Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2016

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2016

Building a Gold Standard for Temporal Entity Extraction from Medieval German Texts

Natalia Korchagina

Institute of Computational Linguistics
University of Zurich
Binzmühlestrasse 14, 8050 Zürich
korchagina@cl.uzh.ch

Abstract

We present a corpus of gold standard annotation of temporal entities in Early New High German texts. This resource addresses the lack of a gold standard temporal annotation for historical German. Such a corpus is necessary for our research. The ultimate goal of our project is to develop an effective system for temporal entity extraction from historical texts. The manually annotated corpus will serve as base for quality estimation of the temporal annotation produced during the experiments.

1. Introduction

Time is a crucial dimension not only in information processing, but in humanities as well, e.g., a description of a person or a place should contain temporal terms. Not only modern texts, but also historical texts may be rich in temporal expressions. Manual extraction of this information is time-consuming, therefore some facts might still be undiscovered, and thus unknown to scientists. This research will contribute to the development of a tool for temporal entity extraction from historical texts, assisting historical text-mining.

Typical application examples exploiting temporal tagging include information extraction, i.e., the described events are summarized and chronically ordered; and information retrieval, where time is used as a query topic. Temporal annotation of historical texts would allow the digital humanities community to benefit from both scenarios, enabling a faster analysis and a time-framed search through the ever-growing amount of historical corpora available in digital form.

The project is funded by the Swiss Law Sources Foundation. As material for our research, we use historical legal texts (i.e., decrees, regulations, court transcripts) kindly provided by the Foundation. This organization has been publishing critical editions of Swiss historical legal texts in German, French, Italian, Romansh, and Latin for over a hundred years. By today 28 of 118 published volumes are available for digital processing, with a roughly estimated total of 7 million tokens of historical data. The texts' creation time ranges from the 10th to the 18th century. The biggest part of the available digital data is in German, therefore in this project we work with German texts.

There are systems for temporal information extraction from modern texts. Effective taggers such as SUTime (Chang and Manning, 2012) and HeidelTime (Strötgen and Gertz, 2013) use handcrafted rules and dictionaries for recognition and normalisation of temporal expressions. Applied to a corpus of modern narrative texts, WikiWarsDE, HeidelTime achieves f-scores of 91.3 and 85.8 for the extraction (lenient and strict, respectively) (Strötgen and Gertz, 2011). However, the application of an off-the-shelf tool developed for a modern language to a historical corpus is unlikely to lead to good results. Scheible et al. (2011) evaluated the TreeTagger (Schmid, 1994) developed for modern German on Early Modern German corpus, achieving a tagging accuracy of 69.6%, which is far from the 97% reported for modern German.

Lexical and spelling differences are some of the most evident properties of historical texts and a substantial obstacle to the application of the existing NLP tools. The example below shows some of the manually extracted expressions meaning or referring to "evening" ("Abend" in modern German).

Abend	abentt	stübgloge
abende	abentts	zenacht
abends	abent	gessen[2]
abendes	aebent	zenacht essen
ábend	aebents	znacht essen
aubent	abentz	Nacht essen
aubend	stübglogge[1]	nachtessen
aubends	stübgloggen	schlaff trunck[3]

Example 1: Expressions in medieval German with the meaning "evening".

The most common approach for dealing with the non-standard spelling is normalisation, i.e., the process of mapping historical word forms to their modern equivalents. After spelling normalisation, expressions in the range of "Abend" – "abendes" in the example above will be recognized, while those like "stübglogg" will remain unidentified because they are no longer used in temporal context or disappeared from the modern language, and thus there is no pattern to be

[1] Betzeitglocke am Abend [*en*: Bedtime bells ringing in the evening]. Schweizerisches Idiotikon – Wörterbuch der schweizerdeutschen Sprache, "S t äu b (i) g l o g g (e^n)" (II, Sp. 617), 1885.

[2] Abendbrot [*en*: supper]. Schweizerisches Idiotikon – Wörterbuch der schweizerdeutschen Sprache, "Z (e) n a c h t ë s s e^n" (I, Sp. 527), 1881.

