Published in: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/stevenson, Mark/Sick, Theresa (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 2017. Pp. 1-6

Intra-connecting an exemplary literary corpus with semantic web technologies for exploratory literary studies

Andreas Dittrich
Academiae Corpora (Austrian Academy of Sciences) / Sonnenfelsgasse 19/8, 1010 Vienna
andreas.dittrich@oeaw.ac.at

Abstract

Many (modernist) works of literature can be understood by their associativeness, be it constructed or "free". This network-like character of (modernist) literature has often been addressed by terms like "free association", connotation", "context" or "intertext". This paper proposes an experimental and exemplary approach to intraconnect a literary corpus of the Austrian writer Ilse Aichinger with semantic web-technologies to enable interactive explorations of word-associations.

1 Introduction

> "Nearly all poetry is strongly associative." (Cuddon, 2013, p. 58)

Large corpora are rich corpora. Following the etymological routes of the word, Latin *largus* does not mean "thick" and "coarse", like the root of the word "great", but "plentiful" and "abundant". The difference between large and small corpora thus is not the simple measure of quantitative size, but the question of how to deal with it: a methodological question.

For John Sinclair, for whom "the difference [between small and large corpora] must be methodological" (Sinclair, 2001, p. xi), "[t]he main virtue of being large in a corpus is that the underlying regularities have a better chance of showing through the superficial variations" (Sinclair, 2004, 189). In the field of literary studies this "underlying regularities" can be various: a theme, plot, motif, sujet and fabula, device, meaning, rhetoric, trope, style, metric, sound or others. But all these refer to a specific text, which can be gathered as a corpus — and, as a digital corpus, analysed with computational methods. A traditional approach of analysing texts, called "close reading", has been extended by a method roughly labelled as "distant reading", which tries to analyse not just one text, but a plenty. If one doesn't understand these terms as opposites, but as different moments of the same process, one can get to read texts close via distant readings and vice versa (Jänicke et al., 2015; Scrivner and Davis, 2017; Jockers, 2013), more or less as Hans-Georg Gadamer describes the structure of understanding as a "circle of whole and part" (Gadamer, 2004, p. 302–5) (although "whole" probably is a hole).

This constant moving between macro- and micro-structure, requires an interactive workframe without delay, which, depending on the size of the corpus, can be difficult to obtain and the idea of lessen the corpus may occur. One of the apparently most natural processes before or after Natural Language Processing (*NLP*) is the exclusion of stop-words. This crucial intervention alters the corpus drastically and deletes merely seemingly 'meaningless' words like the copula "and", which could be a decisive stylistic factor for an author. Such filtering methods, which are important for making corpora suitable for analysis, reduce the richness and thereby the largeness of a corpus. Usual literary corpora may not reach the quantitative size of comparable corpora from Linguistics in their quantitative scale, but may tend there, when they focus on connections between words.

In the following, I want to discuss a project that deals with texts of a specific author (Ilse Aichinger), whose corpus, which we finished to build in TEI-XML[1] (Text Encoding Initiative, Extensible Markup Language), is small in quantitative size (about

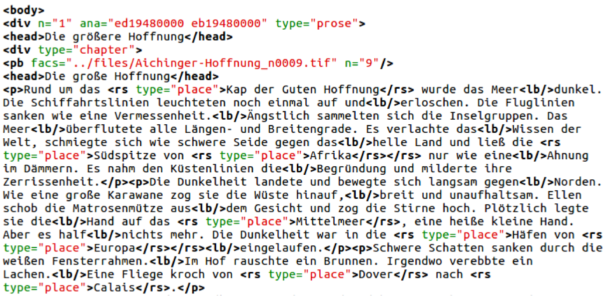
[1] We is a group of students under supervision of Christine Ivanovic from the Institute of Comparative Literature at the University of Vienna and Hanno Biber from Academiae Corpora at the Austrian Academy of Sciences: Marlene Csillag, Katharina Godler, Mathias Müller, Katrin Rohrbacher, Gilbert Waltl and myself.

