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

Andy Lücking, Manuel Stoeckel, Giuseppe Abrami, Alexander Mehler

Goethe University Frankfurt | Text Technology Lab


21 – 23 June 2022
Marseille, France

## 1. Motivation

- We found temponyms that are outside of the extension of HeidelTime's rule systems:
  - spelling variants of mundane temponyms like *Herbste* 'fall'
  - duration-forming constructions such as *in den letzten beiden Jahren* 'in the past two years'
  - set-forming constructions such as *in einem zweijährigen Turnus* 'on a biennial basis'
- The examples have been found within the *Fachinformationsdienst Biodiversität* (BIOfid; see also https://sammlungen.ub.uni-frankfurt.de/biodiv), which uses HeidelTime

temponym: temporal expression which has – possibly aided by context information – a unique interpretation on a time line (Kuzey et al. 2016).

## 2. Procedure

- Extending HeidelTime to HeidelTime$_{ext}$ based on missing patterns of temporal expressions.

### 2.1 Manual Exploration

- Generalizing over false negatives found in BIOfid texts, four missing patterns or expressions have been identified:
  - Spelling variants: seldom and old-fashioned dative suffix *-e*, as in *dem Herbste* 'the-DAT fall', or punctuation in time expressions (e.g., 9.30 pm vs. 9:30 pm).
  - Lexical extensions: the modifier *täglich* 'everyday' which has the same meaning as the quantified noun phrase *jeden Tag* 'every day'
  - Compounds: for instance, compounds where the modifying noun is a known temponym, such as *Winterzeit* 'wintertime' or *Sommermonate* 'months of summer'.
  - Rule extensions: relative times such as *letzten Freitag* 'last Friday' are captured, but the synonymous expression *vorheriger Freitag* 'previous Friday' had to be added

### 2.2 Negative Rules for Proper Names

- Negative rule: a rule which removes its matched expressions from the output.
- Example: *Herr Sommer* 'Mister Summer' can be excluded by a negative rule that says that if a season term follows *Herr* 'Mister' or *Frau* 'Miss', then remove it.
- But also a profession term can mark a season term as a proper name: *Assistent Sommer* 'assistant Summer'
- Therefore, we collected a list of profession terms from the German agency for employment and added them to HeidelTime$_{ext}$'s pattern files (in a tidied up form).
- Profession terms still fall short of capturing *Ehepaar Sommer* 'the married couple Summer', however.
- 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). We used the sentence containing *Assistent Sommer* as input, masked the noun *Assistent*, and collected the 30,000 words which according to BERT fit best into the placeholder position.

### 2.3 Harvesting TimeBanks

- We extracted the content of TIMEX3 tags from the French TimeBank (Bittar et al. 2011), the Basque TimeBank (Altuna, Aranzabe, and Ilarraza 2020), and the MEANTIME newsreader corpus (Minard et al. 2016) (Dutch, English, Italian, Spanish).
- We then used www.deepl.com to translate them into German.
- We fed the list into HeidelTime and inspected the outcome
- We chose 83 sample patterns to extend HeidelTime$_{ext}$.

## 3. Evaluation Corpus

- 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`).
- 766 articles from the *Süddeutsche Zeitung* (SZ)
- 10,000 randomly selected sentences from Wikipedia (WP) from the *Leipzig Wortschatz* (Goldhahn, Eckart, and Quasthoff 2012).

## 4. Pipeline

All texts have been pre-processed using TextImager (Hemati, Uslu, and Mehler 2016):

1. Normalization: All white spaces have been normalized to single spaces.
2. Segmentation: Sentences have been segmented using the OpenNLP Max Entropy Model.
3. Tokenization: Word forms have been tokenized by using the Stanford CoreNLP (Manning et al. 2014) *via* DKPro (Castilho and Gurevych 2014).
4. Part-of-speech Tagging: Parts-of-speech (POS) have been assigned by using the POS tagger from MateTools (Bohnet and Nivre 2012) *via* DKPro (Castilho and Gurevych 2014)

## 5. Results and Discussion

- HeidelTime$_{ext}$ found 4,458 more TIMEX3 expressions than the original HeidelTime, a gain of 8.5 % – see left tabular.
- Right tabular: Instances of true and false positives in newly detected TIMEX3 expressions:

| 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 |

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

- Overgeneration mainly due to *nun* 'now', which cannot be distinguished from *nun* 'well'.
- Removing the counts for *nun* and for the overgeneralizing TIME rules we get a gain of 1,416 TIMEX3 expressions, or 2.7 %.
- In biological and digital humanities contexts, what is rare is as interesting as what is frequent...

## Get HeidelTime$_{ext}$

`https://github.com/texttechnologylab/heideltime`

## References

Altuna, Begoña, María Jesús Aranzabe, and Arantza Díaz de Ilarraza (2020). "EusTimeML: A mark-up language for temporal information in Basque." In: *Research in Corpus Linguistics* 8.1, pp. 86–104.

Bittar, André, Pascal Amsili, Pascal Denis, and Laurence Danlos (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*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 130–134.

Bohnet, Bernd and Joakim Nivre (July 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*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1455–1465.

Castilho, Richard Eckart de and Iryna Gurevych (Aug. 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*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 1–11.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL].

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff (2012). "Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. European Language Resources Association (ELRA), pp. 759–765.

Hemati, Wahed, Tolga Uslu, and Alexander Mehler (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. Osaka, Japan.

Kuzey, Erdal, Vinay Setty, Jannik Strötgen, and Gerhard Weikum (2016). "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. Montréal, Québec, Canada, pp. 915–925.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit." In: *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60.

Minard, Anne-Lyse, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son (2016). "MEANTIME, the NewsReader Multilingual Event and Time Corpus." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. LREC'16. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4417–4422.