ Overall Accuracy	Δ Known Acc.	Δ % Unk. Tokens	Δ Unk. Acc
Fold 2	1.71%	0.61%	-3.87%	1.08%	1.63%	0.22%	-4.33%	2.51%
Fold 3	0.63%	0.28%	-1.77%	0.12%	0.70%	0.14%	-2.20%	0.70%
Fold 4	0.56%	0.25%	-1.19%	0.07%	0.42%	0.07%	-1.39%	0.27%
Fold 5	0.12%	0.00%	-0.81%	-0.11%	0.33%	0.06%	-1.02%	0.89%
Fold 6	0.55%	0.18%	-0.85%	0.86%	0.26%	0.00%	-0.78%	0.83%
Fold 7	0.15%	0.09%	-0.50%	-0.10%	0.13%	0.05%	-0.55%	-0.13%
Fold 8	0.15%	0.00%	-0.54%	0.31%	0.15%	0.02%	-0.55%	0.23%

Table 14: increases of each accuracy value over the previous, along with changes in percentage of unknown tokesn, plotted against training set size for each corpus half, “specific” configuration.

	First half (1654-1729)				Second half (1730-1799)			
Training set	Δ Overall Acc.	Δ Known Acc.	Δ % Unk. Tokens	Δ Unk. Acc	Δ Overall Acc.	Δ Known Acc.	Δ % Unk. Tokens	Δ Unk. Acc
Fold 2	2.20%	0.68%	-4.41%	2.11%	1.60%	0.17%	-4.24%	1.95%
Fold 3	0.88%	0.29%	-1.98%	1.02%	0.88%	0.18%	-2.40%	1.09%
Fold 4	0.51%	0.14%	-1.52%	0.24%	0.35%	0.11%	-1.07%	0.49%
Fold 5	0.27%	0.14%	-0.82%	0.10%	0.22%	0.06%	-0.96%	-0.61%
Fold 6	0.42%	0.15%	-1.07%	0.55%	0.21%	-0.05%	-0.77%	0.73%
Fold 7	0.18%	0.07%	-0.52%	0.06%	0.21%	0.04%	-0.66%	0.35%
Fold 8	0.22%	0.06%	-0.64%	0.16%	0.11%	0.00%	-0.42%	0.08%

Table 15: increases of each accuracy value over the previous, along with changes in percentages of unknown tokens, plotted against training set size for each corpus half, “incorrect” configuration.

4.1. Error analysis

I began the analysis of errors by considering the first question of which tags were most often involved in errors of some type. In this and further analysis, unless otherwise specified, I am looking at the first half with the incorrect configuration, although the patterns of errors were quite similar across configurations and to a somewhat lesser degree similar across corpus halves as well. I used Excel formulas to produce lists and counts of the occurrence of particular tags in errors, both in the TnT output and in the gold standard. The most common tags were $, and $(respectively, both of which refer to punctuation. The $, tag was by far the most common item in the list of erroneous tags, with more than twice as many occurences there as any other, almost always replacing a $(in the gold standard. The $(in the ground truth was by far the most commonly confused character, with more than four times as many occurrences there as any other. Neither of these are listed in the STTS-EMG tagset table in the GerManC documentation (Appendix A in the present work), although they are both elements in the original STTS. They refer to certain non-sentence-ending punctuation, with $, referring to commas and slashes. Slashes in this historical context are often placed where a modern reader would expect a comma.

Mistaking $(in the gold standard for $, was also far more common than any other specific error in the first half, with over 1,000 occurrences, although it was almost nonexistent in the second. This was so common that one might at first suspect that these tags simply share the same meaning, i.e. that $(in the gold standard would almost always match $, in the tagger output and might as well be normalized to the same character. However, this is not so for several reasons. Firstly, $, was most often listed as a mistake made for some third tag rather than $(. Second, the gold standard does not lack the option of $,; it uses both "$(" (to refer to open and closed parentheses) and "$," (to refer to commas and slashes, which tend to act as commas).

After the $, the next most common erroneous tag was NN, which refers to a common noun. A very common tag throughout the results was NN. The documentation notes that this category does not include adjectives used as nouns, which the STTS-EMG tagset includes as a new tag, NA. The prominence of NN in German POS tagging errors

has been remarked on repeatedly in the literature (Hollenstein & Aepli, 2014, p. 92), some of which suggests that this is due to a pecularity of German orthography. Namely, German is the only modern language in which all nouns are required to be capitalized. This makes the distinction between common nouns and proper nouns more difficult. Indeed, after the $, and $(mistakes already discussed, mistaking a proper noun (NE) for a common noun was the most common specific error in my results. Interestingly, mistaking a common noun for a proper noun was not as common; the opposite had more than five times as many occurrences.

I followed Hinrichs & Zastrow's (2012) example in my own analysis regarding the focus on a particular tag in the gold standard and the relevant errors made by the tagger. In my own analysis, I counted the occurrence of particular tags in two columns, with one, the error column, referring to mistakenly assigned tags and another, which I will call the gold standard column, referring to those tags in the gold standard to which mistaken tags were assigned. I began with the NN tag, which Hinrichs & Zastrow (2012) also focused on. With the NN tag in the gold standard column, I found that NE (proper noun) and ADJA (attributive adjective) were by far the most common mistakes made. This is in line with the "received wisdom that nominal, verbal, and adjectival categories are hard to tag for German" that Hinrichs & Zastrow (2012) noted (p. 12). VVINF (full verb infinitive) was also notable in these results, although with less than half the frequency of either of the top two.