400.000 tokens), but rich regarding its literary interconnectivity (Fässler, 2011; Thums, 2013; Pelz, 2009; Markus, 2015). After discussing the work of Ilse Aichinger and which problems occurred to us in the process of annotating place-names, I want to propose an interactive visualisation-method, which is based on technologies of the semantic web. For this purpose the XML-files had to be converted to a RDF-format (`Resource Description Framework`). Finally, I present an exemplary, very short study of three words from the corpus in an open-source visualization-framework, named "RelFinder" (Heim et al., 2010). This not only offers 'new' questions for the field of literary studies, but enables us to see other connections between texts, discovering mediating terms and second-order mediations.

2 Places in the corpus :`aichinger`

Places play an eminent role in the writing of Ilse Aichinger (1921–2016). In order to protect her mother, whom the Nazi-regime labelled as "half-Jewish", she did not emigrate from National Socialist Vienna, where she survived second world war. Places trigger a process of remembrance, and thus "the vanished" acquire a literary presence in their absence (Fässler, 2011, p. 26). The places 'touch' Aichinger in her present being (Thums, 2013, p. 193): "The places, which we looked at, look at us" (Aichinger, 2001, my translation, AD), as she writes in a short text. But place-names are not simply uttered or just staging the scene, they also "carry the plot", as she once noted herself (Aichinger, 1991b, p. 179, my translation, AD).

The difficulties in the annotating-process have been diverse and can be summed up in the question: How to define a literary place? This question arose probably because of the very different 'styles' Aichinger exhibits in her entire oeuvre, which spans over 60, transformative years, from her first published text 1945 to her last one 2005. The annotating group faced texts, where very different place-types turned up: fictitious place-names, moving places, acting places, existing place-names, which do not refer to their real place-reference, but also place-names that can be located on a traditional map. The group agreed, that, at least as a first step, only place-names should be annotated, which can be located on a map. Additionally to a light TEI-encoding (with page-breaks, line-breaks, divisions and headings with corresponding publication dates and genre, paragraphs, stage directions, speaker and speeches, line-groups and lines), place-names were manually annotated by using the "referencing string"-tag (`rs`) with the attribute (`type`) "place" (see Fig. 1).

