An End-to-end Environment for Research Question-Driven Entity Extraction and Network Analysis

Andre Blessing$^{\heartsuit}$ and **Nora Echelmeyer**$^{\clubsuit}$ and **Markus John**$^{\diamondsuit}$ and **Nils Reiter**$^{\heartsuit}$

$^{\heartsuit}$ Institute for Natural Language Processing
$^{\clubsuit}$ Institute for Literary Studies
$^{\diamondsuit}$ Institute for Visualization and Interactive Systems
Stuttgart University
{andre.blessing@ims, nora.echelmeyer@ilw}.uni-stuttgart.de
{markus.john@vis, nils.reiter@ims}.uni-stuttgart.de

Abstract

This paper presents an approach to extract co-occurrence networks from literary texts. It is a deliberate decision not to aim for a fully automatic pipeline, as the literary research questions need to guide both the definition of the nature of the things that co-occur as well as how to decide co-occurrence. We showcase the approach on a Middle High German romance, *Parzival*. Manual inspection and discussion shows the huge impact various choices have.

1 Introduction

The main contribution of this paper is the presentation of a conceptualized and implemented workflow for the study of relations between entities mentioned in text. The workflow has been realized for multiple, diverse but structurally similar research questions from Humanities and Social Sciences, although this paper focuses on one from literary studies in particular. We see this workflow as exemplary for research involving Natural Language Processing (NLP) and Digital Humanities (DH), in which operationalization and modularization of complex research questions often has to be a first step. It is important to realize that this modularization can not be guided by NLP standards alone – the interests of the respective humanities discipline need to be considered, and practical considerations regarding timely availability of analyses as well: If a large portion of the funding period is spent with developing, adapting and fine-tuning NLP tools, the analysis of the results (with often leads to new adaptation requests) risks being missed out.

Our workflow combines clearly defined tasks for which we follow the relatively strict NLP paradigm (annotation guidelines, gold standard, evaluation) with elements that are more directly related to specific Humanities research questions (that often are not defined as strictly). The final module of this workflow consists in the manual exploration and assessment of the resulting social networks by literary scholars with respect to their research questions and areas. In order to enable scholars to explore the resulting relations, we make use of interactive visualization, which can also show developments and changes over time.

More generally, this workflow is the result of ongoing work on the modularization and standardization of Humanities research questions. The need for modularization is obvious for computer scientists (and computational linguists), as they are often consciously restricting their tasks to clearly defined problems (e.g., dependency parsing). However, this opposes typical Humanities research style, which involves the consideration of different perspectives, contexts and information sources – ignoring the big picture would be a no-go in literary studies. This makes research questions seemingly unique and incomparable to others, which in turn leaves little room for standards applied across research questions.

Our ultimate goal is to develop methodology that supports the work of humanities scholars on their research questions. This in turn makes interpretability of the results of NLP-tools an important constraint, which sometimes goes against the tendency of NLP research to produce methods that are solely judged on their prediction performance. However, we intentionally do not focus on tool development: The appropriate use of tools and adequate interpretation of their results is of utmost importance if these form the basis of hermeneutical interpretations. To that end, scholars need to understand fundamental concepts of quantitative analysis and/or machine learning.

The trade-off between interpretability and pre-

diction performance has also been discussed in other projects, e.g. in Bögel et al. (2015). In our project we follow two strategies: (i) Offering visualization and inspection tools as well as a close feedback loop and (ii) integrating humanities scholars early into the development cycle, such that they are involved in the relevant decisions.

Parzival We will use *Parzival* as an example in this paper, because it involves a number of DH-related challenges. The text is an Arthurian grail novel and has been written between 1200 and 1210 CE by Wolfram von Eschenbach in Middle High German. The text comprises of 25k lines and is divided into 16 books. The story of the books mainly follows the two knights Parzivâl and Gâwân and their interaction with other characters. One of the key characteristics of *Parzival* is a large inventory of characters that have complex genealogical patterns and familial relations. This led to an ongoing discussion about the social relations in *Parzival* (Bertau, 1983; Delabar, 1990; Schmidt, 1986; Sutter, 2003), which are much more complex than in other Arthurian romances (*Erec*, *Iwein*). The systematic comparison of the social/spatial relations in different narrations of a similar story is one of our goals. With that in mind, we investigate various operationalization options for these networks.

2 Workflow

Given the above discussion about *Parzival*, we are aiming to establish a workflow to extract social networks from text, such that scholarly/domain experts are enabled to compare the resulting networks from different narrations. Therefore, the steps in this workflow need to be reasonably transparent, errors traceable and the overall results interpretable for scholars without deep technical background.