Based on another aspect of Hinrichs and Zastrow's (2012) work, I chose to use each of the five prominent tags noted in the methodology section in reference to p. 12 as the basis for further analysis similar to what I had already conducted for the NN tag. Namely, these were NE (proper noun), ADJA (attributive adjective), ADJD (predicate adjective, or adjective used adverbially), VVFIN (finite full verb), and FM (foreign language material). I also included ADV (adverb), as it occurred very commonly in errors in my results. For the NE tag in the gold standard, I noted that being incorrectly tagged as NN was quite common, far more so than any other error in this column. This is in line with Hollenstein & Aepli's (2014) previously noted comment about the NN/NE confusion being the single most common tag confusion (p. 92). The next two most common false tags here were FM and CARD. The only other error worth mentioning,

with less than half the frequency of the previous two, was ADJA. For the ADJA tag, the most common errors by far were NN and CARD, while the only other notable error was ADJD. Common errors for ADJD were ADV, followed at about half that frequency by ADJA. VVFIN and VVPP were also notable errors here. For VVFIN, by far the most common error was VVINF, followed with about half that frequency by VVPP. Next there was FM, which refers to words in a language other than that of the main body of the corpus. The mistakes for FM were overwhelmingly dominated by NN and NE in approximately equal measure, while ADJA was also notable. Finally, for the ADV tag, PTKVZ (a separable prefix, something which does not exist in English) was most common, followed by a near tie between APPR (preposition), ADJD and ADJA.

I also performed an analysis opposite to this, namely choosing each of these tags from the error column and analyzing which tags in the gold standard column they were most often associated with. I included not only the six tags listed in the above paragraph but also NN. For NN, not surprisingly, the related tag NE was the most common correct tag, followed by FM and at a much lower rate by ADJA. For NE, only FM and NN were notable as ground truth tags, with FM at about double the frequency of NN. For ADJA, NN was dominant, followed by FM, NA, and ADJD in that order, with NE and ADV somewhat less common. For ADJD, a less common occurrence than most of the others under consideration here, ADV and ADJA were most common. VVFIN was associated with the closely related VVINF and VVPP most often, followed by ADJD. For FM, NE was overhwelmingly dominant with all others being negligible. Finally, for ADV, ADJD and KON (coordinating conjunction, such as “or” in English) were the two most common, followed by PIS (indefinite pronoun) and PTKREL (indeclinable relative particle).

4.2. Categorical error analysis

I also analyzed tagger errors with a confusion matrix based on broader categories, detailed in Appendix C. These categories were given in a paper on the use of STTS tags for German (Schiller et al. 1999). Although the paper also gave sub-categories within several of these, such as four tags for types of infinitive verbs within the broader verbs

category, for the sake of simplicity I have ignored these for the present work.

The most common categories in this confusion matrix were as follows. I am again dealing here with the first corpus half in the "incorrect" configuration. In counting overall tags in the gold standard, nouns and "others" far surpassed most of the remaining categories, with the next largest being punctuation, followed by verbs. Particle and numbers were particularly rare. In counting erroneous tags, punctuation is by far the most common. This is followed by conjunctions, and more distantly by nouns, adverbs, and pronouns and adjectives. The category least likely to show up as an erroneous tag is particles, followed by numbers, adpositions, articles and "other." The presence in this list of adpositions and articles is interesting, as unlike the other two categories they are not rare categories as gold standard tags.

The general pattern in the confusion matrix developed for broader categories was that tagging a token with a tag from the correct category was far more likely than any particular category error by more than an order of magnitude, even with those categories which had far more errors than others. Further, errors in categories adjacent to that of the gold standard were much more frequent than in categories more distant from it. The ordering of categories was based on the two orderings given by Schiller et al. (1999) and detailed in Appendix C but was altered slightly in response to certain relatively common errors. Namely, I chose to place adjectives and adverbs adjacent to each other, as these are similar parts of speech which are often confused. Particularly in German, the same word can often serve as either part of speech without any alternation, depending on context. I also noted relatively frequent confusion between nouns and adjectives in both directions, and attempted to keep these categories adjacent to each other for this reason. Again, the cause of this confusion is not difficult to imagine given German grammar. Some German nouns, known as adjectival nouns, are almost identical to the relevant adjectives and even inflected as if they were adjectives. My new ordering of categories resulted in errors in more non-adjacent categories for the pronoun gold standard category, but I felt this was acceptable as two of these were adjective and adverb, which hardly seem similar to pronouns.

Another notable error which is explicable by particulars of German grammar is the confusion between nouns and verbs. Some nouns are simply capitalized forms of what

would otherwise be verbs, such as Wissen or Verstand.

The same general pattern, to a greater degree, is notable in the specific confusion matrix, i.e. that which contains every individual tag. Some categories in the gold standard stood out as having particularly few errors, such as articles, numbers and punctuation. However, it was unusual for items in the category "other" to be correctly tagged; they were far more often tagged as punctuation (by more than an order of magnitude). The "other" category had far more tagging errors than any other. "Others" were also commonly tagged as conjunctions, followed by nouns, pronouns, and adverbs, all of which had much higher totals for this particular confusion than most other mistakes. Further, the errors for "other" were more chaotic than most in terms of their distance from the "other" category; perhaps this category simply did not fit in a particular place to the degree that other categories did. Some categories were notable as errors in more gold standard categories than most, for example adjective and adverb. These two categories were also notable as being frequently confused for a wide variety of categories. Pronoun was also a relatively common category, both as an error and a category in the gold standard mistaken for various others. These results are to some degree in line with the "received wisdom that nominal, verbal, and adjectival categories are hard to tag for German" (Hinrichs & Zastrow, 2012, p. 12).

Punctuation tagging has not been the subject of as much attention as more typical POS tagging and tends to be more unpredictable due to the irregular use of some punctuation. In the present work, $(being mistaken for $, was particularly common for the first corpus half, accounting for about 1,000 errors there although it was absent in the second. Both of these tags refer to punctuation which does not end a sentence, in contrast to the sentence-ending "$.". The value of the distinction between the two is not entirely clear, and confusion between the two apparently adds significantly to the error rate. Although this is not reflected in the results discussed here, if I replace every instance of $(with $, I find an increase of approximately 5% in overall accuracy in the first half, from 86.5 to 91.1%. This is quite a dramatic effect; this single change overcomes the majority of the discrepancy between the first and second halves. Therefore, it is conceivable to include an option in future development of the tagger to simply ignore the distinction between the two.