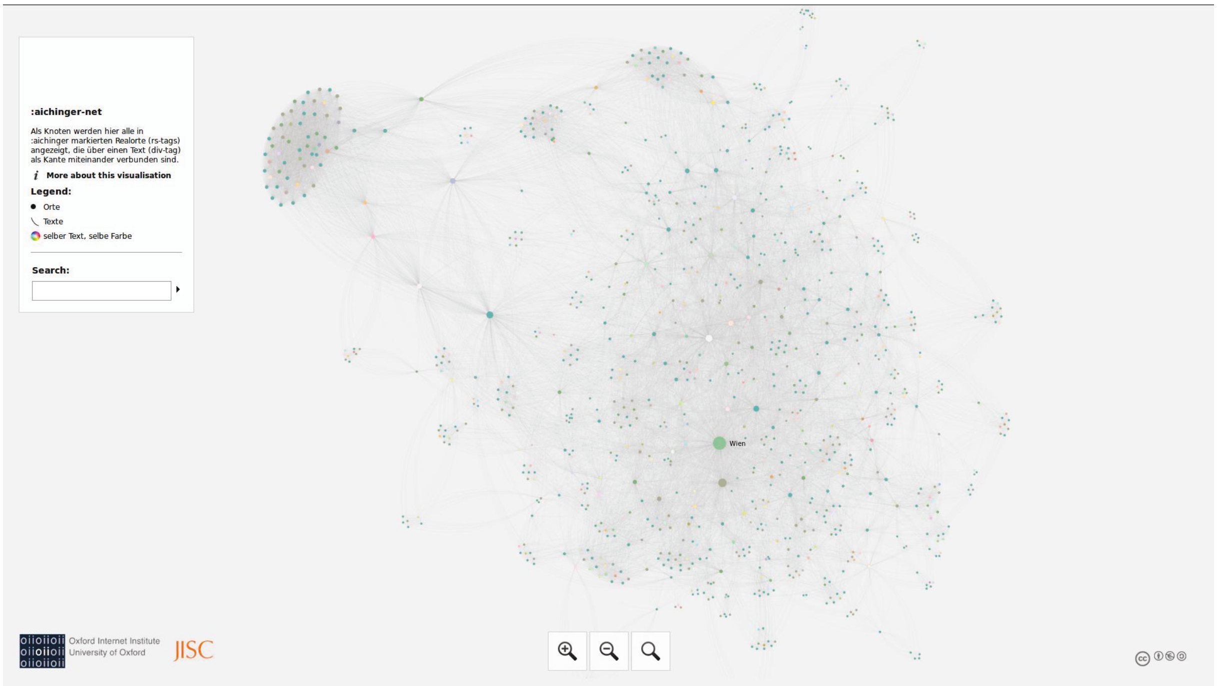


```
<body>
<div n="1" ana="ed19480000 eb19480000" type="prose">
<head>Die größere Hoffnung</head>
<div type="chapter">
<pb facs="../files/Aichinger-Hoffnung_n0009.tif" n="9"/>
<head>Die große Hoffnung</head>
<p>Rund um das <rs type="place">Kap der Guten Hoffnung</rs> wurde das Meer<lb/>dunkel.
Die Schiffahrtslinien leuchteten noch einmal auf und<lb/>erloschen. Die Fluglinien
sanken wie eine Vermessenheit.<lb/>Ängstlich sammelten sich die Inselgruppen. Das
Meer<lb/>überflutete alle Längen- und Breitengrade. Es verlachte das<lb/>Wissen der
Welt, schmiegte sich wie schwere Seide gegen das<lb/>helle Land und ließ die <rs
type="place">Südspitze von <rs type="place">Afrika</rs></rs> nur wie eine<lb/>Ahnung
im Dämmern. Es nahm den Küstenlinien die<lb/>Begründung und milderte ihre
Zerrissenheit.</p><p>Die Dunkelheit landete und bewegte sich langsam gegen<lb/>Norden.
Wie eine große Karawane zog sie die Wüste hinauf,<lb/>breit und unaufhaltsam. Ellen
schob die Matrosenmütze aus<lb/>dem Gesicht und zog die Stirne hoch. Plötzlich legte
sie die<lb/>Hand auf das <rs type="place">Mittelmeer</rs>, eine heiße kleine Hand.
Aber es half<lb/>nichts mehr. Die Dunkelheit war in die <rs type="place">Häfen von <rs
type="place">Europa</rs></rs><lb/>eingelaufen.</p><p>Schwere Schatten sanken durch die
weißen Fensterrahmen.<lb/>Im Hof rauschte ein Brunnen. Irgendwo verebbte ein
Lachen.<lb/>Eine Fliege kroch von <rs type="place">Dover</rs> nach <rs
type="place">Calais</rs>.</p>
```

Figure 1: Exemplary screenshot of a TEI-XML-file.

This resulted in about 1.800 references to real places. Previous scholarly works have not seen this multitude of references in the text (Schmid-Bortenschlager, 2001). Moreover the text with the most quantity and diverse real-place-references ("Nachricht vom Tag") is, surprisingly or not, one which among Aichinger-scholars is very rarely discussed (see figure 2). Further it could be shown, that real-place-references are not exclusive but predominant in Aichinger's later work (see Fig. 3). Previously similar results have been shown with simple text-query-statistics (Frank and Dittrich, 2015, p. 52–53).