Next to our example case *Parzival*, we believe that many research questions in Humanities and Social Sciences can be cast as such a network/relation extraction task, at least on a structural level: Studying the relation of characters in narrative texts is structurally similar to the relation of concepts in philosophical texts, for instance. The workflow we employ consists of the following steps:

1. Identification of textual references to entities of various types (Sect. 3),

2. Grounding of detected entity references (e.g.,

identifying "the knight" as a reference to the main character Parzivâl; Sect. 4),

3. Segmentation of the texts in appropriate parts (e.g., story taking place at a specific location; Sect. 5),

4. Manual, interactive exploration of proto-networks for validation (Sect. 6), and

5. Creation and analysis of networks of entities that co-occur within a segment (e.g., the characters that take part in a great feast; Sect. 7).

It is important to note that this workflow is impacted by the Humanities research question at multiple stages. The notion of **entity** is relatively generic and we have applied it to a number of different genres. However, in order to group entity references into entity types, one has to determine what entity types are actually relevant in the text at hand and for the specific research question. While we assume intersubjective agreement on entity annotations, we make no such assumption for **segment** annotations. Different segmentation criteria can be tested and the resulting networks compared.

In general, we make no assumptions on every step being automatic. Semi-automatic, manual, interpretative or other kinds of work packages can be integrated in such workflows (and, given the nature of (Digital) Humanities, often need to be).

The *Parzival* corpus is preprocessed by several webservices from the CLARIN infrastructure[1] (Mahlow et al., 2014) to obtain a sentence splitted, tokenized and part-of-speech tagged corpus for the previously described workflow steps.

3 Detecting Entity References

3.1 Conceptualisation and Annotation

We define entities as individually distinguishable objects in the real or a fictional world. Words in texts may refer to entities and are thus called entity references (ERs). Linguistically, entity references can be expressed as proper names, pronouns and appellative noun phrases (which together are typically called *mentions* within coreference resolution). Our annotations include only proper names and appellative noun phrases, pronouns have been excluded by definition. Therefore, the task described here is situated in between

---

[1]European Research Infrastructure for Language Resources and Technology: https://www.clarin.eu/

the well-known NLP tasks of named entity recognition (NER) and coreference resolution. This was a pragmatic decision, in order to avoid the most difficult coreference resolution challenges (as pronouns are the most ambiguous mentions) and still include more occurrences than just names. In addition, the referents for appellative noun phrases can be resolved with only a limited amount of context, which makes their grounding (cf. Sec 4) faster for Humans and more promising to automatically support. Our annotation scheme distinguishes between different types of entities. Entity references are marked with the type of the entity they refer to.

We annotate manually five books of *Parzival*, following the annotation guidelines developed in parallel with the annotation process[2]. The manual annotation is done in parallel by two different annotators. Annotation differences have been adjudicated by a third person, after discussion with the annotators. Difficult cases have been discussed with annotation groups for different texts.

In *Parzival*, two different types are actually appearing: Persons and locations. Table 1 shows the distributions of the entity types across the five books that constitute our gold standard. As can be seen, the variance across the books is quite low.

Book	Lines	Tokens	PER	LOC
III	1,898	12,015	610	120
IV	1,338	8,035	464	122
V	1,682	10,441	472	140
VI	1,740	10,918	594	144
VII	1,800	11,358	687	134
Mean	1,691.6	10,553.4	565.4	132
SD	213.2	1,522.5	95.7	10.7

Table 1: Corpus Statistics. PER/LOC: References to persons or locations, SD: Standard deviation.

3.2 Automatic Entity Reference Detection

Our entity reference tagger is built using ClearTK (Bethard et al., 2014), which in turn employs mallet CRF (McCallum, 2002) and the BIO scheme.

Feature set The features presented in Table 2 are extracted for the current, two preceding and one succeeding tokens. Since we are applying this

---

[2]The (German) guidelines are available on the project web site http://www.creta.uni-stuttgart.de.

tagger to different corpora in different languages, we use language-, genre- or text-specific resources only in two, clearly defined cases: part of speech and names gazetteers. Part of speech taggers are available even for many low resource languages (or are among the first being created), gazetteers can often be created by domain experts.

Id	Feature	Description
F_1	Surface	The surface form of the token
F_2	PoS	The part of speech tag of the token. For *Parzival*, we are using a fairly new, publicly available model (Echelmeyer et al., 2017) for tree-tagger (Schmid, 1994).
F_3	Case lookup	Do tokens written in upper case also exist in lower case?
F_4	Unicode character pattern	A canonicalized list of Unicode character properties that appear in the token. "Obilôte", for instance, is represented as `LuLl`, signaling upper case letters followed by lower case letters.
F_3	Gazetteer	A list of names. The gazetteer in our experiment has been collected from various MHG texts by extracting tokens with upper case letters and manually removing the non-names. Generally, this feature allows the inclusion of domain knowledge in a simple and broadly applicable manner.

Table 2: Feature set

3.3 Evaluation

The entity reference tagger is evaluated using book-wise cross-validation (i.e., 5-fold CV). In the **strict** setting, we only count exactly matching boundaries as correct, while in the **loose** setting, we count a true positive as long as there is a one-token overlap between system and reference.

Baselines We compare the entity reference tagger against two baselines: The Stanford named entity recognizer for Modern German (BL_{NER}) and marking every upper case word as a majority class reference (i.e., Person; BL_{Case}).