5. Conclusion

The most notable result here was the major difference between results for the first half and the second half of the corpus, which was not surprising due to the greater standardization of the language during the latter. This dwarfed the differences in results caused by changes in training data, that is, between the "specific" and "incorrect" configurations. Also, like those of Brants (2000, p. 228), my accuracy results seem to have reached a point of diminishing returns well before reaching my maximum training size. I estimate that this point was at the fourth fold at the latest, and occurred earlier for the second half than for the first.

Based on its accuracy, this tagger might be sufficient for tagging on historical German text from the second time period, but not the first half with the current configurations.

For future research, given the far greater difficulty in correctly tagging the first half of the corpus, this era of text (1650-1729) deserves more attention in the future, whether this is a matter of improved corpora, tagset alterations, or further tagger configuration. Further, a larger gold standard would be useful; the present paper deals with gold standards which are less than 10% of the largest of the training sets.

Notes

1. The project partners include the Berlin-Brandenbury Academy of Science and Humanities (BBAW), DAASI International GmbH, University of Applied Sciences Worms (FH Worms), the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), the Institute for the German Language (IDS), the Max Planck Institute for the History of Science (MPI WG), the Göttingen State and University Library (SUB), the Technische Universität Berlin, the Technische Universität Darmstadt Department of Linguistics and Literary Studies, and the Julius Maximillian University of Würzburg Institut für Deutsche Philologie.
2. These confusion matrices are available upon request for the interested reader.

Appendix A: STTS-EMG Tags

Durrell et al., 2012, p. 9-10

Tag	Description	Example
ADJA	attributive adjective (including participles used adjectivally)	das **große** Haus die **versunkene** Glocke
ADJD	predicate adjective; adjective used adverbially	der Vogel ist **blau** er fährt **schnell**
ADV	adverb (never used as attributive adjective)	sie kommt **bald**
APPR	preposition left hand part of double preposition	**auf dem Tisch** **an der Straße entlang**
APPRART	preposition with fused article	**am Tag**
APPO	postposition	meiner Meinung **nach**
APZR	right hand part of double preposition	an der Straße **entlang**
ART	article (definite or indefinite)	**die** Tante; **eine** Tante
CARD	cardinal number (words or figures); also declined	**zwei; 526; dreier**
FM	foreign words (actual part of speech in original language may be appended, e.g. FM-ADV/ FM-NN)	**semper fidem**
ITJ	interjection	**Ach!**
KON	co-ordinating conjunction	**oder** ich bezahle nicht
KOKOM	comparative conjunction or particle	**er arbeitet als Straßenfeger** **so gut wie du**

KOUI	preposition used to introduce infinitive clause	**um den König zu töten**
KOUS	subordinating conjunction	**weil er sie gesehen hat**
NA	adjective used as noun	**der Gesandte**
NE	names and other proper nouns	**Moskau**
NN	noun (but not adjectives used as nouns)	**der Abend**
PAV [PROAV]	pronominal adverb	**sie spielt damit**
PAVREL	pronominal adverb used as relative	**die Puppe, damit sie spielt**
PDAT	demonstrative determiner	**dieser Mann war schlecht**
PDS	demonstrative pronoun	**dieser war schlecht**
PIAT	indefinite determiner (whether occurring on its own or in conjunction with another determiner)	**einige Wochen** **viele solche Bemerkungen**
PIS	indefinite pronoun	**sie hat viele gesehen**
PPER	personal pronoun	**sie liebt mich**
PRF	reflexive pronoun	**ich wasche mich** **sie wäscht sich**
PPOSS	possessive pronoun	**das ist meins**
PPOSAT	possessive determiner	**mein Buch** **das ist der meine/meinige**
PRELAT	relative depending on a noun	**der Mann, dessen Lied ich singe** **[...], welchen Begriff ich nicht verstehe**
PRELS	relative pronoun (i.e. forms of der or welcher)	**der Herr, der gerade kommt** **der Herr, welcher nun kommt**
PTKA	particle with adjective or adverb	**am besten, zu schnell, aufs herzlichste**
PTKANT	answer particle	**ja, nein**

PTKNEG	negative particle	**nicht**
PTKREL	indeclinable relative particle	**so**
PTKVZ	separable prefix	**sie kommt an**
PTKZU	infinitive particle	**zu**
PWS	interrogative pronoun	**wer kommt?**
PWAT	interrogative determiner	**welche Farbe?**
PWAV	interrogative adverb	**wann kommst du?**
PWAVREL	interrogative adverb used as relative	**der Zaun, worüber sie springt**
PWREL	interrogative pronoun used as relative	**etwas, was er sieht**
TRUNC	truncated form of compound	**Vor- und Nachteile**
VAFIN	finite auxiliary verb	**sie ist gekommen**
VAIMP	imperative of auxiliary	**sei still!**
VAINF	infinitive of auxiliary	**er wird es gesehen haben**
VAPP	past participle of auxiliary	**sie ist es gewesen**
VMFIN	finite modal verb	**sie will kommen**
VMINF	infinitive of modal	**er hat es sehen müssen**
VMPP	past participle of auxiliary	**sie hat es gekonnt**
VVFIN	finite full verb	**sie ist gekommen**
VVIMP	imperative of full verb	**bleibt da!**
VVINF	infinitive of full verb	**er wird es sehen**
VVIZU	infinitive with incorporated zu	**sie versprach aufzuhören**
VVPP	past participle of full verb	**sie ist gekommen**

