The GDR Through the Eyes of the Stasi
Data Mining on the Secret Reports of the State Security Service of the former German Democratic Republic

Christoph Kuras[1], Thomas Efer[1], Christian Adam[2] and Gerhard Heyer[1]
[1]*Natural Language Processing Group, Department of Computer Science, University of Leipzig, Leipzig, Germany*
[2]*Agency of the Federal Comissioner for the Records of the State Security Service of the former German Democratic Republic, Berlin, Germany*

Keywords: GDR, Digitization, Humanities, Visualization, Named Entity Recognition.

Abstract: The conjunction of NLP and the humanities has been gaining importance over the last years. As part of this development more and more historical documents are getting digitized and can be used as an input for established NLP methods. In this paper we present a corpus of texts from reports of the Ministry of State Security of the former GDR. Although written in a distinctive kind of sublanguage, we show that traditional NLP can be applied with satisfying results. We use these results as a basis for providing new ways of presentation and exploration of the data which then can be accessed by a wide spectrum of users.

1 INTRODUCTION

As reaction to the June crisis in 1953 the Ministry of State Security in the GDR (the Stasi) set up information groups. Their task was to collect information and to write “mood reports”. On the basis of these reports the Leadership of the GDR wanted to be able to react fast on future uprisings. Since 1953 the ”Central Evaluation and Information Group” (ZAIG) of the Ministry of State Security was compiling information for the party and state leadership. These secret reports were completed in different forms and with varying frequency for more than 36 years – and are today a contemporary source of great historical value. They reveal the Stasi’s specific view of the GDR: they contain references to real and perceived oppositional conduct as well as to economic and supply problems. They also include statistics on currency exchanges, ”illegal” emigration and border violations. Trivial information is presented alongside discussions of both minor and major ”difficulties” created by the effort to institute and maintain SED rule and to establish ”actually existing socialism”. In order to provide an overall impression of the different types of sources and their importance to scholarly research, the publication started with one year from each of the four decades of the GDR under the title: ”The GDR through the Eyes of the Stasi. The Secret Reports to the SED Leadership” (dt.: Die DDR im Blick der Stasi. Die geheimen Berichte an die SED-Führung), (Münkel, 09ff). At the completion of the pilot phase, the other years will be published in irregular order. The education and research department of the BStU[1] decided to publish these documents in a hybrid edition. A year after the publication of a volume, the respective year will be presented online at *www.ddr-im-blick.de*. By now the volumes of 1953, 1961, 1976, 1977 and 1988 are available online (open access). The documents (between 800 and 1500 pages per year) are presented in chronological order. The texts are transcripts of the original files, processed in Microsoft Word and for the final publication converted in XML and published cross media. The digital version of the documents can be accessed by a full-text search, while no further indices are provided, since their creation would cause too much expenses in the manual editing workflow. The edition includes annotations, explanations of abbreviations and an introduction to each year. Each document contains a list of recipients (including personal names, e.g. of ministers, or departments). Due to the Stasi Records Act personal data shall not be published unless they are obvious, they concern employees or beneficiaries of the State Security Service. That’s why some personal data has to be blackened

[1]Agency of the Federal Commissioner for the Records of the State Security Service of the former German Democratic Republic

Kuras C., Efer T., Adam C. and Heyer G..
The GDR Through the Eyes of the Stasi - Data Mining on the Secret Reports of the State Security Service of the former German Democratic Republic.
DOI: 10.5220/0005136703600365
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 360-365
ISBN: 978-989-758-048-2

out in the edition. At the end of the publication process 37 volumes will be available, approximately 50 000 pages of documents, offering an inside-view of more than three decades of the East German History as seen through the eyes of the Stasi. The cooperation with the Natural Language Processing Group, Department of Computer Science at the University of Leipzig is an attempt to get more, innovative and long term approaches to these important data.

2 THE COOPERATION

The perspective of the BStU is mainly historical and political aiming for the reappraisal of the past whereas the NLP group provides methods and technical resources to process text automatically and to extract information from it. This is a result of different focuses. We see a lack of automation and use of information systems concerning the editorial process on the part of the BStU - a lack of historical expertise and a constant need of data available to apply classic automatic language processing methods on the part of the NLP Group. We believe, that by combining the competences of each partner we are able to create added value on both sides.