		Person		Location	
		Prec	Rec	Prec	Rec
strict	BL_{NER}	27.3	1.2	27.6	2.4
	BL_{Case}	36.2	19	0	0
	ERT	71.2	56.8	71.8	48
loose	BL_{NER}	72.9	3.6	41.9	3.9
	BL_{Case}	74.8	38.5	0	0
	ERT	91.6	76.1	85.3	57.9

Table 3: Evaluation results for the entity reference tagger (ERT), compared with two baselines (NER and case-based).

Discussion The results achieved purely automatically can be seen in Table 3. As expected, evaluation scores for the loose setting are higher than for the strict setting. The loose setting in fact is more representative for the actual performance of the tagger in our workflow. As the domain experts perform semi-manual grounding anyway, the exact boundaries of the found entity reference are not that important. In addition, manual inspection revealed that in many cases the entity tagger in fact marked the head of the noun phrase.

Performance scores are higher for Person references, which can be attributed to their frequency. Both baselines are clearly outperformed, although both have their presumed strengths for the proper noun references. Manual inspection also revealed that most of the remaining recall errors are appellative noun phrases (e.g., "des burcgrâven tohterlîn"/"the burgrave's daughter").

3.4 Semi-Automatic Labeling

Although automatic labeling of entity references is an important part of our workflow, a recall error of about 25% of the persons severely limits its usefulness for applications in digital literary studies. We therefore implemented a user interface (not shown) in which scholars can inspect the found entity references on unseen texts and mark them as either *correct*, *incorrect*, or *boundary-incorrect*, for span errors as well as subsequently annotate missing entity references (recall errors). For once, these annotations are then stored as manual annotations that can be used in subsequent workflow steps and secondly, they can be employed as additional training material.

This procedure also gives clear guidance on where to focus the improvements of the tagger:
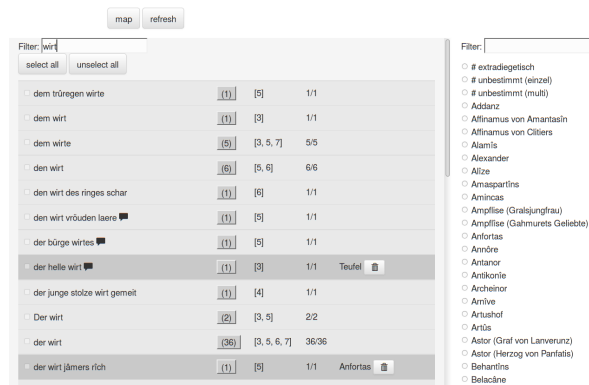

Figure 1: Grounding view: Annotated entity reference types (left) mapped to characters (right)
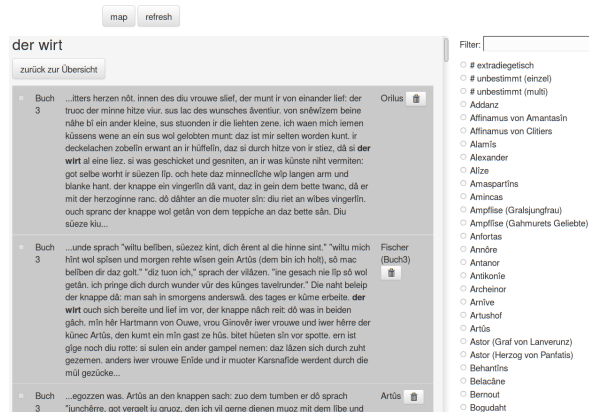

Figure 2: For each mention type all occurrences including textual context can be shown

Precision errors are much easier to spot and correct in this fashion, making improvements in terms of recall our priority.

4 Entity Grounding

Each annotated entity reference is mapped to a pre-defined list of characters[3]. This task can be seen as entity linkage or entity grounding (Ji and Grishman, 2011). In this paper we restricted the mapping to persons but it can easily be applied to other entity classes. While the entity reference detection task was supported automatically, grounding is done manually. Fig. 1 displays the user interface of the mapping tool. The detected entity references are listed on the left, and the characters on the right. Each surface form can appear more than once in the text (e.g., 36 times "der wirt"). The user has two options for the grounding: a) map all occurrences to one character; b) consider

---

[3]The character list was created in several iterations by merging already existing lists and automatically extracted candidates from the corpus.

Character	#ER	#Proper nouns	Ratio
Parzivâl	427	111	25.8
Gâwân	185	118	63.8
Artûs	128	88	68.8
Jeschûte	103	30	29.1
Clâmidê	74	47	63.5
Herzeloyde	69	9	13

Table 4: Ratio of proper nouns among references

each textual context and map each occurrence differently (Fig. 2).

Table 4 shows the grounding result for some main characters. These characters can be divided into two classes: i) characters which are often refered by their name (Gâwân, Artûs, Clâmidê); ii) characters (Parzivâl, Jeschûte, Herzeloyde) which are mainly referred to by appellative noun phrases.

