I still have Time(s): Extending HeidelTime for German Texts

Andy Lücking, Manuel Stoeckel, Giuseppe Abrami, Alexander Mehler
Goethe University Frankfurt, Text Technology Lab
Robert-Mayer-Str. 10, D-60325 Frankfurt am Main, Germany
{luecking, m.stoeckel, abrami, mehler}@em.uni-frankfurt.de

Abstract

HeidelTime is one of the most widespread and successful tools for detecting temporal expressions in texts. Since HeidelTime's pattern matching system is based on regular expression, it can be extended in a convenient way. We present such an extension for the German resources of HeidelTime: HeidelTime_ext. The extension has been brought about by means of observing false negatives within real world texts and various time banks. The gain in coverage is 2.7 % or 8.5 %, depending on the admitted degree of potential overgeneralization. We describe the development of HeidelTime_ext, its evaluation on text samples from various genres, and share some linguistic observations. HeidelTime_ext can be obtained from `https://github.com/texttechnologylab/heideltime`.

Keywords: HeidelTime, German, TimeBanks, BIOfid

1. Motivation

Biodiversity literature exhibits frequent references to times. Hence, processing biodiversity texts, as is done in the context of specialized information services (Driller et al., 2020), involves recognizing temponyms, that is, temporal expressions which have – possibly aided by context information – a unique interpretation on a time line, like toponyms do in terms of geo-spatial entities (Kuzey et al., 2016a). To this end, HeidelTime (Kuzey et al., 2016b; Strötgen and Gertz, 2013) is used within the BIOfid project (`www.biofid.de`) to detect mentions of time-denoting expressions in mainly German biological texts from the 19th century until today – see (Lücking et al., 2021) for an overview. These texts, however, contain tempponyms that are outside of the extension of HeidelTime's rule system and include spelling variants of mundane tempponyms like *Herbste* 'fall'[1], duration-forming constructions such as *in den letzten beiden Jahren* 'in the past two years'[2], and set-forming constructions such as *in einem zweijährigen Turnus* 'on a biennial basis'[3]. Based on such missing patterns of temporal expressions, among others, we extend the German component of HeidelTime (section 2.). We want to emphasize that HeidelTime is designed in such a way that such extensions can be implemented in a comparatively easy manner. The gain in coverage in assessed on various kinds of corpora (section 3.). A manual inspection of the novel instances found by HeidelTime_ext brought two overgeneralizing rules to the light, however. Depending on how much potential overgeneralization is admitted, the relative coverage improvement of HeidelTime_ext is 2.7 % (excluding one source of overgeneralization) or 8.5 % (including the source of overgeneralization). The HeidelTime extension – called HeidelTime_ext – is available at `https://github.com/`

---

[1]E.g., `https://sammlungen.ub.uni-frankfurt.de/3728500` and `https://sammlungen.ub.uni-frankfurt.de/4497114`

[2]E.g., `https://sammlungen.ub.uni-frankfurt.de/3721623`

[3]E.g., `https://sammlungen.ub.uni-frankfurt.de/3748527`

`texttechnologylab/heideltime`.

2. Procedure

HeidelTime is developed to detect a certain kind of tempponyms, namely tempponyms that are categorized as TIMEX3 expressions (ISO, 2012). TIMEX3 expressions comprise dates (e.g., *10 January, 1999*), times (e.g., *12 o'clock*), durations (e.g., *two weeks*), sets (e.g., *every year*). To extend HeidelTime, we first looked for TIMEX3 patterns which are not yet covered.