The benefit of this cooperation has effects in two directions: on the one hand there are improvements of the existing workflow. The introduction of automation and support systems during the editorial process can lead to a more efficient workflow and can also reduce errors and redundancies. On the other hand there are applications that would not be possible without the use of NLP methods, e.g. when the effort is too high to be done with human resources only. Furthermore the NLP group is able to develop new methods and to solve new problems on real world data. Therefore we believe, a cooperation is neccessary for advancing in the reappraising of the past as well as in NLP. In the following we present first results of processing these data with advanced NLP methods. In particular, applying methods for latent semantic indexing when analysing recipients yields highly interesting and promising results.

3 DATA AND STRUCTURE

As a result from the digitization process we get digital documents which does not mean that they are structured in a way that is best for NLP. The original documents are stored in an XML-like format which is optimized for printing thus containing also layout information. This format can be considered semi-structured in the common XML-sense. The tags used do not support the semantics of the document structure but the semantics of layout and printing. Retrieving information from those documents is associated with high search costs. In addition to the documents' contents there are metadata stored for each document. Those include the date, the subject, a list of recipients of that document and other information. To create a format that is less complicated concerning querying and flexible to use for a wider range of applications we parsed the original documents and designed a new XML schema to which we transfered them. Transferring the documents to a database seems natural, however we found designing a new XML schema to be the first step in that direction since it makes further processing easier. The whole collection is categorized by year. Table 1 lists the available XML-formatted documents we extracted from the original format for each year.

Table 1: Currently available Stasi documents by year.

Year	#Documents
1953	198
1961	260
1976	320
1977	337
1988	279
Total	1394

The transformed documents are structured in a way that the document contents, possible attachments and the different metadata elements have their own tags. New annotations can be added easily, for example POS tags or named entities which then can be used as features for advanced NLP methods. Some problems still remain after the transformation. The list of recipients is a plain string list which can contain information annotated during the editorial process. This makes it difficult to split the list into individual recipients which is needed for tasks like automatic indexation. We extracted the recipients from the list by applying regular expressions on the string which works in most cases. However, it was not possible to use exactly the same regular expressions on the documents of every volume so a more refined approach is required to improve the quality of the extracted data. Despite existing challenges, the transformation was a first step to structure the data. Still, further work is necessary.

Without previous attempts to digitally analyse the ZAIG reports, only little is known about the quantity, coverage and relations of certain topics and about the feasibility of employing certain NLP methods. This can be illustrated by Figure 1. It shows the frequency of the words "Sozialismus" (eng. socialism)

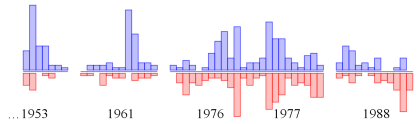

Figure 1: Sentence occurrences of ”Freedom” (blue) and ”Socialism” (red) from 1953 to 1988.

and ”Freiheit” (eng. freedom) from 1953 to 1988. We can clearly see patterns which need to be interpreted. Analyses like this would hardly be possible without the use of information systems. Therefore it is expedient to begin with a series of explorative and transdisciplinary approaches. Visual Analytics (Keim et al., 2010) provides a methodological framework for finding abstractions from the numerical statistics, for aiding the interpretation of its results and for the semi-automatic extraction and refinement of knowledge in an interdisciplinary context.

4 CLASSIC NLP

The transformation process described above focused on the generation of the raw material which can be handled by classic NLP applications. This section focuses on the further processing. As the result of the transformation described above we are now able to process the content of the documents themselves. We process all documents in the same way defined by our chain of tools. At first, each document is segmented into sentences. After that we tokenize the sentences[2]. After tokenization it seems natural to do POS tagging as it is a good basis for more complex tasks, such as parsing.

4.1 POS-Tagging and NER

For POS-tagging we use the Stanford POS-Tagger described in (Toutanova and Manning, 2000) and (Toutanova et al., 2003). We use a pre-trained german model that comes with the tagger. The accuracy is estimated by manually reviewing a sample of 250 words from randomly chosen sentences. It can be estimated between 94% and 96%. For named entity recognition we used the Stanford named entity tagger described in (J. R. Finkel and Manning, 2005). To estimate the accuracy and to see whether problems arise we also evaluated the result on a random sample of 250 named entities. The accuracy can be estimated between 82% and 88% which is higher than we expected as the data contains many specific phrasings and the model was trained on completely different data. Table 2 shows examples of correctly recognized named entities. Even MISC-entities like ”SED-Genosse” (Member of the SED party) are recognized. These are of particular interest concerning our data as they could be used to extract social networks.