5   Text Segmentation

For the later network analysis a segmentation is needed to define windows in which relations between characters are extracted. In contrast to the task of entity reference detection, we do not cast text segmentation as a 'real' NLP task, for which we create annotation guidelines, train annotators etc. The reason for this is that this task is directly related to the research question a scholar wants to investigate. It is difficult to imagine context- and text-independent criteria for the segmentation. Annotation guidelines created for *Parzival* might not generalise to other texts.

We therefore explore segmentation approaches based on linguistic and structural criteria and with regards to the content. For all segmentation settings, we (manually) removed non-narrative sections (in which the heterodiegetic narrator gives comments; cf. Coste and Pier (2014)).

Linguistics   Straightforwardly, one can segment according to sentences. Sentence boundaries have been detected automatically using rules based on punctuation. The lack of abbreviations in *Parzival* also removes the most frequent error source for punctuation-based sentence splitting. Each sentence is considered an individual segment.

Structure   *Parzival* is structured into strophes of 30 lines each. There is no apparent meaning to these strophes, and sometimes sentences are split

over multiple strophes. In this segmentation setting, each strophe is used as a segment.

Content   Content-based segments are designated as episodes. An episode is a self-contained and homogenous segment of the story. Typical indicators of an episode break are changes in character constellations, time and/or space. Episodes have been annotated manually by one of the authors.

6   Interactive Visual Exploration

To inspect and verify automatic results we use a web-based tool that supports close and distant reading. It provides different views including word clouds, plot views, and graph visualizations that allow analyzing entities and exploring their relationships. Each view allows to directly access the corresponding text passage(s).

In this context particularly relevant is the interactive graph visualization, through which processing errors become apparent quickly. The graph visualization uses a force-directed graph layout and represents the relations between entities, as depicted in Fig. 3 (on the left side). The nodes represent characters/persons and the edges the relations between them. The view is complemented by a fingerprint visualization (A) that indicate where the characters are mentioned in the text. A range slider (B) enables users to select a certain range of the text, for example a single chapter. This way, users can analyze not only the overall text but also the development of the relationships between characters. In a list view (C) users can dynamically adapt the network by selecting or deselecting the entities in the list. Furthermore, they can select an edge in the network view to highlight the co-occurrences of two related entities in the fingerprint visualization. By selecting an occurrence, users can jump to the corresponding text passage as depicted in Fig. 3 (right side).

In the text view the selected entities are highlighted in their assigned color. The background color (orange) represent the respective text segmentation. Next to the scrollbar, a vertical fingerprint displays the further co-occurrences. By hovering over an occurrence the corresponding text passage is displayed in a tooltip, as depicted in (D). After clicking on one, the text view jumps to the corresponding position. This way, users can easily analyze and compare the relevant text passages of the selected entities. With the aid of both views, users can determine incorrect relationships
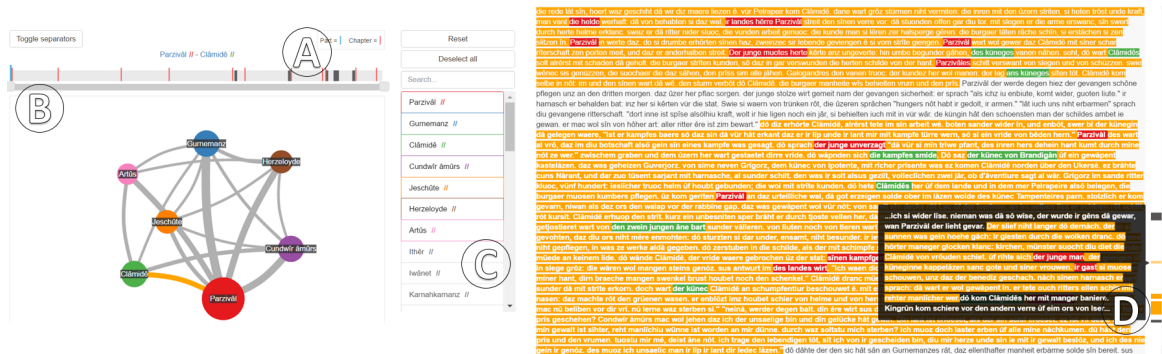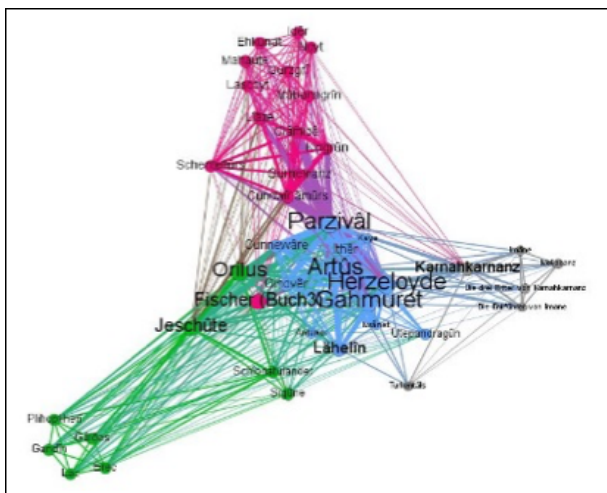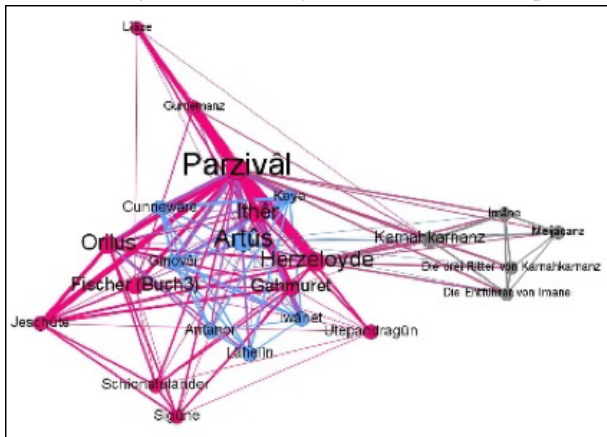
Figure 3: Graph visualization and the text view with selected entities Parzivâl and Clâmidê.