Although promising techniques of automatic place-name-recognition are in development (Bornet and Kaplan, 2017) the annotations have been made manually. The special challenge in this case was, to get to terms relating to the different types of place-references. Mahler and Dünne proposed to differ between "topography" and "topology" (Dünne and Mahler, 2015, p. 6). Topographical entities operate in a semantic reference system and can therefore be mapped. Topological entities operate in a syntactic relation system and are therefore able to get located in a network. It is not easy to differ between those two categories in every case. The notion of "Dover" for example can refer both to the real place in the south of Great Britain and be an empty signifier not referring to anything at all (Aichinger, 1991a). Only out of this undecidable entanglement the playful meaning of the text arises. To grasp the interwoven conjunc-



:aichinger-net

Als Knoten werden hier alle in :aichinger markierten Realorte (rs-tags) angezeigt, die über einen Text (div-tag) als Kante miteinander verbunden sind.

Figure 2: Network-view of all tagged place-names: nodes are place-names, edges are text-divisions. The cluster in the upper left corner represents the place-names in the text "Nachricht vom Tag". This graph can be explored online: `http://homepage.univie.ac.at/andreas.dittrich/aichinger-rsnet`. Visualization made with *Gephi* and *sigmajs.org*.

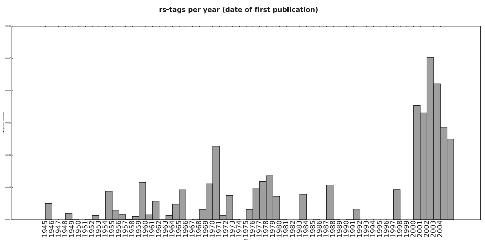


Figure 3: A time-based view (first publication) of place-name-frequency per characters.

tion of topographic and topologic entities it is imperative to first address them separately. But although this distinction is useful for first steps, it does not exhaust the many possibilities of place-references in literature. To name just a few, which we encountered:

- place-names referring to real places and which are mappable;
- common place-names like "kitchen" or "park";
- fictional places like a "fan";
- and place-names that simply cannot be located like "Port Sing".

3 Towards an exploratory framework

Heinz Schafroth suggested to read the texts of Aichinger "associative" (Schafroth, 1976, p. 130), that is to say: reading the intra-connectivity of the different texts. Following this proposal, we can represent the texts as a network and make it explorable as such. Simple methods in Corpus-Studies work with types or word-forms and answer questions like: where, how often and in which context can I find a specific word in the corpus. Even queries about co-occurrences of words are possible. But how about words that share the same co-occurrences, but not the same words?

Say, for example, the word "Vater" occurs in a set of texts A, "Mutter" in a set of texts B. Let us call the overlapping of shared words AB. Now, there are texts, which share words with the set of text A, not B, but share words with a set of texts C, which share words with B (see Fig. 4). This set of texts C can be interesting for analysis — and maybe this is, what Peirce called "abductive reasoning" —, but it would be difficult to reach within the boundaries of conventional queries.

A SPARQL-server (Apache Jena Fuseki), which stores RDF-files, is used. If a simple RDF-turtle-file would contain the following informa-

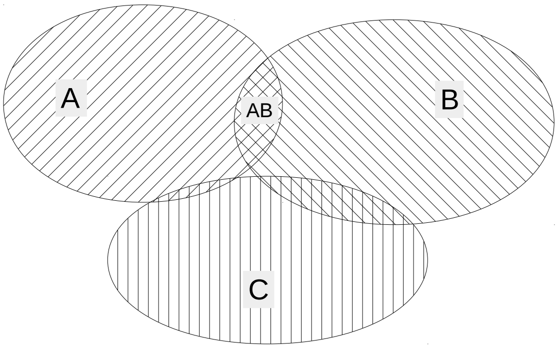


Figure 4: Illustration of set of texts.

tion:

```
@prefix word: <ia.net/words#>.
@prefix rel:  <ia.net/relat#>.
word:father rel:str  "Father".
word:father rel:tr word:vater.
word:vater  rel:str   "Vater".
```

a simple query could look like this:

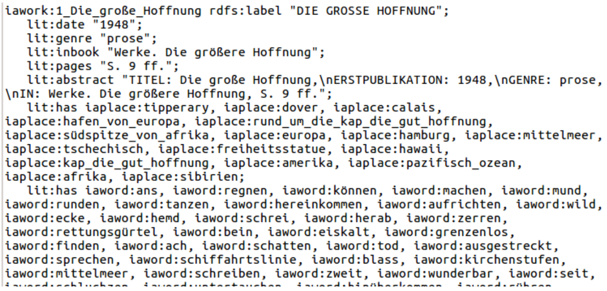
```
SELECT ?o1 ?p2 ?o2
WHERE {
  ?m1 ?p1 ?o1 .
  ?m1 ?p2 ?m2 .
  ?m2 ?p3 ?o  .
}
```

which would result in a formatted output like this:

```
"Father" rel:tr "Vater"
```

To this end, the TEI-XML-files have to be converted in RDF, for which an special Python-program had to be written. In the RDF-file all words, which are in the same division of text, are connected with each other (this is useful, because Aichinger mainly wrote short texts); date-, genre- and place-annotations are linked to the division (see Fig. 5).

To explore such a network not only by its "most frequent" or "most linked" terms one needs to be able to move inside of this network intuitively. Text-corpora of about 400.000 tokens could result in a network of about 81 billion connections. But immediate interactive and visual exploration of these networks is needed. It should be possible to alter the graph (for example to add, drag or remove certain nodes or edges) and see the results without delay. Developed for the so-called semantic web, which works with structured data, the open-source software "RelFinder" offers a suitable framework to make such interactive queries (see `http://relfinder.visualdataweb.org`).



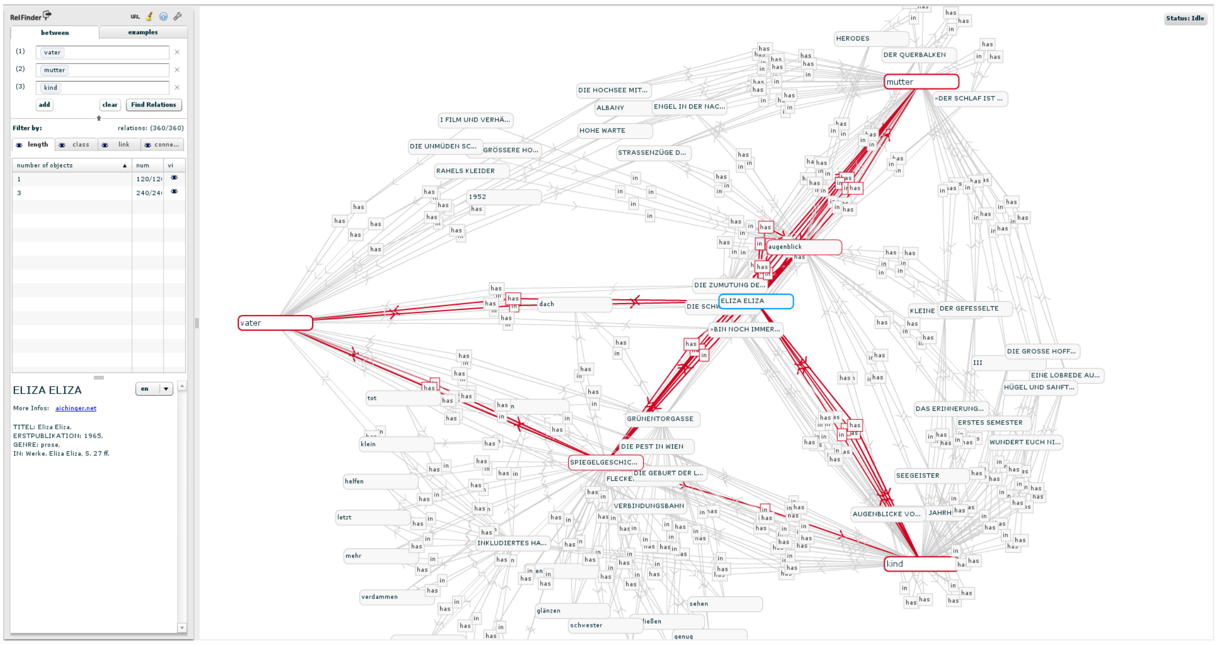
```
iawork:1_Die_große_Hoffnung rdfs:label "DIE GROSSE HOFFNUNG";
   lit:date "1948";
   lit:genre "prose";
   lit:inbook "Werke. Die größere Hoffnung";
   lit:pages "S. 9 ff.";
   lit:abstract "TITEL: Die große Hoffnung,\nERSTPUBLIKATION: 1948,\nGENRE: prose,
\nIN: Werke. Die größere Hoffnung, S. 9 ff.";
   lit:has iaplace:tipperary, iaplace:dover, iaplace:calais,
iaplace:hafen_von_europa, iaplace:rund_um_die_kap_die_gut_hoffnung,
iaplace:südspitze_von_afrika, iaplace:europa, iaplace:hamburg, iaplace:mittelmeer,
iaplace:tschechisch, iaplace:freiheitsstatue, iaplace:hawaii,
iaplace:kap_die_gut_hoffnung, iaplace:amerika, iaplace:pazifisch_ozean,
iaplace:afrika, iaplace:sibirien;
   lit:has iaword:ans, iaword:regnen, iaword:können, iaword:machen, iaword:mund,
iaword:runden, iaword:tanzen, iaword:hereinkommen, iaword:aufrichten, iaword:wild,
iaword:ecke, iaword:hemd, iaword:schrei, iaword:herab, iaword:zerren,
iaword:rettungsgürtel, iaword:bein, iaword:eiskalt, iaword:grenzenlos,
iaword:finden, iaword:ach, iaword:schatten, iaword:tod, iaword:ausgestreckt,
iaword:sprechen, iaword:schiffahrtslinie, iaword:blass, iaword:kirchenstufen,
iaword:mittelmeer, iaword:schreiben, iaword:zweit, iaword:wunderbar, iaword:seit,
```