2.1. Manual Exploration

In the context of BIOfid, we manually sampled sentences from the current BIOfid corpus and processed the sample with HeidelTime. We then inspected the outcome, focusing on false negatives, that is, TIMEX3 expressions that have not been detected. Generalizing over these false negatives, missing patterns or expressions have been identified. In contexts like natural language processing in biodiversity or the humanities it is important to cover all these instances irrespective of their frequency (we come back to matters of frequency as part of the evaluation in Sec. 3.). In general, false negatives fall in one of four classes:

- Spelling variants. This rather trivial class pertains to typographic variation. In German, the dative case of masculine singular nouns can be marked by the suffix *-e*, as in *dem Herbste* 'the-DAT fall'. This spelling variation, however, is a bit old-fashioned, so that it frequently occurs in the older texts from the BIOfid corpus, but less so in contemporary writings. Another example is punctuation in time expressions. Although using a dot instead of a colon to separate hour and minute – as in *21.30 Uhr* '9.30 pm' – does not comply with international standards (ISO, 2006), it is nonetheless used in texts and for that reason should be detected.

- Lexical extensions. Time spans are sometimes partitioned according to business or financial concerns, such as fiscal years. Corresponding nouns (e.g.,

Geschäftsjahr) are a straightforward extension to HeidelTime's time units. Further lexical extensions are due to temporal adjectives or adverbs. For instance, the modifier *täglich* 'everyday' has the same meaning as the quantified noun phrase *jeden Tag* 'every day', but, in contrast to the noun phrase is not yet covered by HeidelTime's lexicon.

- Compounds. German is well-known for its "tapeworm words" (Twain et al., 2016). However, HeidelTime is not concerned with compounds, and presumably for a good reason: the risk of overgeneration is large. For instance, *Jugendzeit* 'young days'/'adolescence'/'youth', although ending on *-zeit* '-time', is not a TIMEX3 expression but denotes a developmental stage. However, there are a couple of "well behaved" compounds. We count compounds where the modifying noun is a known temponym among them, such as *Winterzeit* 'wintertime' or *Sommermonate* 'months of summer'. Arguably, the head nouns do not contribute much in these cases so that, for instance, a combination of a season term and *-zeit* '-time' can be normalized to the value of the season term in a straightforward way. Compounding also underlies the formation of temporal expressions of a set type. An example is *-basis*: *Wochenbasis* 'on a weekly basis' means every week.

- Rule extensions. Some of HeidelTime's rules are restricted to a certain class of expression. For instance, while quantifying over years is covered, quantifying over seasons is not. For that reason, the duration denoting expression *viele Winter* 'for many winters' is not recognized as such, but has become recognizable by adding corresponding duration rules. A related observation can be made with regard to relative times. There are rules that capture times such as *letzten Freitag* 'last Friday', but the synonymous expression *vorheriger Freitag* 'previous Friday' had to be licensed by an additional rule.

All extensions are marked as such within the source files making up HeidelTime_ext. It should be mentioned that not much emphasis is put on grammatical well-formedness or common usage: seldom or questionable compounds are recognized as well as phrases that lack morpho-syntactic agreement. After all, HeidelTime, as HeidelTime_ext, is an annotator, not a grammar.

2.2. Populating Negative Rules

In writing regular expressions, care has to be taken to not to overgenerate. To this end, HeidelTime employs so-called *negative rules*, that is, rules which, when apply, remove their matched expressions from the output. To give an example: while season names are welcome temponyms, they can also figure as family names such as in *Herr Sommer* 'Mister Summer'. Such instances can be excluded by a negative rule that says that if a season term follows *Herr* 'Mister' or *Frau* 'Miss', then remove it. We added such a negative rule. However, one cannot stop here: the BIOfid example *Assistent Sommer* 'assistant Summer',