[3] Trunk, <...>, vor dem Schlafgehen eingenommen [*en*: Nightcap, bedtime drink]. Schweizerisches Idiotikon – Wörterbuch der schweizerdeutschen Sprache, "S c h l ä f f t r u n k" (XIV, Sp. 1212), 1987.

matched in the set of rules. To overcome these limitations, at the second stage of our experiments we will use statistical methods capable to learn possible patterns of temporal expressions from a manually annotated Gold Standard corpus.

This paper describes the creation of a Gold Standard sample corpus (of about 32,000 tokens) of Early New High German containing manual annotations of temporal entities. This corpus is used in our research as base for quality estimation of temporal tagging at all stages of our experiments. Section 2 introduces the contents and design of the corpus. In Section 3, we will describe the annotation process and summarize our experience of the adaptation of the annotation guidelines for historical data. The first stage of experiments based on the corpus will be presented in Section 4.

2. Corpus Design

Two major types of documents are present in our data: legal cases and transactional documents. Legal cases describe incidents of the law violation and legal consequences that followed. Documents of this time contain, e.g., date and time when the event took place. Transactional documents represent contracts, sales agreements, and purchases. They are especially reach in temporal information, important for the legal value of the document. Schilder and McCulloh (2005) mention the following kinds of temporal information in transactional documents: the date when the transaction takes effect, the execution date, and duration clauses.

For the Gold Standard annotation, we selected manually 50 articles, corresponding to various kinds of legal documents described above. The texts were taken from 9 volumes, representing 5 Swiss cantons. This set of texts covers the period between 1450 and 1550. This particular period of time was chosen, first, because of a large number of articles created at this period (total of 4,175 articles were created between 1450 and 1550), available in digital format in the collection of the Swiss Law Sources Foundation. If our preliminary experiments will prove to be effective, larger datasets from the same period may be involved for further experiments. Figure 1 shows the distribution of the number of articles regarding the year they were issued.

Although the biggest amount of the articles belongs to the year 1425, after a closer examination of the contents of these texts we opted for a later period of time. In order to create a properly annotated Gold Standard corpus, it is important for annotators and supervisors of the task to understand well the contents of the corpus. Articles written before the second half of the 15th century were very complicated to understand even for native speakers of German, therefore the second reason for our choice of period is the language state.

The 50 chosen articles are relatively evenly distributed between 1450 and 1550. The choice of material was motivated by the idea to optimize our system for work on diachronically close texts, yet capturing a certain variety in language state due to their diverse origins. We realize that a system adapted for recognition of temporal expressions in the material from a particular period will show lower results, if applied to a text from another period of German, as spelling is period dependent. Our research should be seen as experimental ground, attempting to find a state-of-the-art method for recognition of temporal information in historical texts.

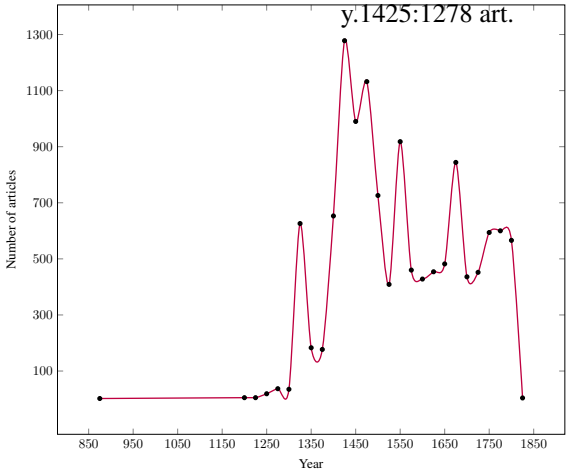

Figure 1: Distribution of the articles available in digital format with regard to the year they were issued.

Period	1450 – 1550
Language period	Early New High German
Domain	legal
Number of articles	50
Aver. length of an art., tokens	950
Total, tokens	32338

Table 1: Characteristics of the Gold Standard corpus.

3. Annotation Process and Results

3.1. Dataset Annotation

To facilitate the annotation task for the human annotators, the corpus was first processed with the rule-based temporal tagger HeidelTime adapted for the Text+Berg corpus (Rettich, 2013), (Volk et al., 2010) containing Swiss alpine texts from 1864 to 2009. Instead of the default configuration for German, the adapted version of the tagger was used because it covers some diachronic variation since the 19th century, and thus the chances of a successful extraction of temporal expressions were higher. When HeidelTime with the adapted set of the resources was applied to our corpus, 200 text segments were identified as temporal expressions.