Figure 5: Exemplary screenshot of a RDF-file.

See Fig. 6 for an example of how the words "vater", "mutter" and "kind" are related: five texts appear in the center of the graph, which share the three words. What may catch the eye of an Aichinger-scholar is the centrality of the term "augenblick" (blink of an eye, moment), which is central to her concept of "hope" (Thums, 2013, p. 193–196), which leads to her novel "The Greater Hope" (Aichinger, 2016). And it not only seems to connect all other nodes but it connects most of the nodes, which are connected with the three searched ones. "Die Zumutungen des Atmens" for example is connected to "mutter", "vater" and "kind", but also to "augenblick", which it shares with "Die Spiegelgeschichte". (The same can be said for "Die größere Hoffnung" (chapter), "Eliza Eliza" (text) and "Die Schwestern Jouet" (drama). The only Text, which does not share "augenblick", but all other words, is "Bin noch immer positiv!".) Although some text do not share all the searched words, many of them share the word "augenblick". It is possible to lessen the graph's output by different mechanisms. See Fig. 7 where all relations, that are direct only, are faded out. The nodes "augenblick" and "Die Spiegelgeschichte" stay in the center and suggest a high connectivity.

4 Conclusion and Discussion

Although the presented approach is in developement and not all possibilities are exhausted yet, it could be shown what the basic idea enables: Finding maybe unexpected connections between texts and by this, enabling new insights into already known connections and discovery of unknown interrelations. The concept of "Augenblick" has already been in the focus of Aichinger-scholars, but



RelFinder

between | examples

(1) vater
(2) mutter
(3) kind

add | clear | Find Relations

Filter by: relations: (360/360)

length | class | link | conne...