[2]These two steps are done by tools developed by the NLP group in Leipzig.

Table 2: NER Evaluation.

Sequence	Tag
Großhennersdorf	I-LOC
Potsdamer Platz	I-LOC
VEB Blechbearbeitung Berlin	I-ORG
MfS	I-ORG
Walter Ulbricht	I-PER
Bischof Schönherr	I-PER
SED-Genosse	I-MISC
deutsch	I-MISC

It is characteristic for the data to contain a large number of location and person names since most documents report about specific events or persons. This can also be a challenge when applying NER. The next section describes some problems we could already identify.

4.2 Specific Problems of Named Entity Recognition

For the task of named entity tagging it is even more difficult to validate the result than in POS tagging because expert knowledge may be required. In some cases it is not obvious, even for human readers, to identify the correct tag. Common problems when tagging the documents are listed below.

- different levels of granularity
  - ”Staatsgrenze West” (engl. ”western country border”) vs. ”Berlin” as locations.
  - street names vs. city names
- incomplete sequences due to specific phrasings
  - only ”VEB” is tagged instead of ”VEB Blechbearbeitung Berlin”
- anonymization
  - anonymous entities (e.g. ”[Name 1]”) are not being recognized yet
- same names for different persons
  - some of the persons have identical last names (e.g. relatives)
  - ”Herbert Krolikowski” vs. ”Werner Krolikowski” (his brother), both were active politicians in overlapping time periods

These challenges could be faced by adding extra knowledge before or after the tagging or by training

the tagger on a different tagset. However, it is very important to map the sequences to the right persons as errors could make the data unusable. Despite that, the recognition of entities works reasonably well and even other tasks that make use of the tagger are possible. The next section describes an approach for tagging another kind of entities.

4.2.1 Tagging Politically Motivated Phrases

Using the Stanford Named Entity Tagger it is possible extract other information than named entities. By manually creating training data we were able to exploit the tagger to identify common phrases which are often found in contexts where persons are accused of acting against the regime. Some reports contain special phrasings like ”provokativ” (eng. ”provoking”) or ”Hetze” (eng. ”agitation”) which appear in a number of different combinations. A person can act provoking as well as a pamphlet may be provoking, resulting in ”defamatory publications”. To find many of those combinations automatically we trained the Stanford NE-Tagger on manually annotated training data taken from about 30 documents. Table 3 shows the most frequent combinations we extracted from documents of the year 1976 automatically.

Table 3: Top 5 political motivated phrases for the year 1976.

Phrase
ungesetzlichen Grenzübertritts (engl. ”illegal” border crossing)
Hetze (engl. agitation)
Einmischung in innere Angelegenheiten (engl. intervention in internal affairs)
provokatorischen Handlungen (engl. provoking actions)
politisch-ideologischen Diversion (engl. political-ideological diversion)

This approach enables us to access political opinions directly in the documents. Such phrasings can then be brought into context with other named entities and recipients. Figure 2 is an example of an interactive graph showing relations between political phrasings, documents, recipients and different persons.

This graph representation can answer many different questions at the same time: Which documents contain most phrasings of this kind? Which persons are involved most in documents with reference to such phrasings? Is it possible that there may be a protagonist-antagonist-relationship between specific person entities and individual recipients? - just to name a few. The interactive viewing of data can also be considered inspiring as new research questions may arise when exploring the graph. After tagging the

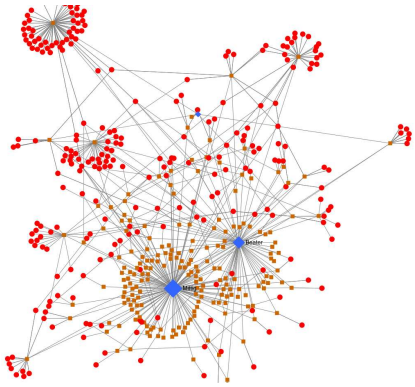

Figure 2: Interactive graph of political phrasings, documents and recipients.

named entities it may be obvious to ask which topics the documents deal with and which documents belong to the same topic. For this task we applied clustering which is described in the next section.