Appendix B: STTS Tags

Schiller et al. (1999), p. 6-7

POS =	Beschreibung	Beispiele
ADJA	attributives Adjektiv	*[das] große [Haus]*
ADJD	adverbiales oder prädikatives Adjektiv	*[er fährt] schnell* *[er ist] schnell*
ADV	Adverb	*schon, bald, doch*
APPR	Präposition; Zirkumposition links	*in [der Stadt], ohne [mich]*
APPRART	Präposition mit Artikel	*im [Haus], zur [Sache]*
APPO	Postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	Zirkumposition rechts	*[von jetzt] an*
ART	bestimmter oder unbestimmter Artikel	*der, die, das,* *ein, eine*
CARD	Kardinalzahl	*zwei [Männer], [im Jahre] 1994*
FM	Fremdsprachliches Material	*[Er hat das mit "]* *A big fish [" übersetzt]*
ITJ	Interjektion	*mhm, ach, tja*
KOUI	unterordnende Konjunktion mit "zu" und Infinitiv	*um [zu leben],* *anstatt [zu fragen]*
KOUS	unterordnende Konjunktion mit Satz	*weil, daß, damit,* *wenn, ob*
KON	nebenordnende Konjunktion	*und, oder, aber*
KOKOM	Vergleichspartikel, ohne Satz	*als, wie*
NN	Appellativa	*Tisch, Herr, [das] Reisen*
NE	Eigennamen	*Hans, Hamburg, HSV*
PDS	substituierendes Demonstrativ–pronomen	*dieser, jener*
PDAT	attribuierendes Demonstrativ–pronomen	*jener [Mensch]*
PIS	substituierendes Indefinit–pronomen	*keiner, viele, man, niemand*
PIAT	attribuierendes Indefinit–pronomen ohne Determiner	*kein [Mensch],* *irgendein [Glas]*
PIDAT	attribuierendes Indefinit–pronomen mit Determiner	*[ein] wenig [Wasser],* *[die] beiden [Brüder]*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, dir*
PPOSS	substituierendes Possessiv–pronomen	*meins, deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein [Buch], deine [Mutter]*
PRELS	substituierendes Relativpronomen	*[der Hund,] der*

POS =	Beschreibung	Beispiele
PRELAT	attribuierendes Relativpronomen Relativpronomen	*[der Mann ,] dessen [Hund]*
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PWS	substituierendes Interrogativpronomen	*wer, was*
PWAT	attribuierendes Interrogativpronomen	*welche [Farbe],* *wessen [Hut]*
PWAV	adverbiales Interrogativ– oder Relativpronomen	*warum, wo, wann,* *worüber, wobei*
PAV	Pronominaladverb	*dafür, dabei, deswegen, trotzdem*
PTKZU	"zu" vor Infinitiv	*zu [gehen]*
PTKNEG	Negationspartikel	*nicht*
PTKVZ	abgetrennter Verbzusatz	*[er kommt] an, [er fährt] rad*
PTKANT	Antwortpartikel	*ja, nein, danke, bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am [schönsten],* *zu [schnell]*
TRUNC	Kompositions–Erstglied	*An– [und Abreise]*
VVFIN	finites Verb, voll	*[du] gehst, [wir] kommen [an]*
VVIMP	Imperativ, voll	*komm [!]*
VVINF	Infinitiv, voll	*gehen, ankommen*
VVIZU	Infinitiv mit "zu", voll	*anzukommen, loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen, angekommen*
VAFIN	finites Verb, aux	*[du] bist, [wir] werden*
VAIMP	Imperativ, aux	*sei [ruhig !]*
VAINF	Infinitiv, aux	*werden, sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*[er hat] gekonnt*
XY	Nichtwort, Sonderzeichen enthaltend	*D2XW3*
$,	Komma	,
$.	Satzbeendende Interpunktion	. ? ! ; :
$(	sonstige Satzzeichen; satzintern	– [](){}

Appendix C: STTS Broader Categories

Schiller et al., 1999, pp. 1-2, 4

1. Nomina (N)
2. Verben (V)
3. Artikel (ART)
4. Adjektive (ADJ)
5. Pronomina (P)
6. Kardinalzahlen (CARD)
7. Adverbien (ADV)
8. Konjunktionen (KO)
9. Adpositionen (AP)
10. Interjektionen (ITJ)
11. Partikeln (PTK)

Appendix D:

	Gold Standard											
Error	article	pron.	numb.	verb	noun	adj.	adv.	conj.	adpos.	part.	punct.	other
article	1752	56	0	0	0	0	0	0	0	0	1	0
pron.	9	3088	1	1	4	9	26	10	2	0	0	180
numb.	0	0	202	0	4	23	1	0	0	0	0	0
verb	0	5	1	3184	10	37	6	2	4	1	0	4
noun	6	11	6	20	7258	38	7	1	4	0	0	52
adj.	3	19	5	19	49	1620	27	7	7	0	0	5
adv.	1	18	3	2	13	24	1630	33	5	3	0	71
conj.	0	1	0	1	0	0	13	1384	15	0	0	1404
adpos.	0	0	11	2	0	1	20	3	1624	8	0	0
particle	0	0	0	0	4	1	21	3	12	396	0	4
punct.	0	0	0	0	11	0	0	0	0	0	3813	3433
other	0	0	0	2	14	1	4	1	0	1	0	125

Table 16: Gold standard broader categories for tags and errors, second half, “specific” configuration, averaged over eight iterations.

	Gold Standard											
TnT	article	pron.	numb.	verb	noun	adj.	adv.	conj.	adpos.	part.	punct.	other
article	1754	52	0	0	1	0	0	0	0	0	1	0
pron.	8	3105	0	1	2	8	28	8	2	0	0	181
numb.	0	0	201	0	7	25	0	0	0	0	0	0
verb	0	2	0	3173	7	43	13	2	3	1	0	2
noun	0	11	18	31	7324	51	11	0	3	0	0	51
adj.	3	12	8	25	63	1601	26	1	3	0	0	5
adv.	3	12	1	2	7	23	1614	34	4	3	3	63
conj.	0	1	0	0	2	0	14	1391	16	0	0	1411
adpos.	0	0	0	2	0	1	21	3	1629	8	7	1
part.	0	0	0	0	0	1	23	3	14	396	0	2
punct.	0	0	0	0	10	0	0	0	0	0	3906	3433
other	0	1	0	0	9	0	4	1	0	1	0	129

Table 17: Gold standard broader categories for tags and errors, second half, “incorrect” configuration, averaged over eight iterations.

