Interpersonal Relations in Biographical Dictionaries. A Case Study.

Sophia Stotz*, **Valentina Stuß***, **Matthias Reinert**‡, **Maximilian Schrott**‡

*University of Paderborn
stotz,stuss@upb.de

‡Historische Kommission München
reinert,schrott@hk.badw.de

Abstract

Adopting the concept of "Local Grammars" (M. Gross), which were successfully applied in practice by (Geierhos, 2010) to biographical information extraction in English our project aims to detect, encode, and finally visualize relations between persons. Our corpus consists of the digitised biographical lexicon "Neue Deutsche Biographie (NDB)", roughly 21.000 biographies in 25 volumes in print since 1953. We developed local grammars and suitable dictionaries to describe interpersonal relations and applied them to the corpus with Unitex 3.1. The local grammars were designed to integrate existing TEI-XML structures in the corpus. Using the ability of local grammars in Unitex to act as transducers we were able to produce XML-tags and encode semantic information. Based on grammars for personal names and places we described interpersonal relations like *to study*, predecessors and successors as well as friends and circles. Afterwards we identified persons (as given in the authority file or index). Finally we displayed relations on our website in an interactive and dynamic way. Utilizing the Javascript library D3.js we represented named relations between identified individuals as ego centred network graphs.

Keywords: Local Grammar, Relation Extraction, Visualisation

1. Introduction

Biographical dictionaries comprise accounts of lives in a condensed, often abbreviated form. They list the most important events in an individual's life, as well as achievements and contacts with others. Events are expressed in predicates or sometimes idioms. Both carry one or more arguments, at least one of them representing an individual. This we call predicate-argument-structure (Geierhos, 2010, 7f.). Other statements about the influence of publications, innovations or intellectual impact brought about by the subject of biography are not taken into account.

A subset of these predicate-argument structures contain relational expressions: a second argument representing another person and the predicate - possibly accompanied by temporal or modal modifiers - representing the relation.

We consider academic teachers, friends, colleagues as *direct* interpersonal relations and relations constituted by peer-groups attending the same school and university or share the same profession and professional institution as *indirect* relations. Another dimension is hierarchy (patrons, teachers) vs equality (friends, colleagues) expressed in direct relations and hereditary (familiar background) vs transcendence (intellectual influence, schools of thought) in indirect relations. Obviously these relations are manifold and occur in modified forms therefore we have to normalise them. In this paper we will demonstrate the extraction of relations expressed by the verb *to study*.

In order to visualize relations between individuals we need to identify their names. We achieved this be applying simple matching techniques using indexes and scores and we undertook tests using topic similarities.

Finally we show the potential of relation extracting between identified individuals by visualizing them online using common force-directed graph libraries.

1.1. Method

In the huge field of information extraction we operate on *named entity recognition*, *named entity disambiguation* and *relation extraction*. But we restricted our efforts to detect personal names and a restricted set of relations. Interesting relations are accompanied with predicates containing further nameable entities as arguments. Our disambiguation aims primarily to align personal names with a knowledge base, namely an index of people, already qualified with profession, dates of birth and death and references to pages where they occur in the printed volumes.

In order to extract relations we applied methods described by Gross (1997), an approach called *local grammars*. Gross promoted the idea that idioms tended to be predominant over syntactic rules in language and demanded to examine large corpora in order to extract typical phrases. It is a combined *dictionaries and graph* approach, whereby graphs describe linguistic structures on a sub-sentence level. Linguistic structures or predicate-argument-structures are considered as verbal or noun phrases comprising entities carrying information. This reflects the influence of (Harris, 1974) who put the focus on argument structures.

Recent research into this approach has been undertaken on organization names in English by (Mallchok, 2005), on descriptors for humans in German by (Geierhos, 2007), on toponyms in German by (Nagel, 2008), on biographical facts in English by (Geierhos, 2010) and on biographical facts in French by (Maurel et al., 2011) and (Maurel and Friburger, 2013).