(a) including embedded entity references and direct speech

(b) without embedded entity references and direct speech. Several nodes and sub-networks (around Gurzgrî or Gâlôes) disappear; some characters (Gahmuret) become less important because they are often embedded in other entities' expressions.

Figure 4: Entity network of Book III, using content-based segmentation

(graph visualization) and inspect or verify them in detail (text view).

| | Direct speech | + | − | − |
	Embedded ERs	+	+	−
Nodes		41	24	24
Edges		431	144	130
Density		0.526	0.522	0.47
Avgerage degree		21	12	10.8

Table 5: Influence of removing direct speech and embedded entity references on network parameters. Density: Edges / Possible edges; degree: Number of edges in a node.

7 Network Analysis

In this section we compare different parameters of network visualization and analyze their interdependencies and influences on the network results. In the moment, we focus on person-based networks and leave aside the spatial information, which can be used in further analysis, e.g., to distinguish between static and dynamic characters or to detect events (cf. Lotman, 1977).

All network graphs are created with Gephi (Bastian et al., 2009), which provides various layout algorithms, offers statistics and network metrics, and supports dynamic graph visualization. Plots and tables in this section are based on Book III of *Parzival*.

7.1 Embedded entities and direct speech

As a first step, we explore the influence of ERs within a) other ERs (embedded) and b) direct speech. This is due to the fact that neither embedded entities (as Gahmuret in "vil li roy Gahmuret"/"son of the king Gahmuret") nor entities mentioned in (direct) speech are neccesarily taking part in the narrated story or event.

Fig. 4 demonstrates the influence of embedded

Segmentation	by sentence		by strophe		by content	
Entity grounding	+	−	+	−	+	−
Edges	24	18	24	19	130	74
Nodes	65	25	79	26	24	20
Density	0.18	0.15	0.27	0.16	0.47	0.39
Avgerage degree	5.4	2.74	6.48	2.84	10.83	7.4
Connected components	1	3	1	3	1	1
Network diameter	4	4	3	6	2	3

Table 6: Influence of the different segmentation types on the network parameters. Comparison of the network parameters with and without entity grounding. Connected components: Isolated groups of nodes (lower number: stronger connectivity); diameter: Largest distance between two nodes.

entites and direct speech visually, Table 5 provides a numerical view. The network without embedded ERs or direct speech is less dense (0.47 vs. 0.526), the average degree is much lower (10.8 vs. 21), and the number of nodes and edges decreases from 41 nodes and 431 edges to 24 nodes and 130 edges.

7.2 Segmentation

Figures 5a-c show the effects of the different segmentation criteria (cf. Sect. 5) on the networks, quantitative network properties are displayed in Table 6. First we observe a decrease in density from the largest content-based segmentation (0.47) to the medium-sized segmentation in strophes (0.27) to the smallest segmentation in sentences (0.177), which comes along with a reduction of the number of edges (from 130 to 79 to 65). The highest average degree can be found in the content-based network (10.8), in the strophe-based network it is reduced to 6.5 and in the sentence-based network it decreases further to 5.5. As the nodes are less and less connected, the network diameter becomes bigger the smaller the segmentation gets. Since the chosen segmentation serves as basis for the extraction of co-occurrences of characters, it has a huge impact on the network properties which is important to reflect for later interpretations of the network.

7.3 Entity grounding

To estimate the influence of entity grounding on the networks we compare networks based on entity grounding (Sect. 4) to those only based on proper names. We identify an interdependency between entity grounding and segmentation. The sentence-based and strophe-based segmentation are relatively small and therefore more dependent on entity grounding. The co-occurrence of two proper names in a sentence is rare (in Book III of *Parzival*: 25 co-occurrences in over 750 sentences), even in a strophe it rarely appears (26 co-occurrences in 63 strophes). As we see in Fig. 5, the sentence- and strophe-based networks without entity grounding disintegrate into three components, they become less dense and most of the relations get lost.