The automatically annotated corpus was verified manually by two annotators. Their task was to correct erroneous annotations and add missing tags. For our annotation, we adopted a customized XML-like format based on TimeML language (Pustejovsky et al., 2003). It is a robust specification language for events and temporal expressions in natural language text. Several tags and their attributes are defined in TimeML adressing event markup, time stamping of events, ordering of events in time, reasoning with contextually underspecified temporal expressions (e.g., “ten days”) and reasoning about the persistence of events.

Since the work time was limited to 40 hours, it was important to provide annotators with concisely written guide-

lines, strictly relevant to their assignment. Although our instructions were based on the existing guidelines (Saurí et al., 2006) for temporal and event annotations in TimeML standard, it was not possible to reuse them entirely. First, these guidelines are very detailed: they contain descriptions of many attributes and tags which may be of use for the development of a more complex system. Second, being developed for the annotation of modern texts, the TimeML guidelines do not reflect particularities of historical corpora.

Our guidelines covered the two most important points: what to annotate and how. First, different kinds of temporal expressions (explicit/implicit/relative/markable/non-markable) were introduced. According to the TimeML annotation guidelines, only markable expressions (which can be situated on a timeline) should be annotated. Non-markable expressions are less amenable to being situated on a timeline, e.g., later, previous, sooner. It is difficult to understand old German texts even for a native speaker of German. In order not to miss a markable expression, considering it to be a non-markable one, the annotators were asked to tag any expressions with temporal semantics.

The annotators used a subset of the TimeML mark-up language, as it was implemented in HeidelTime, i.e., temporal expressions were tagged with TIMEX3 tags. In addition, the annotators were to put SIGNAL tags to mark a token signaling a relationship between two temporal expressions, and non-consuming (meta-tags, not containing any text) TIMEX3 tags detailing this relationship. These features also belong to the TimeML language. According to our guidelines, each TIMEX3 tag should contain three compulsory attributes: ID number, type and value. The TimeML standard distinguishes four different types of temporal expressions:

- DATE, for expressions describing a calendar date, e.g., December 25, 2015;
- TIME, for expressions referring to a time of the day, e.g., half past midnight;
- DURATION, for expressions describing a duration, e.g., three days;
- SET, for expressions describing a set of times, e.g., twice a week.

The value attribute specifies which temporal information is contained in the tagged span of text. In the TimeML guidelines, the value attribute should be in the form of an ISO8601 format for date and time, supposing that only markable expressions are to be annotated. Following our guidelines, the annotators were allowed to underspecify the value attribute for cases when the meaning of an expression is not entirely clear, e.g., use *"XXXX-12-25"* for "Christmas" of a year unknown from the context, or *"XXXX"* for non-markable expressions. Annotators were asked to pay special attention to the tagging of the saint feast days. They represent a large part of the temporal information in historical texts from this period and were often used instead of the calendar dates. Due to the spelling variation none of these expressions was detected during the automatic annotation using HeidelTime. Annotators were asked to assign a value attribute in ISO8601 format for fixed feasts (they normally refer to a feast of a particular saint). As for moveable feasts, i.e., relative to the Easter Sunday of a particular year, annotators could underspecify the value attribute.

old de: Verkunden uff den 8 januarii anno etc. 1545
mod. de: Verkünden auf den 8 Januar Jahr etc. 1545

en: 'Announced on the 8 January year etc. 1545'

Verkunden uff den <TIMEX3 tid="t620"
type="DATE" value="1545-01-08" >
8 januarii anno etc. 1545 </TIMEX3>.

Figure 2: Sentence in historical German (old de), its modernised spelling (mod. de), translation into English (en) and temporal annotation.

The example above shows an annotated sentence from the Gold Standard, preceded by a gloss pairing the original sentence in Early New High German with its modern equivalent and a translation into English.

3.2. Dataset Analysis

After the annotation process was finished, we calculated the inter-annotator agreement. We present values for average observed agreement and chance-corrected agreement (Cohen's Kappa) in Table 2. Relaxed matches of text spans were allowed during the calculation of the agreement on the detection of temporal expressions.