Just like these studies we rely on Unitex corpus-processor (Paumier, 2013). Unitex adopts the early efforts of W. A. Woods on applying graphs to linguistic phenomena (William A Woods, 1970). Already in 1980 he proposed to draft and apply subsequent graphs step by step (Woods, 1980). Among others, those ideas and the ability to call sub-graphs and morphological filters have been

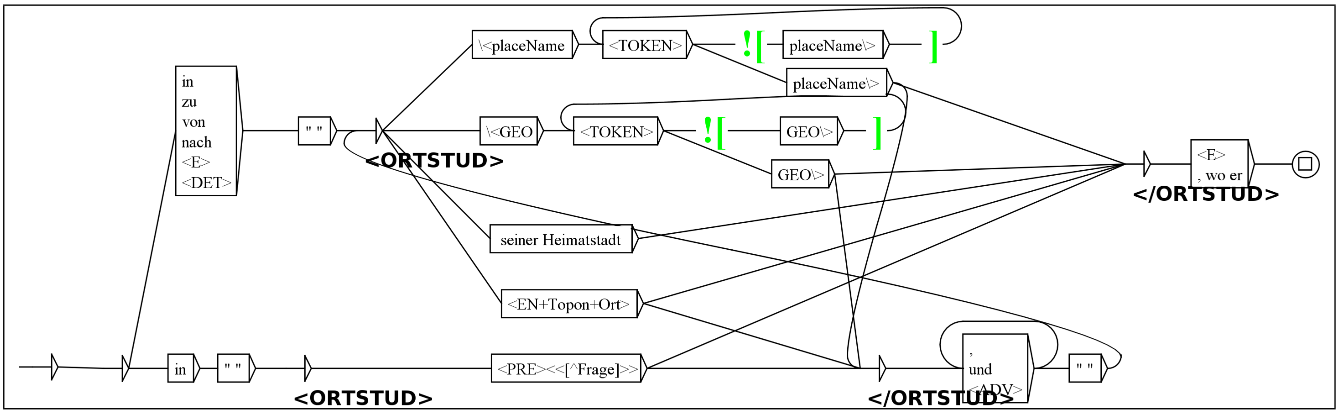

Figure 1: Example of a simple bootstrap graph detecting place names

implemented in Unitex.

We constructed local grammars in two steps. First we drafted preliminary graphs to describe and detect the specific vocabulary around interesting phrases. This was helpful to set up auxiliary dictionaries. Like the electronic dictionaries distributed with Unitex we use the DELA syntax (Dictionnaires Electroniques du LADL [Laboratoire d'Automatique Documentaire et Linguistique] (Paumier, 2013, 29)).

Secondly we had to cope with TEI-XML-markup already present in the corpus. We decided not to clean up this information because abbreviations had been tagged and facilitated the detection of sentence boundaries. This was achieved by using subsequent local grammars graphs, a mode of "cascade" available in Unitex and described by (Maurel and Friburger, 2013).

1.2. Dictionaries

Dictionaries are crucial for the adoption of local grammars. We used the general dictionary CISLEX for German developed at Center for Information and Language Processing (Centrum für Informations- und Sprachverarbeitung - CIS) Munich (Guenthner and Maier, 1994). CISLEX contains syntactic information about 150.000 entries encoded in DELA format (Paumier, 2013, 47ff).

In addition we extracted dictionaries of denominators for named entities from indices (list of names, professions) and an authority file (Gemeinsame Normdatei[1]. The Gemeinsame Normdatei (GND)[2] provided personal names and name parts, names for places, regions and organisations. We could derive dictionaries with roughly 1.9 mio surnames, 1.5 mio forenames and 9.3 mio full names for individuals as well as 1.36 mio entries for organisational names. Describing simple local grammars in a bootstrap manner (Gross, 1999) we could extract lists of entities for fields of study, institutions and place names (see 2). These bootstrapped dictionaries are specific to the given corpus and linguistically simply structured. They contain almost no syntactic information or declined forms but carry semantic information. We put together another 32.000 descriptors for occupation, 2.000 of them in declined form; 15.000 geographical names, 3.500 institutional names, mostly multi word chunks. A special vocabulary (1000 entries) covered disciplines and adjectives accompanying them; another individual school names who otherwise interfere with the relation *to study*.

```
Wetzlar,.EN+Topon+ORTSTUD
Wismar,.EN+Topon+ORTSTUD
Witzenhausen,.EN+Topon+ORTSTUD
Włocławek,.EN+Topon+ORTSTUD
Worpswede,.EN+Topon+ORTSTUD
Zerbst,.EN+Topon+ORTSTUD
```

Figure 2: Example of a simple dictionary entries, denoting place names with lexical category EN (named entity), semantic categories Topon and ORTSTUD

Bootstrapping dictionaries from the corpus gives the opportunity to revise and optimize the dictionaries.

