When Does (German) Literature Take Place? On the Analysis of Temporal Expressions in Large Corpora

Frank Fischer[1] and Jannik Strötgen[2]

[1]Göttingen Centre for Digital Humanities, University of Göttingen, Germany
[2]Institute for Computer Science, Heidelberg University, Germany

1 Introduction

Exact date specifications are a feature of many types of prose. However, literary texts seem to prefer approximate dates opening room for interpretation. For instance, all 19 mentions of month names in Theodor Storm's novella *Der Schimmelreiter* (*The Rider on the White Horse*) are of an approximate nature ("on an October afternoon", "at the end of March", etc.). Immediate exceptions to this rule are epistolary, diary, adventure and historical novels: Goethe's *Sorrows of Young Werther* or Jules Verne's *Around the World in Eighty Days* are, due to their genre, filled with exact dates. For other prose genres we can try to formulate as a hypothesis: If an exact date occurs in literary texts, it is a narrative statement that calls for further analysis. With this as a premise, we can try to pose and answer questions that can be regarded typical research questions in the field of literary studies:

- Is the avoidance of exact dates really a continuous feature of certain literary genres? How do exact dates relate to approximate dates, frequence-wise?
- Can a frequency analysis of date specifications be used for genre studies?
- Which meanings can dates have other than providing the temporal setting of a story? ('semiotisation' of dates)
- Are there accumulations of certain dates in large corpora? If so, why?
- What is the role of fictitious dates? (cf. Erich Kästner's novel *The 35th of May* or Shakespeare's 80th of April in *The Winter's Tale*, Autolycus' ballad in Act 4: "Here's another ballad of a fish, that appeared upon / the coast on Wednesday the four-score of April")

In our presentation, we interpret date specifications as an isolatable feature of literary corpora as proposed by Matthew Jockers: "Indeed, the very object of analysis shifts from looking at the individual occurrences of a feature in context to looking at the trends and patterns of that feature aggregated over an entire corpus" (Jockers 2013, p. 24). Looking at this feature (the explicit mentioning of a date) as a single analysable unit will add to the methods of literary macroanalysis.

2 Workflow

Our workflow consisted of four steps that were performed in parallel: 1. compilation of a suitable (German-language) corpus. 2. collection of data using the temporal tagger HeidelTime (Strötgen & Gertz 2012) for the automatic extraction of temporal

expressions according to the guidelines of the temporal markup language TimeML (Pustejovsky et al. 2003). 3. data analysis (from heat maps to individual cases). 4. development of an Android app for exploring the “literary year”.

Out of the two biggest corpora with German literary texts, the TextGrid Repository[1] and Gutenberg-DE[2], we chose the latter. We prepared the corpus so it would only contain fiction and ended up with 2735 works by 549 authors, the majority of which had been published between 1840 and 1930. The resulting 900 MB of text were fed into HeidelTime to extract date specifications. Just using the explicit (and therefore very correct) expressions, we created a calendar heatmap (where ‘1’ means 0–9 occurrences, ‘2’ means 10–19 occurrences, etc., and ‘+’ means 90 or more occurrences; days with more than 50 occurrences are highlighted). Some expected and unexpected accumulations turned up:

```
JAN: +333222323131323223222222222131
FEB: 433222221332123323322123222231
MAR: 7333432223324342243363232322252    (21st)
APR: +332332234322232233323222223323
MAY: +3544332353643532324243232232 44   (12th)
JUN: 7332333233333244323433243 33233
JUL: 9444332333243652333432224223223    (14th)
AUG: 836442232724446334453333232322 2   (3rd, 10th, 15th)
SEP: 8544332333223423323322122232 3
OCT: +353322224223552253432222222133
NOV: 944233333723225213232222222224     (10th)
DEC: 552234121322323213223339 2122224   (24th)
```

As we can see, first days of a month and fixed holidays (New Year, Christmas) have a much higher frequency. But some other days also stick out, an example being the ‘10th of August’. A look into our results files showed that 74% of its occurrences provide a temporal setting for the fictional plot. The above-average frequency of the 10th of August, though, has to do with a historical event, the storming of the Tuileries Palace on August 10, 1792. About 21% of the occurrences are references to this historical date (cf. Georg Büchner’s play *Danton’s Death*: “FIRST CITIZEN: Danton was with us on the 10th August, Danton was with us in September.”). Therefore, it is necessary to distinguish between historical dates that left their traces in literary texts, and explicit dates that provide a temporal setting of a fictional plot. The collection and analysis of such date accumulations will be systematically expanded, in regard to specific authors, languages, nations.

[1] http://www.textgridrep.de/
[2] http://projekt.gutenberg.de/

3 Significance for Literary Studies

Along the lines of Hans Ulrich Gumbrecht's study on the year *1926*, it would be useful and feasible to assemble the literary history of individual days. Every date has its own literary history as is indicated by the example of Paul Celan and the '20th of January'.

In Celan's prose poem *Conversation in the Mountains* (1960), he alludes to Georg Büchner's short story *Lenz* which also describes a passage through the mountains. Büchner's text starts with the sentence: "The 20th of January, Lenz walked through the mountains." In *Der Meridian*, Celan's acceptance speech for the Georg Büchner Prize (Germany's most prestigious literary accolade), he stresses that Lenz's hike through the mountains takes place on a 20th of January and extends the temporal frame by referring to another 20th of January, the one of 1942, when the Wannsee Conference took place in Berlin. Celan concludes: "Perhaps one may say that every poem has its '20th of January' inscribed?" (cf. Sieber 2007, pp. 114).