The last example (Fig. 6) shows the high influence of entity grounding and its importance for an appropriate representation of the character configurations of a narrative text. The unequal ratio of proper names and nouns (cf. Table 4) underlines the importance: By only taking into account the proper names, we found that Parzivâl is mentioned 111 times (in Books III-VII) and that Gâwân is mentioned 118 times. Whereas the consideration of other references to both characters leads to the fact that Parzivâl amounts to 427 mentions and Gâwân to 185 mentions. This means that the primacy of Parzivâl only becomes apparent by including entity grounding.

Thus, being aware of the chosen parameters is a precondition for an adequate analysis and interpretation of the networks. To represent the plot and narrative structure of *Parzival* by analyzing the development of character configurations over time, for instance, it is necessary to exclude embedded entities and direct speech as well as to include entity grounding. The appropriate way of segmentation still needs to be reconsidered.

8 Related Work

Several researchers have extracted co-occurrence networks from dramatic texts or screen plays (Moretti, 2011; Trilcke et al., 2016; Wilhelm et al.,
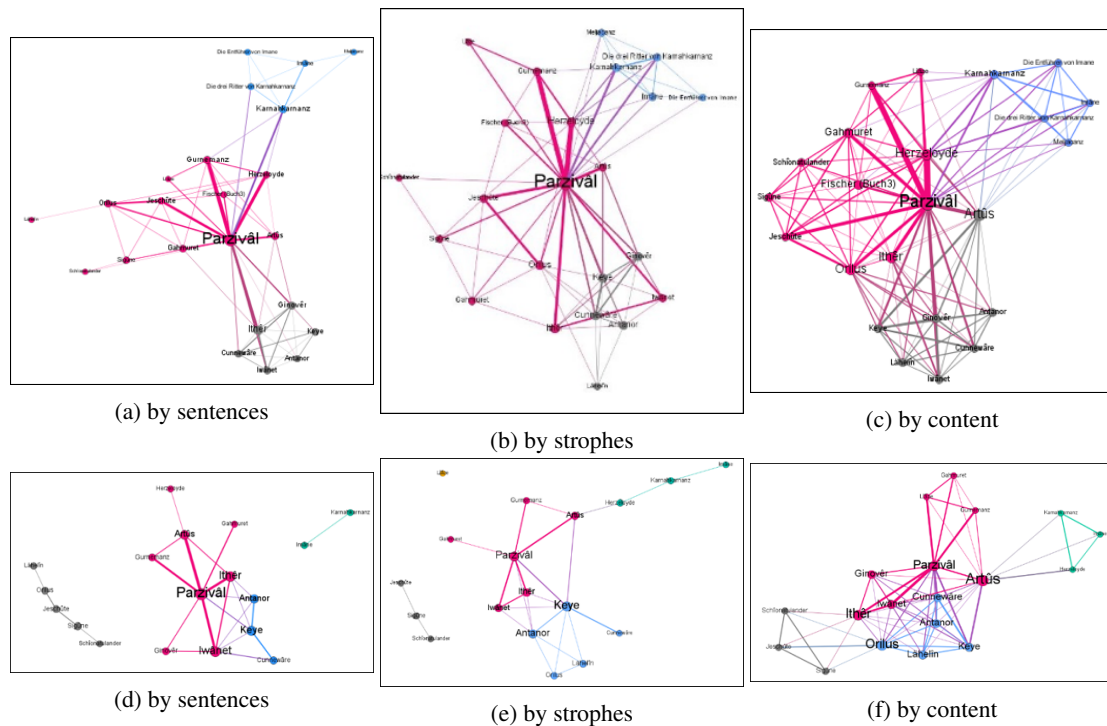
(a) by sentences

(b) by strophes

(c) by content

(d) by sentences

(e) by strophes

(f) by content

Figure 5: Different segmentation options. (a)-(c): With entity grounding; (d)-(f): Without grounding.
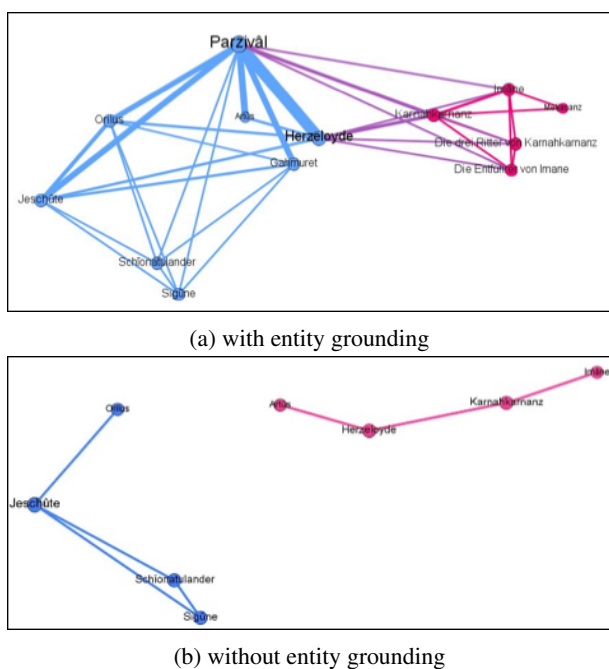

(a) with entity grounding

(b) without entity grounding

Figure 6: Parzivâls childhood, content-based segmentation with and without entity grounding. Without entity grounding (b) he is not present despite being a central character (a).