1.3. The corpus Neue Deutsche Biographie

Our corpus is provided online at www.deutsche-biographie.de. The website consists of the digitised biographical dictionaries "New German Biography" (NDB). The dictionary recently reached the letter T (Tecklenborg) and has published 25 volumes in print since 1953. Available online are about 21.000 articles of the first 24 volumes (A-Stader). These biographical articles have been selected in a peer review process by the editorial team under guidance of the editor in chief. They are composed of a headline, a short genealogy, the account of life and further technical paragraphs on awards, works, secondary literature and depictions. All articles are signed by an author. Articles are written in modern German (pre 2006 style) in full sentences but show many abbreviations of frequent words (adjectives, nouns) and the lemma itself (surname or personal name of the subject of the biography).

In addition to the NDB its precursor "Allgemeine Deutsche Biographie" finished 1912 in 55 volumes plus an index volume enlarges the amount of articles available in the website by 27.000. These older articles are written in an outdated orthography and style and have not been taken into account.

We heavily used auxiliary databases listing the individuals

[1] http://www.dnb.de/gnd

[2] http://www.dnb.de/lds GND as Linked Data Service in March 2013.

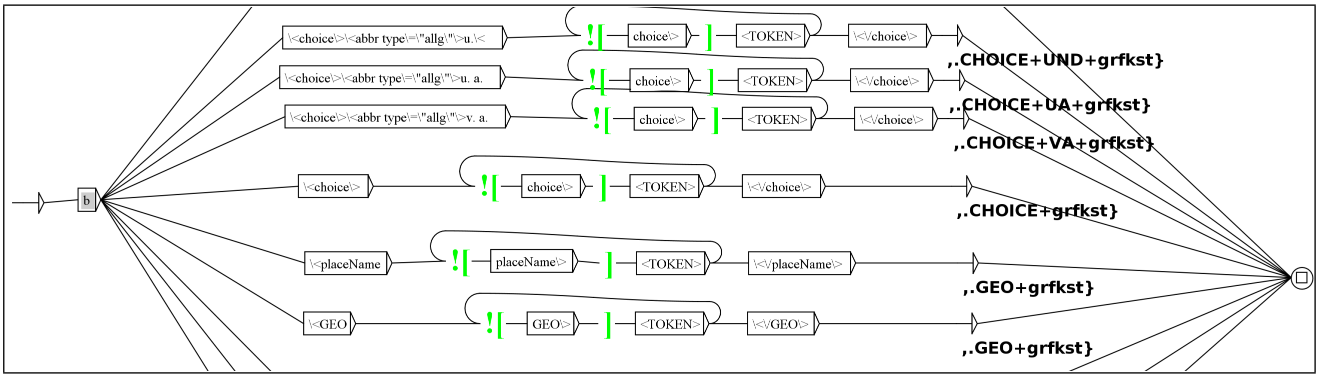

Figure 3: Masking pre-tagged text and entities, using TOKEN-loops with ![ ]-negative context

mentioned in the text along with profession or position in life, their birth and death dates and references to the printed volumes. All in all the core data base consists of 92.000 individuals and several hundred families. Almost each entry has been aligned with or added to the bibliographic authority file Gemeinsame Normdatei (GND).

The articles were digitised and typographically tagged by an exernal firm and afterwards structurally tagged in XML according to the TEI guidelines (Text Encoding Initiative, 2009) in the project. For reasons like human read-ability, easier proof-reading, and tagging of pre-existing XML, we decided neither to follow the stand-off mark-up approach nor the habit of computational linguistics of working on plain text but to keep up the whole tagging, re-use it on occasion and add further tags in line.

2. A local grammar for the verb *to study*

In German, there are several ways to express someone has studied. The verb *studieren* as well as *ein Studium beginnen, aufnehmen, absolvieren, beenden* or *(sich) an der Universtität einschreiben/ Vorlesungen (an der Universität) belegen, besuchen*, *jemanden hören* each sets a certain focus to the activity and determines possible arguments. We restrict our grammar to the verb *to study* and its forms. Our analysis of the corpus resulted in the following structure:

The predicate-argument structure of *to study* is accompanied by several types of entities, like institution, university (*Universität Wien, Akademie der bildenden Künste*), place, discipline (*Physik, Kulturwissenschaften*, teacher (*bei Virchow und Naunyn, <persName>...</persName>*), time (*1813, 4 Semester, ab Juli 1876*) and student colleagues.