	Gold Standard											
TnT	article	pronoun	number	verb	noun	adj.	adv.	conj.	adpos.	part.	punct.	other
article	1543	41	0	0	2	0	1	9	0	0	0	2
pronoun	21	2898	0	5	12	21	29	17	0	0	0	162
number	0	1	255	0	44	36	8	0	0	0	0	9
verb	0	5	0	3220	28	42	22	2	6	0	0	14
noun	1	24	9	23	6783	74	14	5	6	0	1	215
adj.	2	29	5	32	88	1528	29	5	3	0	0	38
adv.	0	42	1	6	9	46	1580	44	12	17	0	128
conj.	0	4	0	2	3	0	11	1625	13	0	0	1651
adpos.	0	1	2	2	3	5	16	5	1554	16	0	11
part.	0	1	0	0	1	0	25	1	18	435	0	2
punct.	0	0	0	1	4	1	1	1	0	0	4767	3966
other	0	1	1	0	36	2	6	3	2	2	0	367

Table 18: Gold standard and errors in broad categories, first corpus half, “specific” configuration, averaged over eight iterations.

References

- Andersson, E. (1994). Swedish. In König, E. and van der Auwera, J. *The Germanic Languages* (pp. 271-312). London and New York: Routledge.

- Apache Software Foundation. Apache openNLP – Welcome to Apache openNLP. Retrieved from https://opennlp.apache.org/ on July 5, 2015

- Askedal, J. O. (1994). Norwegian. In König, E. and van der Auwera, J. *The Germanic Languages* (pp. 219-270). London and New York: Routledge.

- Barnes, M.P., and Weyhe, E. (1994). Faroese. In König, E. and van der Auwera, J. *The Germanic Languages* (pp. 190-218). London and New York: Routledge.

- Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik*, *20*(1), 41-67.

- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*(3), 209-226.

- Beißwenger, M. (2007). Corpora zur computervermittelten (internetbasierten) Kommunikation. *Zeitschrift für Germanistische Linguistik*, 35, 496–503.

- Bennett, P., Durrell, M., Scheible, S., & Whitt, R. J. (2009). Annotating a Multi-Genre Corpus of Early Modern German. *Corpus Linguistics 2009*, 20-23.

- Berlin-Brandenburg Academy of Sciences and Humanities. The Deutsches Textarchiv. Retrieved from http://www.deutschestextarchiv.de/.

- Berlin-Brandenburg Academy of Sciences and Humanities. Deutsches Textarchiv (German Text Archive). Retrieved from http://www.bbaw.de/en/research/dta.

- Bird, S. (2006, July). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69-72). Association for Computational Linguistics.

- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. " O'Reilly Media, Inc.".

- Boehm, I. (2005). Unigram Backoff vs. TnT Evaluating Part of Speech Taggers. *Introduction to Computational Linguistics. ByteLABS*.

- Brants, T. (2000a). TnT — A Statistical Part-of-Speech Tagger. In ANLC '00 Proceedings of the sixth conference on applied natural language processing, (pp. 224-231). 10.3115/974147.974178

- Brants, T. (2000b). Inter-annotator Agreement for a German Newspaper Corpus. In *LREC*.

- Brants, T. (2000c, May). TnT – A Statistical Part of Speech Tagger. Saarland University, Computational Linguistics.

- Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. (2002, September). The TIGER treebank. In *Proceedings of the workshop on treebanks and linguistic theories* (Vol. 168).

- Brants, T., Skut, W., & Uszkoreit, H. (2003). Syntactic annotation of a German

newspaper corpus. In *Treebanks* (pp. 73-87). Springer Netherlands.

- Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., & Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, *2*(4), 597-620.

- Carreras, X., Chao, I., Padró, L., & Padró, M. (2004, May). FreeLing: An Open-Source Suite of Language Analyzers. In *LREC*.

- Chambers, W. W. and Wilke, J. R. (1970). A Short History of the German Language. London: Methuen & Co. Ltd.

- Clark, J. W. (1957). *Early English: A Study of Old and Middle English*. New York: W.W. Norton & Company, Inc.

- Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992, March). A practical part-of-speech tagger. In *Proceedings of the third conference on Applied natural language processing* (pp. 133-140). Association for Computational Linguistics.

- Daelemans, W., Zavrel, J., Berck, P., & Gillis, S. (1996). MBT: A memory-based part of speech tagger-generator. *arXiv preprint cmp-lg/9607012*.

- DeSchutter, G. (1994). Dutch. In König, E. and van der Auwera, J. *The Germanic Languages* (pp. 439-477). London and New York: Routledge.

- Demske, U. 2007. "Das MERCURIUS-Projekt: Eine Baumbank für das Frühneuhochdeutsche." Sprachkorpora: Datenmengen und Erkenntnisfortschritt. Kallmeyer and Zifonun (eds). Berlin: Walter de Gruyter, 91-104.

- Deutsch Diachon Digital. Referenzkorpus Altdeutsch. Retrieved from http://www.deutschdiachrondigital.de/.

- Demske, U. Media: Treebank of Early New High German. Linguist List. Oct 29 2014. Retrieved from https://linguistlist.org/issues/25/25-4307.html.

- Dipper, S. (2010, September). POS-tagging of historical language data: First experiments. In *Semantic Approaches in Natural Language Processing. In: Proceedings of the 10th Conference on Natural Language Processing (KONVENS 2010)* (pp. 117-121).