number of objects	num	vi
1	120/12	
3	240/24	

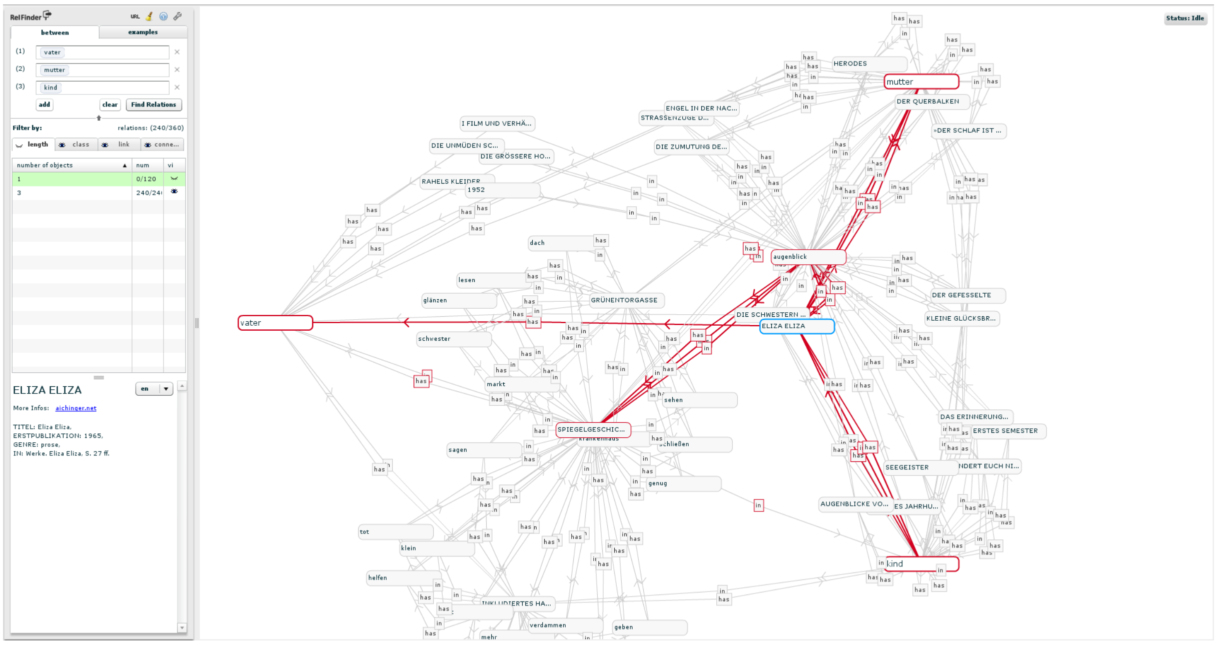
ELIZA ELIZA en

More Infos: aichinger.net

TITEL: Eliza Eliza.
ERSTPUBLIKATION: 1965.
GENRE: prose.
IN: Werke. Eliza Eliza. S. 27 ff.

Status: Idle

Figure 6: A view of all associations of “vater”, “mutter” and “kind” in the corpus `:aichinger` with *RelFinder*.



RelFinder

between | examples

(1) vater
(2) mutter
(3) kind

add | clear | Find Relations

Filter by: relations: (240/360)

length | class | link | conne...

number of objects	num	vi
1	0/120	
3	240/24	

ELIZA ELIZA en

More Infos: aichinger.net

TITEL: Eliza Eliza.
ERSTPUBLIKATION: 1965.
GENRE: prose.
IN: Werke. Eliza Eliza. S. 27 ff.

Status: Idle

Figure 7: A selected view of the associations: Only those nodes are displayed whose connections are not only direct.

not in the perspective of its relatedness to other texts and words.

A crucial point in this types of visualisations is the eclipse of the dimension of time. The graph seems to suggest, that these words are used in a timeless space. I tried to adjust this by listing meta-data with first publications-date and genre on the left side of the screen. To make it easier to read the texts in their entire context, the book and page-number, where the texts can be found, are listed.

Of course some methodological problems do persist in this approach. One, that troubles me basically, is that this approach seems to assume that words mean the same in different contexts. But they don't. Not even in, or maybe most notably not in literature. By unifying different singular occurrences of a word to one word-type, the singular use in a singular context gets covered. One has to be vigilant to not level important differences. Ilse Aichinger wrote a text called "Hemlin", which performs the variability of words by using the untranslatable (or exactly translatable) word "Hemlin" in various ways (Markus, 2015, p. 89–90) and questions the – sometimes undue – unifying drive of scientific methods: "Hemlin must be a monument, round, makes trouble." (Wolf and Hawkey, 2010, p. 191). Hemlin.