Several adverbs and modifiers occur in the phrases as well as uncertainty markers and negative phrases (*studierte wahrscheinlich*).

The position of arguments/entities in the sentence is not fixed, they may occur after the predicate, before or on both sides. One position usually expressed with a pronoun or an abbreviated lemma denotes the subject who has studied. As the corpus contains biographies on individuals we assume that the subject of *to study* and of the biography are the same.

2.1. Masking pre-tagged text and entities

The corpus comprises lots of abbreviations. We masked them using a special grammar. The masking started a short sequence of grammars (cascade). Almost all grammars were acting as transducers - they wrote output back into the recognized chunks of text. In this way new XML tags were introduced to mark extracted entities in each step.

1. masking pre-existing tags
2. masking interfering statements on education
3. *to study* with arguments in pre- and post-position
4. *to study* with arguments in pre-position
5. *to study* with arguments in post-position (common case)
6. deal with the noun *study*.

Figure 5: Schema of the *cascade* (Paumier, 2013, 243ff). Each graph is applied repeatedly until no new match is found and in merge-mode e.g. merging outputs with the detected sequences in the corpus

There is a {multi word expression,.lexical type|mask(+lexical type|mask)*}–notation processed by the Unitex system (Paumier, 2013, 44-46). As shown in fig. 3 Unitex recognizes such kind of meta-syntax in order to treat multi-word expressions on the one hand and assign lexico-semantic types (e.g. CHOICE+UA in fig. 3) to text units on the other hand (Geierhos et al., 2011, 49).

The mask applies to abbreviations already identified and tagged, certain abbreviations are tagged with semantic types. This applies also to personal names which were similarly identified and tagged with local grammars.

The schema of Cassys allows to apply a list of graphs and to run through each graph once or until no further match is detected. By default each graph is applied as a transducer, its output can be given in replace- or merge-mode (Paumier, 2013, 84).

2.2. Recognizing Entities and Relations

By using dictionaries (see 1.2.) and masking graphs we created a sequence of graphs for our target relation. The schema of the cascade starts by masking pre-existing XML-Tags and goes on detecting and encoding composed entities. The Local grammar for the verb *to study* is split up by positional differences.

The main graph (s. fig. 4) is composed of paths and subgraphs (Paumier, 2013, 99). Each path describes a linguistic possibility and for certain arguments the graph de-

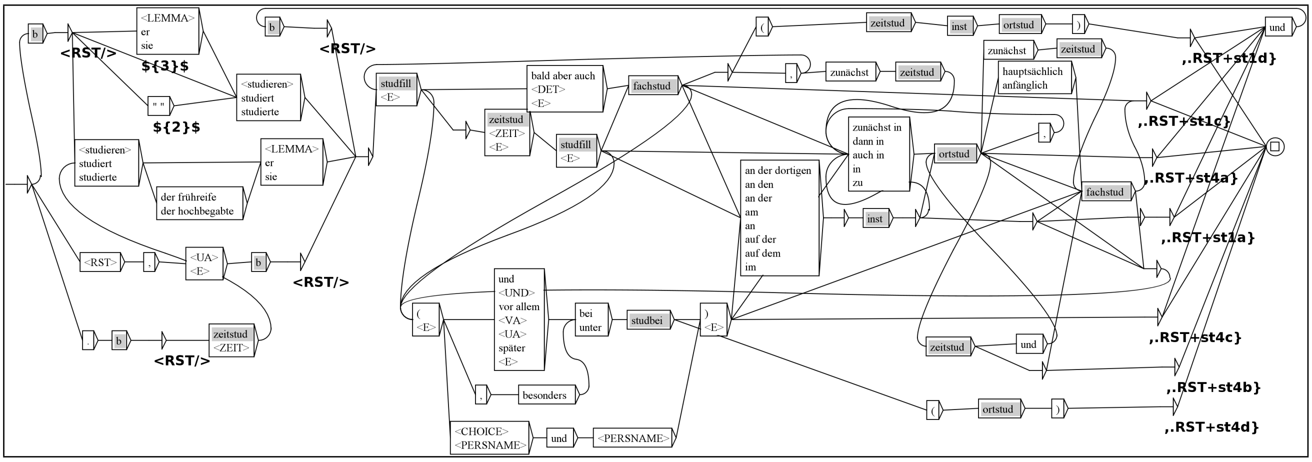

Figure 4: A local grammar describing the post positioned arguments of *studieren/to study*, boxes after prepositions branch into subgraphs

scend into subgraphs describing the structure of the argument more detailed. Obviously the arguments are governed by prepositions; *in* is followed by place names, *bei* precedes teachers. The only object argument - the discipline(s) or field(s) of study - directly governed by *studieren/to study* is rare in a university context.