-Donaldson, B. (2008). *Dutch: A Comprehensive Grammar (2nd ed.).* London and New York: Routledge.

- Dredze, M., & Wallenberg, J. (2008, June). Icelandic data driven part of speech tagging. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*(pp. 33-36). Association for Computational Linguistics.

- Dudenredaktion (2000). Duden: die deutsche Rechtschreibung (22nd ed.). Mannheim: Dudenverlag.

- Eisenberg, P. et al. (eds.) (1998). Duden: Grammatik der deutschen Gegenwartssprache (6th ed.) Mannheim: Dudenverlag.

- Durrell, M., Bennet, P., Scheible, S., and Whitt, R.J. School of Languages, Linguistics and Cultures, University of Manchester. (2012, March 31). *The GerManC Corpus.* Manchester.

- Einarsson, Stefán (1947). *Icelandic: Grammar, Texts, Glossary*. Baltimore and London: Johns Hopkins University Press.

- Eisenberg, P. (1994). German. In König, E. and van der Auwera, J. *The Germanic Languages* (pp. 349-387). London and New York: Routledge.

- Ernst-Gerlach, A., & Fuhr, N. (2006). Generating search term variants for text collections with historic spellings. In *Advances in Information Retrieval* (pp. 49-60). Springer Berlin Heidelberg.

- Ernst-Gerlach, A., & Fuhr, N. (2007, June). Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 333-341). ACM.

- Ernst-Gerlach, A. (2013). *Retrievalmethoden für historische Korpora mit nicht standardisierten Schreibweisen* (Doctoral dissertation, Universität Duisburg-Essen, Fakultät für Ingenieurwissenschaften» Informatik und Angewandte Kognitionswissenschaft).

- European Bureau of Library, Information and Documentation Associations. Digital Agenda for Europe: Digital Libraries Initiative. Retrieved from http://www.eblida.org/news/digital-agenda-for-europe-digital-libraries-initiative.html.

- Faarlund, J. T. (1994). Old and Middle Scandinavian. In König, E. and van der Auwera, J. *The Germanic Languages* (pp. 38-71). London and New York: Routledge.

- Gesmundo, A., & Samardžić, T. (2012a, July). Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 368-372). Association for Computational Linguistics.

- Gesmundo, A., & Samardžić, T. (2012b). Lemmatising Serbian as Category Tagging with Bidirectional Sequence Classification. In *LREC* (pp. 2103-2106).

- Giesbrecht, E., & Evert, S. (2009, September). Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German Web as corpus. In *Proceedings of the Fifth*

Web as Corpus Workshop (pp. 27-35).

- Giménez, J., & Màrquez, L. (2004). Fast and accurate part-of-speech tagging: The SVM approach revisited. *Recent Advances in Natural Language Processing III*, 153-162.

- Google Books. Retrieved from https://books.google.com/?hl=en.

- Gotscharek, A., Reffle, U., Ringlstetter, C., & Schulz, K. U. (2009, September). On lexical resources for digitization of historical documents. In *Proceedings of the 9th ACM symposium on Document engineering* (pp. 193-200). ACM.

- Haberland, H. (1994). Danish. In König, E. and van der Auwera, J. *The Germanic Languages*. (pp. 313-348). London and New York: Routledge.

- Halácsy, P., Kornai, A., & Oravecz, C. (2007, June). HunPos: an open source trigram tagger. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions* (pp. 209-212). Association for Computational Linguistics.

- HathiTrust Digital Library. Retrieved from http://www.hathitrust.org/about.

- Hauser, A., Heller, M., Leiss, E., Schulz, K. U., & Wanzeck, C. (2007). *Information access to historical documents from the early new high german period*. Internat. Begegnungs-und Forschungszentrum für Informatik.

- Helgadóttir, S. (2004). Testing data-driven learning algorithms for pos tagging of icelandic. *Nordisk sprogteknologi*, 257-265.

- Henrisken, C., and van der Auwera, J. (1994). The Germanic Languages. In König, E. and van der Auwera, J. *The Germanic Languages* (pp. 1-18). London and New York: Routledge.

- Hinrichs, E., & Zastrow, T. (2012). Linguistic annotations for a diachronic corpus of German. *Linguistic Issues in Language Technology*, *7*(1).

- Hollenstein, N., & Aepli, N. (2014). Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging. *COLING 2014*, 85.

- Handke, J. (2012, August). Natural Language Processing: The Structure of the Lexicon: Human Versus machine. Walter De Gruyter.

- Ide, N., & Macleod, C. (2001). The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001* (Vol. 3).

- Institut für Deutsche Sprache. Historisches Korpus. Retrieved from http://www1.ids-mannheim.de/lexik/abgeschlosseneprojekte/historischeskorpus/historisches-korpus.html.

- Johnson, S. A. (2005). *Spelling trouble?: language, ideology and the reform of German orthography*. Multilingual Matters.

- Jurish, B. (2010, July). Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology* (pp. 72-77). Association for Computational Linguistics.

- Kroch, A., Santorini, B. and Diertani, A. (2004). Penn-Helsinki Parsed Corpus of Early Modern English. Retrieved from http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html

- König, E. and van der Auwera, J. (1994). *The Germanic Languages*. London and New York: Routledge.

- Kübler, S., & Baucom, E. (2011, September). Fast Domain Adaptation for Part of Speech Tagging for Dialogues. In *RANLP* (pp. 41-48).

- Langendoen, D. T. (1997). English for the computer: The SUSANNE corpus and analytic scheme-Sampson, G.

- LAUDATIO (Long-term Access and Usage of Deeply Annotated Information). Retrieved from http://www.laudatio-repository.org.

- Lavergne, T., Cappé, O., & Yvon, F. (2010, July). Practical very large scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 504-513). Association for Computational Linguistics.