2013; Agarwal et al., 2014b). The strong structuring of such texts (scenes, acts) and clearly defined speakers make identifying co-occurring characters simple. The networks extracted from narrative texts by Elson et al. (2010) only include conversational relations: If two characters appear together in a dialogue, they are connected in a network. The work has been conducted on 19th century British novels and is based on named entities only, with a rule-based co-reference resolution. Agarwal et al. (2012) identify 'social events' in *Alice in Wonderland* and extract different types of networks (interaction- and observation-network) to investigate the roles/profiles of the characters. Using Mouse and Alice as an example they demonstrate the limitations of static networks and the need for dynamic networks that can display change over time. In later publications, they employ automatically created FrameNet frames as a basis to detect social events between named entities (Agarwal et al., 2014a).

Recently, several approaches for visualizing social networks have been introduced. For example, Oelke et al. (2013) analyze prose literature by using the pixel-based literature fingerprinting technique (Keim and Oelke, 2007). The approach visualizes relationships between characters and their evolution during the plot. A related technique is

64

used in FeatureLens (Don et al., 2007), which has been designed to support analysts in finding interesting text patterns and co-occurrences in texts.

There are quite a number of approaches (Vuillemot et al., 2009; Stasko et al., 2008) that provide node-link diagrams to represent social networks. In general, nodes represent entities and edges relations between them. An alternative method is a matrix-based representation which shows relationships among a large number of items where rows and columns represent the nodes of the network (Henry and Fekete, 2007). Both approaches have their drawbacks respective the readability of the structure of the overall network and also for detailed analysis (Ghoniem et al., 2005). Therefore, Henry et al. (2007) introduced a hybrid representation for social networks which combines the benefits of both approaches. It supports a set of interactions which allow users to flexibly change the representation to and from node-link and matrix forms of certain areas in the network.

9 Conclusions

We have presented an end-to-end environment for the extraction of co-occurrence networks based on criteria guided by literary research questions. This guidance not only informs the kinds of entities we are taking into account, but also the different ways of segmenting the text and even the fact that we are including non-named entities in the networks. The given examples in Section 7 demonstrate the influence of these choices on the networks of one and the same narrative text, thus it is important to make these decisions in close collaboration with the domain experts who will use the results. Ultimately, relying solely on named entities can lead to highly skewed impressions of the relative importance of characters in a text, to misleading interpretations of networks and thus, of literary texts. This becomes even more dangerous when large text collections are analyzed, for which a manual inspection is simply not possible. Allowing interactive exploration of aggregated data (networks) mitigates this issue: Domain experts interactively working with a network of a text become aware of such issues quickly. The early integration of scholarly experts even into primarily technical modules is therefore of utmost importance.

The collaboration with experts from different disciplines from Humanities and Social Sciences not only greatly benefits the conceptual development of, e.g., entity reference annotation guidelines: Difficult cases that appear frequently in one text type might appear rarely in another – Researchers working on the latter benefit from the collaboration because it would have taken much longer to come across rare cases. But in addition, this collaboration helps to ensure that technical and methodological developments are not too specialized for one particular text or text type. Too specialized software is relatively expensive to develop and will be outdated quickly. It also runs counter to the often purely methodological computer science goals of 'generic problem solving'. We therefore also concentrate on the fundamental methodological questions rather than on tool development.

We have focused here on one particular research question and corpus, but the above described workflow has been applied to narrative (modern and medieval) texts, theoretical philosophical texts (with the goal of establishing relations between philosophical networks) and parliamentary debates (with the goal of connecting political parties to political issues). We believe that it is worthwhile and feasible to search for common interests across multiple Humanities and Social Sciences disciplines and research questions. The identification of – at least structurally – common research questions allows to develop workflows that are supported by NLP and visualization methods that otherwise would just not pay off due to the development efforts.

References

Apoorv Agarwal, Sriramkumar Balasubramanian, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2014a. Frame semantic tree kernels for social network extraction from text. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, pages 211–219. http://www.aclweb.org/anthology/E14-1023.

Apoorv Agarwal, Sriramkumar Balasubramanian, Jiehan Zheng, and Sarthak Dash. 2014b. Parsing screenplays for extracting social networks from movies. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Association for Computational Linguistics, Gothenburg, Sweden, pages 50–58. http://www.aclweb.org/anthology/W14-0907.