The graph (s. fig. 4) is applied as a transducer (Paumier, 2013, 243ff). In the figure outputs are displayed in boldface letters, each attached to the a box matching possible type, strings or ε on a certain position in the input string. They produce well formed XML which can be processed afterwards.

2.3. Results of Relation Extraction

The LGs were modeled on a subset of the whole corpus (vols. 2–4,12–14,22–24) which covered the wide range of years. Hence the results have been measured twice: once on the model set and again on the test set comprising all other volumes (1,5–11,15–21).

In order to test the results we extracted lines containing the string *studier* which represents the infinitive and present stem (*studier[en]*), past and perfect stems (*studiert*) but not related nouns and composita of (*Studie*, *Studium*).

The matches and errors were counted as follows:

entities of	*found*	*not found*	*false named*
to study	true positive	fault (Recall)	fault (Precision)
not to study	false positive (Precision)	true negative	false (Precision)

Table 1: Assertion of errors to precision and recall

We calculated the common F-measure:

$$\text{F-Measure} = \frac{2 \times precision \times recall}{precision + recall}$$

We achieved a high rate of precision as intended. The small number of errors resulted in an erroneous path in the grammar which could be deactivated. Another remedy for these errors would be another graph applied within the cascade or on top of the result in replacement mode like (Nagel, 2008, 233, see “Antigrammatiken”) has shown.

The recall can be increased by additional grammars which can be applied on top of the result. Missing entities due to early exiting graphs which are generally the consequence of missing entries in the dictionary.

Part of Corpus	model	unseen
$N_{studier...}$	3378	5245
$N_{studier...}$ in sample	148	261
Total nr. of entities in sample	580	1028
Nr. of entities found by LG	427	601
Errors	4	17
Recall	73,62%	58,46%
Precision	99,31%	98,35%
F-Score	84,56	73,33

Table 2: The caption of the table

3. Disambiguating Personal Names

Detecting relations in predicate-arguments structures resulted in named entities as typed sets of strings (literals). The relation extraction already differentiated between personal names, university names, place names and disciplines. One of the next steps was to disambiguate the identity of personal names by aligning them with knowledge bases. We identified “literals” as individuals in our registry of names and the authority file.

To illustrate the problem the single word “Goethe” could refer to the famous writer and public servant Johann Wolfgang von Goethe (1832), but possibly to 5 other articles on persons named “Goethe” in NDB and ADB. The authority file GND provides 129 hits for a person called “Goethe”.

The first approach matched features from index-entries (given name, surname, year of birth, year of death, page and region [headline, biography or genealogy]) and occurrences of names. By simply adding points together for each matching criteria we related the sum to the number of criteria. Matching years scored double, matching initials scored

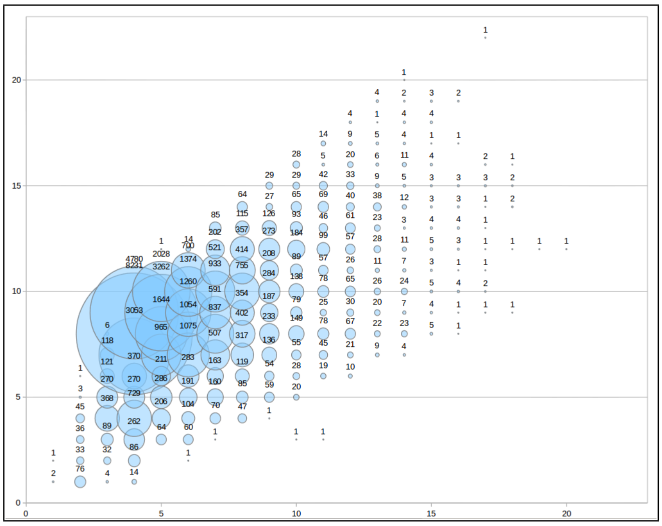

Figure 6: Distribution of numbers of score-baseline pairs, x represents the baseline of features in the string (name parts, dates, pages), y represents the scores achieved by matching and c(x,y) represents the count of matches for the given score and baseline drawn as a circle

half points. This resulted in about 55.000 matches in a distribution given in fig. 6.