The automatic extraction of date expressions from large corpora makes such simultaneities visible and enables their systematic exploration.

4 Development of an Android App to Facilitate the Exploration of Date Expressions in World Literature

To get an idea of the seasonal cycle of literature, we developed an Android application that works like a calendar and, for each day of the year, presents passages from canonised texts of world literature that take place on that very day. Screenshots are shown in Figure 1.

It is well known that James Joyce's *Ulysses* takes place on June 16, 1904. But there is just one inconspicuous mentioning of the day in the novel (the secretary Ms Dunne types it in on the keyboard), it is made visible in the app. Other examples for such passages are June 12 in Günter Grass's novel *The Tin Drum* (birth of Oskar Matzerath's declared son Kurt), February 29 in Thomas Mann's *Magic Mountain* (where the date is of importance as a special variant of the Walpurgisnacht, see Figure 1) and July 27 in Stefan Zweig's *Chess Story* (on that day, imprisoned protagonist Dr. B. gets hold of the chess book).

Our app thus represents a database of fictional dates which will be available for further research. To be able to map the entire "literary year", though, we will also have to take approximate temporal expressions into account which we will be attempting in the next section.

5 The Seasonal Cycle of Literature

We already mentioned the very specific ratio between exact and approximate dates. In the search for anomalies in the works of individual 19th-century authors we came across Theodor Fontane and Theodor Storm. A collection of just the month

specifications in the fictional works of both authors yielded the results shown in Table 1.

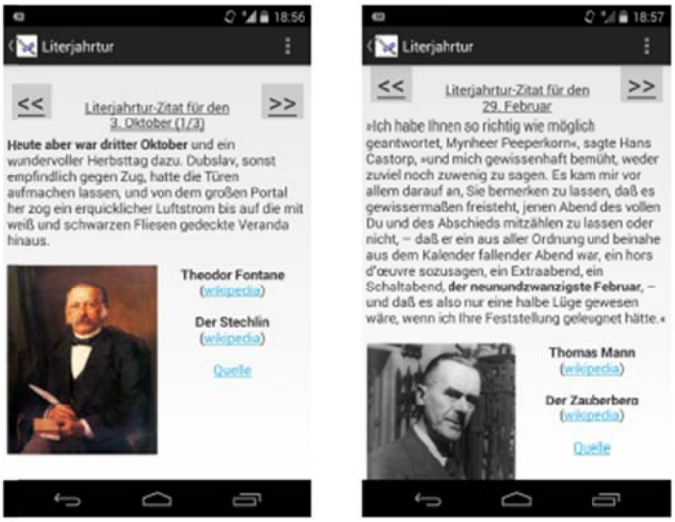

Figure 1: Screenshots of our Android app *Literjahrtur*.

	Fontane	Storm
JAN	30	5
FEB	13	3
MAR	18	7
APR	13	7
MAY	28	11
JUN	17	9
JUL	16	6
AUG	17	10
SEP	36	10
OCT	41	11
NOV	27	13
DEC	25	1

Table 1: Fontane vs. Storm.

In conformity with the popularity of the 1st of May, the whole month is strongly represented in the narratives of both authors. However, the summer months (of the Northern Hemisphere) are not. Fontane's narratives seem to especially take place in-between September and January, Storm's works in-between August and November. Given that every month name has a specific tonal-associative character and creates a stylistic effect, both authors seem to favour autumnal/wintry settings and moods.

6 Conclusions

In this abstract, we presented a method to find date accumulations in large literary corpora. We described the relation between approximate and exact dates and introduced a growing database of exact date specifications in world literature. Furthermore, we approached the seasonal cycle of literature and certain authors to try to answer the question "When does (German) literature take place?" in a macro-analytic fashion. The results obtained to date already show the potential of the ongoing research.

References

Jockers, M. (2013). Macroanalysis. Digital Methods and Literary History. Chicago: University of Illinois Press.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G. and Radev, D. R. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. In: New Directions in Question Answering, pp. 28–34.

Sieber, M. (2007). Paul Celans “Gespräch im Gebirg”. Erinnerung an eine versäumte Begegnung. Tübingen: Niemeyer. http://books.google.de/books?id=KbfF2oIHrjwC&pg=PA114 (accessed 1 March 2015).

Strötgen, J. and Gertz, M. (2012). Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards. In: Proceedings of the 8th International Conference on Language. Resources and Evaluation (LREC 2012), pp. 3746–3753.

In this paper, we investigate the role of temporal signals in temporal relation extraction and provide a quantitative analysis of these expressions in the TimeBank annotated corpus. 1. Introduction. The task of automatically determining the temporal re-lations that hold between events described in a text is a research challenge that has increasingly occupied re-searchers in computational language processing (Set-zer and Gaizauskas, 2000; Pustejovsky et al., 2004; Verhagen et al., 2009; Verhagen et al., 2010). The mechanisms used to convey temporal relational infor-mation in text are complex an The data are taken from Latvian, Russian and German translations of Oscar Wilde and Lewis Carroll. Different techniques are suggested to render a similar effect, such as the use of an equivalent idiom, a loan translation, an extension, an analogue transformation, substitution, compensation, loss of wordplay, and metalingual comment.Â the message and the addressee and on the way the latter is taken by surprise and plunged into something entirely different from what s/he has been prepared for" (Delabatista, 1996:138).Â In this case, the Arabic script, which does not usually indicate short vowels, allows for this kind of pun (orthographic, as mentioned in the literature).