- Light, C. (2015) [this is the latest copyright date; creation date unclear] Parsed Corpus of Early New High German. Retrieved from https://enhgcorpus.wikispaces.com/.

- Lin, X. (2003, August). Impact of imperfect OCR on part-of-speech tagging. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on* (pp. 284-288). IEEE.

- Loftsson, H. (2006). Tagging Icelandic text: An experiment with integrations and combinations of taggers. *Language Resources and Evaluation*, *40*(2), 175-181.

- Loftsson, H. (2007, April). Tagging Icelandic text using a linguistic and a statistical tagger. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers* (pp. 105-108). Association for Computational Linguistics.

- Loftsson, H. (2008). Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, *31*(01), 47-72.

- Loftsson, H., Yngvason, J. H., Helgadóttir, S., & Rögnvaldsson, E. (2010). Developing a PoS-tagged corpus using existing tools. In *Proceedings of "Creation and use of basic*

lexical resources for less-resourced languages", workshop at the 7th International Conference on Language Resources and Evaluation, LREC.

- Loper, E., & Bird, S. (2002, July). NLTK: The natural language toolkit. In*Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*(pp. 63-70). Association for Computational Linguistics.

- Lüdeling, A., Hirschmann, H., & ZELDES, A. (2011). Variationism and Underuse Statistics in the Analysis of the Development of Relative Clauses in German. *Corpus-based Analysis and Diachronic Linguistics*, *3*, 36.

- Lúthersson, S.K. (2010). *Tagging and parsing a large corpus: Research report* (national ID: 071183-2119). Reykjavik University, Department of Computer Science. Retrieved June 2015 from http://www.ru.is/~hrafn/students/BScThesis_taggingParsingLargeCorpus.pdf

- Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2), 313-330.

- Martin Luther University Halle-Wittenberg. Referenzkorpus Frühneuhochdeutsch. Retrieved from http://www.germanistik.uni-halle.de/forschung/altgermanistik/referenzkorpus_fruehneuhochdeutsc/.

- McWhorter, J. (2008). Our Magnificent Bastard Tongue: The Untold History of English. New York: Gotham Books.

- Mihov, S., & Schulz, K. U. (2004). Fast approximate search in large dictionaries. *Computational Linguistics*, *30*(4), 451-477.

- Moeller, J., Adolph, W.R., Mabee, B., & Berger, S. (2007). *Kaleidokop: Kultur, Literatur*

und Grammatik (7th ed.). Boston; New York: Houghton Mifflin Company.

- Neijmann, D. L. (2001). *Colloquial Icelandic: The Complete Course for Beginners.* New York; Oxford: Routledge.

- Neunerdt, M., Trevisan, B., Reyer, M., & Mathar, R. (2013). Part-of-speech tagging for social media texts. In *Language Processing and Knowledge in the Web* (pp. 139-150). Springer Berlin Heidelberg.

- Ngai, G., & Florian, R. (2001, June). Transformation-based learning in the fast lane. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). Association for Computational Linguistics.

- Norwegian Infrastructure for the Exploration of Syntax and Semantics (INESS). Retrieved from http://clarino.uib.no/iness.

- Nugues, P. M. (2006). *An introduction to language processing with Perl and Prolog: An outline of theories, implementation, and application with special consideration of English, French, and German.* Berlin, Heidelberg: Springer. ISBN: 978-3-540-25031-9 (Print) 978-3-540-34336-3 (Online) Cognitive Technologies, 2006.

- Open Content Alliance. Retreived from http://www.opencontentalliance.org/.

- Petrov, S., Barrett, L., Thibaux, R., & Klein, D. (2006, July). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 433-440). Association for Computational Linguistics.

- Pilz, T., Luther, W., Fuhr, N., & Ammon, U. (2006). Rule-based search in text databases with nonstandard orthography. *Literary and Linguistic Computing*,*21*(2), 179-186.

- Pilz, T., & Luther, W. (2009). Automated support for evidence retrieval in documents with nonstandard orthography. *The Fruits of Empirical Linguistics. Sam Featherston and Susanne Winkler (eds.)*, 211-228.

- Pind, J., Magnússon, F., & Briem, S. (1991). Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]. *The Institute of Lexicography, University of Iceland, Reykjavik, Iceland*.

- Piotrowski, M. (2012, September). Natural language processing for historical texts. Synthesis Lectures on Human Language Technologies, 5(2), 1-157. doi:10.2200/S00436ED1V01Y201207HLT017

- Poel, M., & Boschman, E. (2008). A Neural Network Based Dutch Part of Speech Tagger.

- Project Gutenberg. Retrieved from http://www.gutenberg.org.

- Project Gutenberg-DE. Retrieved from http://gutenberg.spiegel.de/.

- Rajimwale, S. (2006). *Handbook of Linguistic Terms*. Sarup & Sons.

- Ratnaparkhi, A. (1996, May). A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing* (Vol. 1, pp. 133-142).

- Rayson, P., Archer, D., & Smith, N. (2005). VARD versus WORD: A comparison of the UCREL variant detector and modern spellcheckers on English historical corpora.

- Rayson, P., Archer, D., Baron, A., & Smith, N. (2006). Tagging Historical Corpora-the problem of spelling variation. In *Digital Historical Corpora*.

-Rayson, P., Archer, D., Baron, A., Culpeper, J., & Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora.

-Rehbein, I., & van Genabith, J. (2007). Why is it so difficult to compare treebanks? TIGER and TüBa-D/Z revisited. In: TLT 2007 - The 6th International Workshop on Treebanks and Linguistic Theories, 7-8 December, 2007, Bergen, Norway.

- Ruge, N. (2005). Zur morphembezogenen Überformung der deutschen Orthographie. *Linguistik online*, *25*(4/05), 66.

- Sampson, G. (1993). The SUSANNE corpus. *ICAME Journal*, *17*(125127), 116.