We examined a sample for each pair in order to detect a threshold of certainty. Names without dates were generally under-determined and have been dropped. In genealogies we assumed everyone shared the surname of the subject of the biography. But plain given names sometimes do refer to another family and the implicit assumption of a common surname led to failures. In headlines and in the biographical description the matching of names bearing at least 3 correct features (f.i. a name and the page and a date, 2 dates and a part of the name, 2 name parts and a date) yielded to reasonable correct results.

3.1. Results of Disambiguation

The simple scoring approach allowed us to match most of the articles and a substantial amount of persons in the biographical descriptions and to a smaller degree in genealogies. Named entities for personal names without dates – very frequent in the early volumes and the preceding ADB – could not be processed. We tested topic modelling and topic similarity measures (cosine similarity) but were not successful due to the lack of biographies for all potentially interesting individuals. Some biographies were not elaborate enough to provide a decent vector of topics.

4. Visualising Relations Online

The visualization of the extracted relations data was realized with D3.js.[3] This javascript library is all about transforming data into graphics, as its name "Data-Driven-Documents" implies. We decided to use this library because of some key advantages. It is modern technology, which creates its graphics user side without the need for any plugin except javascript. It draws into a HTML "div"-container and integrates the different elements of the visualization into the Domain Object Model of the website, making them styleable with CSS and debuggable with standard in-browser developer tools. D3 is also quite flexible: it can process a variety of data formats, as long as the data is structured like an array and then transform it into any kind of visualization, either simple or complex. While potent D3 it should be noted though, that D3 can be quite hard to implement, due to a poor documentation and some unintuitive behaviours.

4.1. Designing the graph

When looking for a way to visualize the interpersonal relations we experimented with displaying the persons on the outside of the perimeter of a circle, with the edges between them running through the inside of the circle itself. We hoped that this would provide a good way to display large numbers of persons and their relations within a delimited space. In the end however we found this this approach lacking in comprehensibility and difficult to implement. Instead we took inspiration from the Social Network and Archive (SNAC) project at the Institute for Advanced Technology in the Humanities at the University of Virginia.[4] Their prototype visualization displays the relations of a person in a classic network graph, in which the persons are nodes with edges between them representing their relations. But while the SNAC visualization arranges the nodes along concentric circles, we decided to use a force-directed graph.

4.2. Force-directed graphs

In a force-directed graph, the layout is determined automatically and dynamically by an algorithm, that calculates simulated forces between the nodes. This algorithm is provided by the D3 library. Normally nodes repel each other and would just spread out evenly across the canvas. Edges, which have a certain length and flexibility, similar to a real-life bungie cord, counteract this repulsion and tie the connected nodes together. These two forces should ideally arrange the graph in a clearly laid out way. Unrelated nodes are kept at a distance from each other, while related ones group closely together, forming clusters that indicate their high level of interconnectedness at first glance.

4.3. The ego-centred network graph

Our graph is centered around one person - the root. When the visualisation is started, only the immediate relations of the root are grouped radially around it. But the graph can be expanded further, like in the visualisation of SNAC. By clicking on the node of a person the user can append their relations to the graph (if they have any within our database). This not only works with the nodes that are directly linked to the root, but with any node in the graph. This way the user can jump from relation to relation, go deep into the graph and discover extensive interpersonal networks.

The nodes can also be collapsed again by clicking on them a second time. This removes all nodes and edges from the graph that are connected to the root only through the clicked node. And by clicking on the root node the graph can always be brought back into its original state, with only the root itself and its immediate relations visible. The deletion