(file `https://sammlungen.ub.uni-frankfurt.de/3673151`) still circumvents our new negative rule. Obviously, in addition to addressing particles, also profession terms can mark a season term as a proper name. Therefore, we collected a list of profession terms from the German agency for employment and added them to HeidelTime_ext's pattern files.[4] Addressing particles and profession terms still fall short of capturing *Ehepaar Sommer* 'the married couple Summer', however. This example is, of course, covered by HeidelTime_ext, but it illustrates that time recognition seems to be an open-ended task. For this reason, we followed a more dynamic approach and used BERT, a transformer-based language model trained for contextual embeddings of words (Devlin et al., 2019). BERT can be used like a cloze test: suggestions for a masked item can be obtained from left and right context information. In this manner, we used the sentence containing *Assistent Sommer* as input, masked the noun *Assistent*, and collected the 30,000 words (in fact, BERT also suggests non-word character sequences) which according to BERT fit best into the placeholder position. We then removed all suggestions that are shorter than 4 characters and excluded fragmentary items (starting with "##"). Finally, we selected the first 5,000 suggestions (using [much] more leads to regular expressions which are too long to handle for the system). The BERT list provides an immediate benefit for the negative rule: since it includes typical given names, full names ending on a season or weekday term are rightly excluded from HeidelTime_ext's temponym recognition.

2.3. Harvesting Time Banks

To obtain indications of further extensions of HeidelTime, we looked at instances of expressions that have been marked as TIMEX3 expressions in several time banks. We extracted the content of TIMEX3 tags from the French TimeBank (Bittar et al., 2011) (Bittar, 2011), the Basque TimeBank (Altuna et al., 2020) (Altuna, 2019), and the MEANTIME newsreader corpus (Minard et al., 2016) (The Newsreader Project, 2015) (Dutch, English, Italian, Spanish). We then used `www.deepl.com` to translate the time expressions from different languages into German ones. We fed the list into HeidelTime and inspected the outcome, most notably lines that lacked a HeidelTime tag.[5] This procedure resulted in 83 sample pattern which underlie the coverage extension of HeidelTime_ext. The examples have been chosen manually. False negatives which do not make up clear temponyms have been ignored. This includes

---

[4] To be more precise, we applied a string manipulation on the profession terms first: we removed subdomain classifications (for instance, distinguishing farmers for different agricultural sectors), and extended each entry into a separate masculine and feminine form. To avoid running into errors due to too long regular expressions, we restricted the list to single-word profession terms.

[5] Since the temporal expression-based input lacks sentential context, ambiguity due to conflicting pattern scopes arises. For instance, if a day term precedes a date, then both are taken to be a single time expression. If, however, a second day term follows, a conflict arises whether the first or the second day mention constitutes a time expression together with the date. Such ambiguities do not do harm to our approach, since the output is manually inspected anyway.

event-denotating expressions (e.g., *l'heure de l'Europe et de la mondialisation* 'in the age of Europe and globalization' or *temps de guerre contre le terrorisme* 'in times of the war on terrorism'), "vacuous" temporal expressions (such as *any time*), and temporal modifiers which typically are used to modify not-temporal nouns (e.g., *recent*).

3. Evaluation

The gain in coverage of HeidelTime and HeidelTime$_{ext}$ (Sec. 3.3.) is assessed and compared on various resources (Sec. 3.1.), which have been pre-processed as described in Sec. 3.2..

3.1. Evaluation Corpus

Texts have been sampled from five different sources to balance potential effects of text type to the frequency of temporal expression use:

- 10 randomly collected protocols of the German Bundestag (https://www.bundestag.de/services/opendata).

- 10 books from the *German Text Archive* (DTA, https://www.deutschestextarchiv.de, namely Dickens, *Weihnachtsabend* (1844), Fontane, *Effi Briest* (1896), Goethe, *Faust 1* (1808), von Humboldt, *Kosmos*, vol. 1 (1845), Kafka, *Die Verwandlung* (1915), Lessing, *Nathan der Weise* (1779), Marx, *Das Kapital*, vol. 1 (1867), Nietzsche, *Homer und die klassische Philologie* (1869).

- 10 randomly selected tests from the *Zoologisch-Botanische Datenbank* (Zobobat, https://www.zobodat.at). The texts have been converted to plain text files from OCR PDFs and hence contain some token errors.