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of alice in wonderland. In David K. Elson, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz, editors, *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, Montréal, Canada, pages 88–96. http://www.aclweb.org/anthology/W12-2513.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. In *Proceeding of International AAAI Conference on Weblogs and Social Media*.

Karl Bertau. 1983. Wolfram von Eschenbach. Versuch über die Verhaltenssemantik von Verwandten im "Parzival". In Karl Bertau, editor, *Neun Versuche über Subjektivität und Ursprünglichkeit in der Geschichte*, München, Germany, pages 190–240.

Steven Bethard, Philip Ogren, and Lee Becker. 2014. Cleartk 2.0: Design patterns for machine learning in uima. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Thomas Bögel, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris, and Jannik Strötgen. 2015. Gleiche Textdaten, unterschiedliche Erkenntnisziele? Zum Potential vermeintlich widersprüchlicher Zugänge zu Textanalyse. In *Proceedings of DHd*. Digital Humanities im deutschsprachigen Raum, Graz, Austria.

Didier Coste and John Pier. 2014. Narrative levels (revised version; uploaded 23 april 2014). In Peter Hühn, editor, *the living handbook of narratology*, Hamburg University Press. http://www.lhn.uni-hamburg.de/article/narrative-levels-revised-version-uploaded-23-april-2014.

Walter Delabar. 1990. *Erkantiu sippe unt hoch gesellschaft: Studien zur Funktion des Verwandtschaftsverbandes in Wolframs von Eschenbach "Parzival"*. Kümmerle, Göppingen, Germany.

Anthony Don, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman, and Catherine Plaisant. 2007. Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*. ACM, CIKM '07, pages 213–222. https://doi.org/10.1145/1321440.1321473.

Nora Echelmeyer, Nils Reiter, and Sarah Schulz. 2017. Ein PoS–Tagger für "das" Mittelhochdeutsche. In *Book of Abstracts of DHd 2017*. Bern, Switzerland, pages 141–147. https://doi.org/10.18419/opus-9023.

David K. Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre, editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, pages 138–147. http://www.aclweb.org/anthology/P10-1015.

Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. 2005. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization* 4(2):114–135.

Nathalie Henry and Jean-Daniel Fekete. 2007. Matlink: Enhanced matrix visualization for analyzing social networks. *Human-Computer Interaction–INTERACT 2007* pages 288–302.

Nathalie Henry, Jean-Daniel Fekete, and Michael J McGuffin. 2007. Nodetrix: a hybrid visualization of social networks. *IEEE transactions on visualization and computer graphics* 13(6):1302–1309.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 1148–1158.

D. Keim and D. Oelke. 2007. Literature fingerprinting: A new method for visual literary analysis. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. VAST '07, pages 115–122. https://doi.org/10.1109/VAST.2007.4389004.

Juri Lotman. 1977. *The Structure of the Artistic Text*. Oxon Publishing Ltd. Translated by Ronald Vroom.

Cerstin Mahlow, Kerstin Eckart, Jens Stegmann, Andre Blessing, Gregor Thiele, Markus Grtner, and Jonas Kuhn. 2014. Resources, Tools, and Applications at the CLARIN Center Stuttgart. In *Proceedings of the 12th Konferenz zur Verarbeitung natrlicher Sprache (KONVENS 2014)*. pages 11–21.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Franco Moretti. 2011. Network theory, plot analysis. Pamphlets of the Stanford Literary Lab 2, Stanford Literary Lab.

D. Oelke, D. Kokkinakis, and D. A. Keim. 2013. Fingerprint matrices: Uncovering the dynamics of social networks in prose literature. *Computer Graphics Forum* 32(3pt4):371–380. https://doi.org/10.1111/cgf.12124.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the conference on New Methods in Language Processing* 12.

Elisabeth Schmidt. 1986. *Familiengeschichten und Heilsmythologie. Die Verwandtschaftsstrukturen in den französischen und deutschen Gralromanen des 12. und 13. Jahrhunderts*. Tübingen, Germany.

John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization* 7(2):118–132. https://doi.org/10.1057/palgrave.ivs.9500180.

Rolf Sutter. 2003. *mit saelde ich gerbet han den gral: Genealogische Strukturanalyse zu Wolframs von Eschenbach "Parzival"*. Ph.D. thesis, Eberhard-Karls-Universität, Tübingen, Germany.

Peer Trilcke, Frank Fischer, Mathias Göbel, and Dario Kampkaspar. 2016. Theatre plays as 'small worlds'? network data on the history and typology of german drama. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pages 385–387. http://dh2016.adho.org/abstracts/360.

R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar. 2009. What's being said near "martha"? exploring name entities in literary text collections. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 2009*. VAST '09, pages 107–114. https://doi.org/10.1109/VAST.2009.5333248.

Thomas Wilhelm, Manuel Burghardt, and Christian Wolff. 2013. "To See or Not to See" - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In *Kultur und Informatik: Visual Worlds & Interactive Spaces*. Verlag Werner Hülsbusch, Glückstadt, Germany, pages 175–185.