References

Ilse Aichinger. 1991a. *Schlechte Wörter*, volume 4. Fischer.

Ilse Aichinger. 1991b. *Zu keiner Stunde. Szenen und Dialoge*, volume 7. Fischer.

Ilse Aichinger. 2001. *Kurzschlüsse*. Edition Korrespondenzen.

Ilse Aichinger. 2016. *The Greater Hope*. Königshausen & Neumann.

Cyril Bornet and Frédéric Kaplan. 2017. A simple set of rules for characters and place recognition in french novels. *Frontiers in Digital Humanities*, 4:6.

John Anthony Cuddon. 2013. *A Dictionary of Literary Terms and Literary Theory*. Wiley-Blackwell, 5 edition.

Jörg Dünne and Andreas Mahler. 2015. Einleitung. In Jörg Dünne and Andreas Mahler, editors, *Handbuch Literatur & Raum*, pages 1–11. Walter de Gruyter.

Andrew U. Frank and Andreas Dittrich. 2015. Flexible annotation of digital literary text corpus with rdf. In Francesco Mambrini, Marco Passarotti, and Caroline Sporleder, editors, *Proceedings of the Workshop on Corpus-Based Research in the Humanitie*, pages 49–58. Institute of Computer Science. Polish Academy of Sciences.

Simone Fässler. 2011. *Von Wien her, auf Wien hin*. Böhlau.

Hans-Georg Gadamer. 2004. *Truth and Method*. Bloomsbury, 2 edition.

Philipp Heim, Steffen Lohmann, and Timo Stegemann. 2010. Interactive relationship discovery via the semantic web. In *Proceedings of the 7th Extended Semantic Web Conference (ESWC 2010)*, volume 6088 of *LNCS*, pages 303–317. Springer.

Matthew L. Jockers. 2013. *Macroanalysis*. Topics in the Digital Humanities. University of Illinois Press.

Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In R. Borgo, F. Ganovelli, and I. Viola, editors, *Eurographics Conference on Visualization (EuroVis) - STARs*. The Eurographics Association.

Hannah Markus. 2015. *Ilse Aichingers Lyrik. Das gedruckte Werk und die Handschriften*. De Gruyter.

Annegret Pelz. 2009. Spracharbeit in meeresnähe. In Ingeborg Rabenstein-Michel, François Rétif, and Erika Tunner, editors, *Ilse Aichinger – Misstrauen als Engagement?*, pages 63–72. Könighausen & Neumann.

Heinz F. Schafroth, 1976. *Die Dimensionen der Atemlosigkeit*, pages 129–133. Fischer.

Sigrid Schmid-Bortenschlager. 2001. Die topographie ilse aichingers. In *Was wir einsetzen können, ist Nüchternheit*, pages 179–188. Königshausen & Neumann.

Olga Scrivner and Jefferson Davis. 2017. Interactive text mining suite: Data visualization for literary studies. In Thierry Declerck and Sandra Kübler, editors, *Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 2017)*, pages 29–38.

John Sinclair. 2001. Preface. In Mohsen Ghadessy, Alex Henry, and Robert L. Roseberry, editors, *Small Corpus Studies*, page vii–xvi. Benjamin Publishing.

John Sinclair. 2004. *Trust the Text: Language, Corpus and Discourse*. Routledge.

Barbara Thums. 2013. Zumutungen, ent-ortungen, grenzen. In Doerte Bischoff and Susanne Komfort-Hein, editors, *Literatur und Exil*, pages 183–209. De Gruyter.

Uljana Wolf and Christian Hawkey. 2010. Notizen beim Übersetzen von aichingers hemlin. *Berliner Hefte zur Geschichte des literarischen Lebens*, (9):188–191.