- 766 articles from the *Süddeutsche Zeitung* (SZ), collected within three TEI files with the following time stamps: 15. 06. 1996, 04. 09. 2002 and 07. 12. 2013.

- 10,000 randomly selected sentences from Wikipedia (WP) from the *Leipzig Wortschatz*[6], dump from the year 2012 (Goldhahn et al., 2012).

The size of the samples is summarized in Tab. 1.

Sample	# sentences	# tokens
Bundestag	188,768	3,682,370
DTA	27,687	810,582
SZ	18,938	359,706
Zobodat	6,231	92,003
WP	10,000	176,775
sum	251,624	5,121,436

Table 1: Number of tokens (incl. punctuation) and sentences within the evaluation samples.

3.2. Pipeline

Since HeidelTime requires single white spaces and, at least for some rules, part-of-speech information, all texts described in the previous section have been pre-processed using a TextImager pipeline (Hemati et al., 2016) as follows:

1. Normalization: All white spaces have been normalized to single spaces.

2. Segmentation: Sentences have been segmented using the OpenNLP Max Entropy Model.[7]

3. Tokenization: Word forms have been tokenized by using the Stanford CoreNLP (Manning et al., 2014) *via* DKPro (Eckart de Castilho and Gurevych, 2014).

4. Part-of-speech Tagging: Parts-of-speech (POS) have been assigned by using the POS tagger from Mate-Tools (Bohnet and Nivre, 2012) *via* DKPro (Eckart de Castilho and Gurevych, 2014)

HeidelTime and HeidelTime$_{ext}$ are run on the pre-processed texts and compared by means of two views in a UIMA CAS.

3.3. Results

HeidelTime$_{ext}$ found 4,458 more TIMEX3 expressions than the original HeidelTime, a gain of 8.5 %, as summarized in Tab. 2. However, as discussed below in Sec. 4., the bare gain in coverage has to put into perspective.

Sample	HeidelTime	HeidelTime$_{ext}$
Bundestag	36,068	37,507
DTA	5,827	7,568
SZ	4,399	5,305
Zobodat	2,110	2,235
WP	3,966	4,213
sum	52,370	56,828

Table 2: Number of TIMEX3 expressions found by original HeidelTime and HeidelTime$_{ext}$ (see also Sec. 4.).

The coverage is detailed in Tab. 3. Column "novel" lists the number of temporal expressions newly found by HeidelTime$_{ext}$, whereas the column "missing" counts the expressions only found by the original HeidelTime. Since there are 298 missing TIMEX3 in total (assuming that these are true positives), this means that the newly added rules in HeidelTime$_{ext}$ interfere with the application of some of the original rules. The extended rule system of HeidelTime$_{ext}$ also generally covers larger token spans, as expressed in the column "extended". Smaller token spans – column "reduced" – can be ignored due to their little frequency of occurrence.

The following examples provide an impression of the kinds of temporal expressions which HeidelTime$_{ext}$ is designed for: *Vorjahr* 'preceding year', *übermorgen* 'the day after tomorrow', *Wintermonate* 'winter months', *eine halbe Stunde*

Sample	novel	extended	reduced	missing
Bundestag	267	70	2	97
DTA	258	37	—	59
SZ	436	75	1	103
Zobodat	86	5	—	8
WP	133	28	—	31
sum	1,180	215	3	298

Table 3: Coverage details. Novel: newly found temporal expressions; extended: "longer" matches; reduced: "shorter" matches; missing: failed detection by HeidelTime$_{ext}$.

'half an hour'. Furthermore, durations as property denotations like *dreitägig* 'three-day' as in *dreitägige Exkursion* 'three-day excursion' are also taken to be TIMEX3 expressions. However, a manual inspection of the outcome of HeidelTime$_{ext}$ revealed two kinds of expressions which need to be discussed.