	Detection	Classification
Average observed agreement	0.75	0.89
Cohen's Kappa	0.74	0.76

Table 2: Inter-annotator agreement values.

The inter-annotator agreement values in Table 2 show that temporal entity annotation in historical texts is a highly context dependent task, and detection of a temporal expression is the most difficult part of it. Identification of a certain expression as temporal requires a thorough understanding of the context. For instance, the word *"jarzit"*, because of its similarity with "Jahreszeit" in modern German (*en* "season"), was tagged by one of the annotators as temporal expression of the type DURATION. However, in the given context *"jarzit"* should be normalised to "Jahrzeit" in modern German, referring to the event of the commemoration of a deceased person, and therefore not being a proper temporal expression.

The annotation process was finished by adjudicating the annotations, i.e., deciding which annotations should be kept in the resulting Gold Standard. According to (Pustejovsky and Stubbs, 2013), the adjudication process should be performed by those who were involved in creating the annotation guidelines, as they will have the best understanding of the annotation purpose. For this reason, the adjudication was performed by the author of the paper. The following features of each tag were adjudicated: extent of the tagged temporal expression, type and value. The tag extent was judged based

on the general rule of span economy: the tagged expressions should contain the smallest number of tokens needed to identify it as temporal expression of a particular type. For example, "1. Januar" (*en*: 1st of January) is preferred to "am 1. Januar" (*en*: on the 1st of January).

Table 3 presents the comparison between the adjudicated Gold Standard and its predecessors, i.e., annotations produced: 1) automatically by HeidelTime (HT); 2) by human annotators (A1 and A2).

	R	P	F	Type	Value
HT	0.26	0.96	0.41	96%	81%
A1	0.93	0.95	0.94	91%	84%
A2	0.88	0.90	0.89	95%	79%

Table 3: Annotation produced by a rule-based system (HT) and manual annotations (A1, A2) evaluated against the Gold Standard. Recall/precision/f-measure scores are calculated for tag extraction, whereas scores in "Type" (correctly classified) and "Value" (correctly normalised) columns are calculated based on the correctly extracted expressions.

4. Experiments Based on the Gold Standard

Several projects in the recent years applied normalisation techniques for the tasks of information extraction.

In (Pettersson et al., 2014) various methods of normalisation (i.e., rule-based, dictionary-based, Levenshtein-based, and based on statistical machine translation) are evaluated d to the task of the verb phrase extraction from Early Modern Swedish texts. The best scores for normalisation and subsequent verb phrase extraction (92.9% accuracy and 87.5 F-score respectively) were achieved by the character-based machine translation approach. Logačev et al. (2014) used a normalisation method based on weighted edit distances to improve part of speech tagging (POS) of several Early New High German texts. The tagging accuracy improved by the average of 2% for the normalised texts. We will follow the steps of these researchers and observe, to what extent the performance of a temporal tagger developed for modern texts can be improved by using normalisation as a pre-processing step.

We started our trial of spelling normalisation methods by an edit-distance based technique described in (Pettersson et al., 2013), used to improve the performance of existing NLP tools (developed for the modern language) for the task of verb extraction from historical Swedish texts, allowing to improve recall from 64.2% for unnormalised text to 86.2%. This approach benefits from context-sensitive weights (lower than 1) for commonly occurring edits and a threshold value for a dictionary entry to be considered as a normalisation candidate, both learned from a parallel corpus of manually normalised data. The only resource for historical German containing relatively large amount of manually normalised data is the GerManC corpus including texts from the period 1650–1800 (Scheible et al., 2011). The normalised subset of this corpus belongs to the period 1659–1780 and contains about 50,000 tokens. We normalised the Gold Standard corpus applying the edit-based method with context-sensitive weights and threshold value for candidates learned from the GerManC parallel data. Table 4 presents the results of the temporal annotation.

	R	P	F	Type	Value
HT	0.27	0.89	0.41	96%	85%

Table 4: Evaluation of the temporal annotation produced by HeidelTime after the Gold Standard corpus was normalised with a weighted edit-distance technique.