[3] `http://d3js.org`

[4] `http://socialarchive.iath.virginia.edu/`

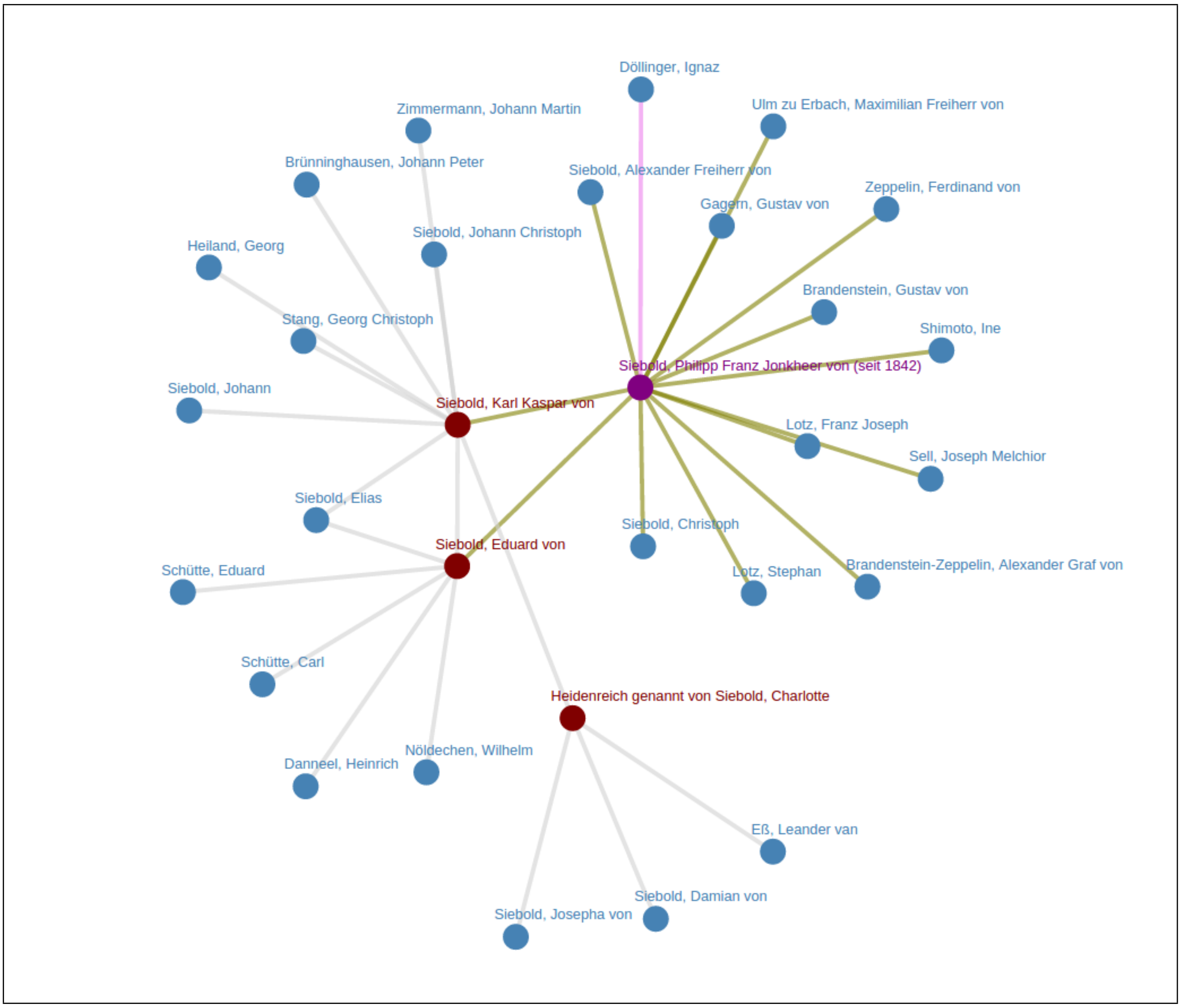

Figure 7: The relations of Philipp Jonkheer von Siebold and three relatives, manually expanded.[5]

of links and nodes has to be done recursively to account for deep trees of relations, which might spawn from a single node. Before deleting a node the program checks if it has "children" of its own. If so these "children" are then checked for deletion or further recursion.

For the recursion to work, the edges in the graph have to be directed. Even though this is not visible in the visualisation, every link has a source and a target node. To prevent the forming of circles within the graph, which could lead to unwanted behaviour during the recursion, links pointing back towards the direction of the root have to be avoided. For this reason edges that connect two already linked nodes but in the opposite direction are quietly dropped. Likewise other links that would close a circle are flipped around by the program to point away from the root. While this manipulation and discarding of data is not ideal, we do not consider it to be very problematic and simply present all relations as mutual to the user.