4. Discussion

To assess the outcome of HeidelTime$_{ext}$ not only in quantitative terms, we manually inspected 1,198 TIMEX3 expressions only found by HeidelTime$_{ext}$.[8] Each instance is labeled as being a true or a false positive – see Tab. 4 for an overview. The relatively high number of false positives, however, is mainly due to two kinds of overgeneralizations. The first overgeneralization is bound up with two TIME rules which wrongly apply not only to times of day expressions but to any measure terms formed with digits and a separating punctuation symbol. From 168 examples of this type, which have been part of manual inspection, 139 are regarded false, and 29 true positives. This overgeneralization is attenuated in the release of HeidelTime$_{ext}$ as of 19th April 2022 and following ones.

Sample	true	false
Bundestag	185	35
DTA	217	95
SZ	210	73
Zobodat	64	64
WP	125	30
sum	801	297

Table 4: Instances of true and false positives in newly detected TIMEX3 expressions.

The second kind of overgeneralization is due to a significant proportion of newly detected occurrences of German *nun*, which has not been part of the original HeidelTime. HeidelTime$_{ext}$ found 2,530 instances in total.[9] This particle word, however, has two broad uses which translate into English as 'now' respectively "well" (that is, the adverb typically used sentence initial). Obviously, only the first

of these uses is a temporal one, namely PRESENT_REF.[10] During manual inspection, 455 instances of *nun* (irrespective of capitalization) have been checked: 333 of them correspond to a temporal, 122 to a discourse use. Since there does not seem to be typical contexts which distinguish between temporal and discourse-functional *nun*, overgeneration cannot simply be prevented by a negative rule. This example therefore exemplifies a limit of temponym recognition based on regular expressions. The question therefore arises whether neural network-based approaches such as CNNs (Lin et al., 2017), which have been trained on date expressions, fare better with regard to temporal particles.[11] For the time being, a user of HeidelTime$_{ext}$ can choose how to proceed by commenting out the rule in question (rule `date_r8a-explicit`).

Removing the counts for *nun* and for the overgeneralizing TIME rules from the figures given in Tab. 2, we get the more "cautious" overview in Tab. 5. There is now a gain of 1,416 TIMEX3 expressions, or 2.7 %.

Sample	HeidelTime	HeidelTime$_{ext}$
Bundestag	36,068	36,386
DTA	5,827	6,429
SZ	4,399	4,748
Zobodat	2,110	2,136
WP	3,966	4,087
sum	52,370	53,786

Table 5: Number of TIMEX3 expressions found by original HeidelTime and HeidelTime$_{ext}$ with problematic cases removed.

Note finally that we still found examples that circumvented our highly extended negative rules, cf. Sec. 2.2.. For instance, the given name *Carl*, as opposed to the form variant *Karl*, is not part of any black list, so that the proper name *Carl Winter* still triggers a season of the year rule.

5. Conclusion

Based on manually collected false negatives from various time banks and biological texts, we developed HeidelTime$_{ext}$, a German extension of HeidelTime. We constructed an evaluation corpus to quantify the gain in coverage of the extension. A manual inspection of novel instances found by HeidelTime$_{ext}$ identified two kinds of rules which tend to overgeneralize to non-temporal instances, however. Depending on how much potential overgeneralization is admitted, the relative coverage improvement of HeidelTime$_{ext}$ is between 2.7 % and 8.5 %. At this point, one could object that our approach mainly discovers rarely occurring expressions of time. But especially these cases are interesting for disciplines of the humanities or

---

[8]Bundestag: 220, Zobodat: 228, DTA: 312, WP: 155, SZ: 283.

[9]Bundestag: 1,010, DTA: 1,119, SZ: 289, Zobodat: 28, WP: 84.

[10]Actually, things are more complicated than that: as is well known, even indexical expressions are evaluated with reference to three temporal indices, namely event time, reference time, and utterance time (Reichenbach, 1947). For a semantics of temporal reference see (Kamp, 1979).