After normalisation, the recall value improved by only 0.01 point, while precision even dropped from 0.96 to 0.89, compared to similar values in Table 3 for the annotation produced by temporal tagger on the Gold Standard before normalisation. From this experiment we concluded that the use of manually normalised resources on texts from a slightly different period of time does not produce a positive effect on the output of the temporal tagger.

5. Conclusion

In this paper we described the process of creation of a Gold Standard corpus of Early New High German, containing manual annotation of temporal entities. Given the absence of similar corpora for historical German of this period, the creation of this annotation was a necessary step in the development of a temporal entity extraction system for historical texts. The Gold Standard corpus is used for quality estimation of the automatically produced temporal annotation. Detection of temporal expressions is a difficult task d to a high lexical and spelling variation in our data, therefore, at this point of our research, we are interested in the ability of the temporal entity extraction system to identify temporal expressions in text, reflected in the recall values. First, we evaluated the performance of the rule-based temporal tagger HeidelTime (with modern resources enhanced with the resources adapted for the Text+Berg corpus, 1964-2009) against our Gold Standard and obtained the recall of 0.26. In attempt to reduce spelling variation preventing temporal tagger developed for contemporary German from recognizing temporal expressions in our Gold standard, at the first stage of our experiments we opted for the spelling normalisation approach. We normalised the Gold Standard corpus using an edit distance metric with context-sensitive weights learned from the manually normalised subset of the GerManC corpus. The recall value after tagging the normalised text only reached 0.27. We assume, that such a little improvement is due to the fact that the subset of the GerManC corpus used for learning edit weights and the threshold for the dictionary candidates matching, belongs to a later state of German, covering the period from 1659–1780, whereas our Gold Standard contains text from 1450 to 1550.

Future work includes further evaluation of various normalisation techniques. After the best-performing normalisation approach or combination of methods will be defined, we will manually correct a portion of the output. We will then apply a modern temporal extraction system on the manually normalised subset of the Gold Standard in order to establish, to what extent spelling normalisation can improve

temporal tagging. We expect a certain portion of expressions not to be matched after spelling normalisation, because they either disappeared from the modern language, or lost their temporal semantics, e.g., *"stübglogge"* from Example 1. Machine learning techniques may be applied to deal with such expressions. For instance, a character-level classifier can be used in order to learn the shape of the word as a sequence of characters. Successful character-based systems for information extraction tasks were described in (Klein et al., 2003) and (Qi et al., 2014).

The presented Gold Standard corpus is available for research purposes. To obtain a copy of the corpus, please contact the author of the paper.

6. References

Angel X. Chang and Christopher Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 180–183, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pavel Logačev, Katrin Goldschmidt, and Ulrike Demske. 2014. Pos-tagging historical corpora: The case of early new high german. In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories*, Tübingen, Germany, December.

Eva Pettersson, Beata Megyesi, and Joakim Nivre. 2013. Normalization of historical text using context-sensitive weighted levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference on Computational Linguistics :*.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 32–41, Gothenburg, Sweden, April. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2013. *Natural language annotation for machine learning*. OReilly Media, Sebastopol, CA.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. Timeml: Robust specification of event and temporal expressions in text. In *in Fifth International Workshop on Computational Semantics (IWCS-5*.

Yanjun Qi, Sujatha G. Das, Ronan Collobert, and Jason Weston. 2014. Deep learning for character-based information extraction. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 668–674.

Katrin Michaela Rettich. 2013. *Automatische Annotation von deutschen und französischen temporalen Ausdrücken im Text+Berg-Korpus Zusammenfassung*. Master thesis, University of Zurich.

Roser Saurí, Jessica Littman, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML annotation guidelines, version 1.2.1.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. 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*, LaTeCH '11, pages 19–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frank Schilder and Andrew McCulloh. 2005. Temporal information extraction from legal documents. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, *Annotating, Extracting and Reasoning about Time and Events*, number 05151 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

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

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Jannik Strötgen and Michael Gertz. 2011. Wikiwarsde: A german corpus of narratives annotated with temporal expressions. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2011)*, pages 129–134, Hamburg, Germany, September.

Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef. 2010. Challenges in building a multilingual alpine heritage corpus. In *LREC*, Valetta, Malta. European Language Resources Association.