- Santorini, B. (2010, January). Annotation manual for the Penn Historical Corpora and the PCEEC. (2nd. ed.) Retrieved from http://www.ling.upenn.edu/hist-corpora/annotation/.

- Scheible, S., Whitt, R. J., Durrell, M., & Bennett, P. (2011a, June). Evaluating an 'off-the-shelf' POS-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 19-23). Association for Computational Linguistics.

- Scheible, S., Whitt, R. J., Durrell, M., & Bennett, P. (2011b, June). A gold standard corpus of Early Modern German. In *Proceedings of the 5th linguistic annotation workshop* (pp. 124-128). Association for Computational Linguistics.

- Scheible, S., Whitt, R. J., Durrell, M., & Bennett, P. (2012). GATEtoGerManC: A GATE-based Annotation Pipeline for Historical German. In *LREC* (pp. 3611-3617).

- Scherrer, Y., & Rambow, O. (2010). Natural language processing for the Swiss German dialect area.

- Scherrer, Y. (2011). Morphology Generation for Swiss German Dialects. In*Systems and Frameworks for Computational Morphology* (pp. 130-140). Springer Berlin Heidelberg.

- Scherrer, Y. (2012). Machine translation into multiple dialects: The example of Swiss German. In *7th SIDG Congress-Dialect 2.0*.

- Schiller, A., Teufel, S., & Thielen, C. (1995). Guidelines für das Tagging deutscher Textcorpora mit STTS. *Manuscript, Universities of Stuttgart and Tübingen*, *66*.

- Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. Schmid, H. (1994, September). In *Proceedings of the international conference on new methods in language processing*, *12*, 44-49.

- Schmid, H. (1995). Improvements in Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*.

- Schnell, R., Bachteler, T., & Bender, S. (2004). A toolbox for record linkage. *Austrian Journal of Statistics*, *33*(1-2), 125-133.

- Shen, L., Satta, G., & Joshi, A. (2007, June). Guided learning for bidirectional sequence classification. In *ACL* (Vol. 7, pp. 760-767).

- Skut, W., Brants, T., Krenn, B., & Uszkoreit, H. (1998). A linguistically interpreted corpus of German newspaper text. *arXiv preprint cmp-lg/9807008*.

- Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997, March). An annotation scheme for free word order languages. In *Proceedings of the fifth conference on Applied natural language processing* (pp. 88-95). Association for Computational Linguistics.

- Sonderforschungsbereich 632. ANNIS. Retrieved from http://korpling.german.hu-berlin.de/annis3/ddd.

- Strunk, J. (2003). Information retrieval for languages that lack a fixed orthography. *Linguistics Department, Stanford University, California*.

- Telljohann, H., Hinrichs, E., Kübler, S., & Kübler, R. (2004). The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

- Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., & Beck, K. (2012, January). Stylebook for the Tübingen treebank of written German (TüBa-D/Z). In*Seminar fur Sprachwissenschaft, Universitat Tubingen, Tubingen, Germany*.

-Text Creation Partnership. Retrieved from http://www.textcreationpartnership.org/.

- Text Creation Partnership. Eighteenth Century Collections Online. Retrieved from http://quod.lib.umich.edu/e/ecco/.

- Text Creation Partnership. Early English Books Online. Retrieved from http://eebo.chadwyck.com/marketing/about.htm.

-Text Grid: Virtuelle Forschungsumgebung für die Geisteswissenschaften. Retrieved from https://www.textgrid.de/.

- Thráinsson, Höskuldur. (1994). Icelandic. In König, E. and van der Auwera, J. *The Germanic Languages* (pp. 142-189). London and New York: Routledge.

- Toutanova, K., Klein, D., Manning, C.D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics 1* (pp. 173-180).

- Tsuruoka, Y., Tateishi, Y., Kim, J. D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. I. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Advances in informatics* (pp. 382-392). Springer Berlin Heidelberg.

- Tufte, E. R., & Graves-Morris, P. R. (1983). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

- UK Data Service. UK Data Service data catalogue record for: GerManC: A historical corpus of German texts, 1650-1800. Retrieved from http://discover.ukdataservice.ac.uk/catalogue?sn=7021.

- Universität Duisburg-Essen, Korpora.org. Das Bonner Frühneuhochdeutschkorpus. Retrieved from http://www.korpora.org/Fnhd/.

- University of Manchester, School of Arts, Languages and Cultures. The GerManC project: A representative historical corpus of German 1650-1800. Retrieved from http://www.alc.manchester.ac.uk/subjects/german/research/projects/germanc/.

- Van Baalen, C., Blom, F. R.E., & Hollander, I. (2012). *Dutch for Reading Knowledge*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

- Van der Wal, M. J., and Quak, A. (1994). Old and Middle Continental West Germanic. In König, E. and van der Auwera, J. *The Germanic Languages* (pp. 72-109). London and New York: Routledge.

- Van Halteren, H., Zavrel, J., & Daelemans, W. (1998, August). Improving data driven wordclass tagging by system combination. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1* (pp. 491-497). Association for Computational Linguistics.

- Volk, M., & Schneider, G. (1998). Comparing a statistical and a rule-based tagger for German. *arXiv preprint cs/9811016*.

- Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., and Rögnvaldsson, E.. 2011. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. http://www.linguist.is/icelandic_treebank

- Wikimedia Foundation. Wikisource. Retrieved from http://wikisource.org/wiki/Main_Page.

- Wikimedia Foundation. Wikisource (German). *Lutherbibel*. Retrieved from http://de.wikisource.org/wiki/Lutherbibel.

- Wilcock, G. (2009). *Introduction to Lingusitic Annotation and Text Analytics*. Morgan & Claypool Publishers.

- Willemyns, R. (2013). *Dutch: Biography of a Language*. Oxford: Oxford University Press.

- Witten, I.H., Frank, E., and Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

- Zoëga, Geir T. (1926*). A Concise Dictionary of Old Icelandic*. Oxford: Oxford University Press.