[11]Neural networks can be used for temponym *detection*, however, but not so easily for time *normalization*.

biodiversity (as outlined in Secs. 1. and 2.1.), which also deal with what is rare rather than with what is frequent. Future work might deal with a compositional approach to written numbers such as *einhundertdreiundfünfzig* 'one hundred and fifty-three'. A difficulty here of course is to generate a corresponding norm value. Another issue is the improvement of negative rules to prevent overgeneralization. We used a list-based approach, making use of governmental material and material generated by a BERT model. HeidelTime$_{ext}$ is available from `https://github.com/texttechnologylab/heideltime`.

6. Acknowledgements

7. Bibliographical References

Altuna, B., Aranzabe, M. J., and Díaz de Ilarraza, A. (2020). EusTimeML: A mark-up language for temporal information in Basque. *Research in Corpus Linguistics*, 8(1):86–104.

Bittar, A., Amsili, P., Denis, P., and Danlos, L. (2011). French TimeBank: An ISO-TimeML annotated reference corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 130–134. Association for Computational Linguistics.

Bohnet, B. and Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.

Driller, C., Koch, M., Abrami, G., Hemati, W., Lücking, A., Mehler, A., Pachzelt, A., and Kasperek, G. (2020). Fast and easy access to Central European biodiversity data with BIOfid. *Biodiversity Information Science and Standards*, Oct.

Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 759–765. European Language Resources Association (ELRA).

Hemati, W., Uslu, T., and Mehler, A. (2016). TextImager: a distributed UIMA-based system for NLP. In *Proceedings of the COLING 2016 System Demonstrations*. Federated Conference on Computer Science and Information Systems.

ISO. (2006). Data elements and interchange formats – information interchange – representation of dates and times. Standard, International Organization for Standardization.

ISO. (2012). Language resource management – semantic annotation framework (SemAF) – part 1: Time and events (SemAF-Time, ISO-TimeML). Standard ISO/IEC TR 24617-1:2012, International Organization for Standardization, Geneva, CH.

Kamp, H. (1979). Events, instants and temporal reference. In Rainer Bäuerle, et al., editors, *Semantics from Different Points of View*, number 6 in Springer Series in Language and Communication, pages 376–417. Springer, Berlin and Heidelberg and New York.

Kuzey, E., Setty, V., Strötgen, J., and Weikum, G. (2016a). As time goes by: Comprehensive tagging of textual phrases with temporal scopes. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 915–925.

Kuzey, E., Strötgen, J., Setty, V., and Weikum, G. (2016b). Temponym tagging: Temporal scopes for textual phrases. In Jacqueline Bourdeau, et al., editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 841–842. ACM.

Lin, C., Miller, T., Dligach, D., Bethard, S., and Savova, G. (2017). Representations of time expressions for temporal relation extraction with convolutional neural networks. In *BioNLP 2017*, pages 322–327, Vancouver, Canada,, August. Association for Computational Linguistics.

Lücking, A., Driller, C., Stoeckel, M., Abrami, G., Pachzelt, A., and Mehler, A. (2021). Multiple annotation for biodiversity: developing an annotation framework among biology, linguistics and text technology. *Language Resources & Evaluation*. Online first.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., and van Son, C. (2016). MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC'16, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).

Reichenbach, H. (1947). *Elements of Symbolic Logic*. The Macmillan Company, New York.

Strötgen, J. and Gertz, M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources & Evaluation*, 47:269–298.

Twain, M., Kersten, H., and Embassy of the USA. (2016). *The awful German language*. Insel Verlag.

8. Language Resource References

Altuna, Begoña. (2019). *EusTimeML*. B. Altuna, distributed via `http://ixa2.si.ehu.eus/eusheideltime/index.html`, 2.0.

André Bittar. (2011). *French TimeBank*. ddenz, distributed via `https://github.com/ddenz/FR-TimeBank`, 1.1.

The Newsreader Project. (2015). *MEANTIME*. Newsreader project, distributed via `http://www.newsreader-project.eu`, 1.0.