4.4. Typed Relations

A new feature currently tested out in closed beta is the typing of relations. Currently we mainly distinguish three types of links. The differentiation is based on the part of the article, from where the relations was extracted. If it's from the genealogy the link is classed as "Familie" (family). "Leben" (life) on the other hand means, that the relation was found in the biography itself. And finally "Literatur" (literature) links come from the bibliographical appendix to the article. The edges in the graph are color-coded according to their type and can be removed from or added to the graph by the user. The next step is to add the relations extracted with the more sophisticated method of computational linguistics described earlier in this paper. These link types are based on the actual nature of relationship rather than their position in our text. We already have added the type "Lehrer/Schüler" (teacher/students) to our beta version and plan to add further types, once they can be extracted with enough confidence. Right now relations like "Lehrer/Schüler" exists separate from the three other types.

[5] Access to test version `http://data.deutsche-biographie.de/beta/lib4/Projects/dtBio/relations/?id=sfz80197&version=ndb` on request.

But as they model a different kind of relationship, we plan to revise the data model, so that a link can have multiple types.

4.5. Further Plans

We also plan to migrate the relation data to a graph database. Right now the data for the ego-graph is produced from the same Apache Solr search index as the rest of our website. While this works sufficiently well for our current implementation, we want to expand the functionality of our visualisation. With the integrated advanced support for graphs in databases like Neo4J we hope to allow for new functions like the automatic computation of the shortest relationship between any two persons, while at the same time reducing the problems with circles and backlinks.

5. Outcomes and Discussion

The laborious description of predicate-argument structures finally payed off. We could retrieve structured information as type named entities and have been able to adopt our grammars to similar unseen corpora with a fair result. Our approach on disambiguation is supported for individual mentions comprising names and dates. Names missing dates and other named entities bearing fewer features were unable to identify.

6. Acknowledgements

The work is funded by the Deutsche Forschungsgemeinschaft (DFG) to establish a biographical information system online (2012-15).

7. References

Michaela Geierhos, Jean-Leon Bouraoui, and Patrick Watrin. 2011. Towards multilingual biographical event extraction - initial thoughts on the design of a new annotation scheme. In *Multilingual Resources, Multilingual Applications. hg.v. Hanna Hedeland, Thomas Schmidt, Kai Wörner*, page 4.

Michaela Geierhos. 2007. *Grammatik der Menschenbezeichner in biographischen Kontexten*. Arbeiten zur Informations- und Sprachverarbeitung. Band 2.

Michaela Geierhos. 2010. *BiographIE - Klassifikation und Extraktion karrierespezifischer Informationen*. Linguistic Resources for Natural Language Processing 05. Lincom.

Maurice Gross. 1997. The construction of local grammars. In E. Roche and Y. Schabès, editors, *Finite-State Language Processing*, pages 329–354.

Maurice Gross. 1999. A bootstrap method for constructing local grammars. In Neda Bokan, editor, *Proceedings of the Symposium on Contemporary Mathematics*, pages 229–250.

Franz Guenthner and Petra Maier. 1994. *Das CISLEX Wörterbuchsystem.*

Zellig S. Harris. 1974. *Lecture Notes on English Transformational Grammar Université de Paris VIII, 1974 (Transl. 1976 by Maurice Gross: Notes du course de syntaxe, Paris: Editions du Seuil.).*

Friederike Mallchok. 2005. *Automatic Recognition of Organization Names in English Business News*. Studien zur Informations- und Sprachverarbeitung Band 9, zugleich Dissertation 2004.

Denis Maurel and Nathalie Friburger. 2013. Utilisation avancée des cascades de graphes sous unitex (cassys). In *2nd Unitex/GramLab Workshop. 10-11 octobre 2013, Université Paris Est-Marne-la-Vallée*.

Denis Maurel, Nathalie Friburger, J.-Y. Antoine, I. Eshkol-Taravella, and D. Nouvel. 2011. Cascades autour de la reconnaissance des entités nommées. In *TAL*, pages 69–96.

Sebastian Nagel. 2008. *Lokale Grammatiken zur Beschreibung von lokativen Sätzen und ihre Anwendung im Information Retrieval.*

Sébastian Paumier. 2013. *Unitex 3.1 (Beta). User Manual.*

Text Encoding Initiative, editor. 2009. *TEI: P5 Guidelines, version 1.5.*

William A Woods. 1970. Transition network grammars for natural language analysis. *Communications of the ACM*, 13(10):591–606.

William A. Woods. 1980. Cascaded ATN grammars. *American Journal of Computational Linguistics*, 6:1–12.