4.3 Clustering Similar Documents

Especially after the long time period passed since these documents were written it is impossible for contemporay readers to instantly overview their topics. One way of bringing more structure to the collection is finding similar documents. We decided to chose a simple approach first and clustered the documents using the k-means algorithm. There are at least three possible feature sets describing a single document. One possibility is to use the recipients of the document to cluster the documents. Another approach could be to use the set of named entities that occur in a document or to combine both approaches. In our case, each document is solely described by a

Table 4: K-means clustering of documents (1976) by recipients.

Cluster	Documents	possible description
1	52	currency exchange
2	108	fleeing, border violation
3	84	church
4	32	border traffic
5	44	catastrophes

binary-valued vector of recipients where each value represents whether the recipient has received the document or not. Table 4 shows the clustering by recipient vectors with manually added descriptions. By this mere example we can already get an overview of the data and even identify different spheres of competence which then could be assigned to a group of recipients. There is, of course, a challenge of choosing the right number of clusters which has to be considered in further work. As an alternative approach a

graph clustering algorithm like the chinese whispers algorithm proposed in (Biemann, 2006) can be used to achieve comparable results, in a more visual way. Focusing more on the recipients, we consider it useful to analyse their characteristics and the network structures among these. As a first step we are able examine recipients that often appear together. The next section describes a first approach to recipient analysis.

5 RECIPIENT ANALYSIS

Since every document has a list of recipients assigned to it we can apply different techniques to analyse the characteristics of individual recipients or to extract networks. For example, we determined which recipients occur together remarkably frequent.

5.1 Frequent Recipient Sets

Therefore, we consider each recipient an item of a basket we want to analyse to get frequent item sets. For this task we use the FP-Growth algorithm proposed in (Han et al., 2000). The top five frequent recipient sets with a minimal support of 0.2 for the year 1976 can be seen in Table 5.

Table 5: Frequent sets of recipients for the year 1976 with min. support 0.2.

Rank	Set
1	HA XX, Schorm
2	Mittig, Schorm
3	Mittig, Mittag
4	HA XX, Verner
5	Mittig, Verner

This is a way of detecting cohesion between different recipients and to reveal underlying structures which can be expected or even new to historians. Table 5 shows that there are HA (Hauptabteilung, engl. main department) in the list. Further processing is needed to split the recipients into organizational units and individual persons. Then, the algorithm can be applied to the group of persons and the group of organizational units individually. Again, this method could also be applied to the named entities alone or the sets of recipients and named entities together.
All these approaches target the fact that there are properties describing the recipients and that individual recipients have different properties. The following section proposes another approach to extract those properties.

5.2 Extracting Recipient Properties

Regarding the document-recipient-matrix which contains information about which recipient received which documents, we assume that there is knowledge about both recipients and documents encoded in the matrix. The extraction of this information can be accomplished by decomposing the document-recipient matrix. The matrix is binary-valued and has the form

$$D_{m,n} = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m,1} & d_{m,2} & \cdots & d_{m,n} \end{pmatrix} \quad (1)$$

where n are the recipients and m the documents.
This matrix is factorized into a factor containing information about the addressees (matrix A), a factor containing information about the reports (matrix R) and the logistic function. This can be seen in Equation 2 where σ is the logistic function which constrains values to lie in the range [0,1].

$$D = \sigma(A \cdot R) \quad (2)$$

We define a cost function (Equation 3) which is then optimized. By optimization we minimize the difference between data and prediction. The cost function is the sum of differences for every element of the prediction and data matrix. In our case we set the number of features to *5*. The logistic function is chosen since the input is a binary-valued matrix.

$$Cost = \sum (D - \sigma(A \cdot R)) \quad (3)$$

This is similar to other matrix factorization techniques like non-negative matrix factorization or principal components analysis (Lee and Seung, 1999; Tipping and Bishop, 1999).

We found that the factorized matrices contain information about specific properties of individual recipients and documents. Given these properties, we can sort the matrix along the properties axis to find members with the highest rank in that dimension. The semantics of those properties can be determined by a manual analysis. However, this task may require a high level of expert knowledge about the persons reviewed. Looking at the top-ranked recipients of one property it is possible to get an idea of what is described by that property. Figure 3 shows a cluster resulting from the application of k-means to the recipients' property vectors. There are negative values for some recipients in dimension *5*. Looking at the individual recipients, "KGB Berlin-Karlshorst" (representation of the KGB in Berlin) has the lowest value. Taking into account that "Ungarn" (Hungary), "Geheimdienste Polen" (security services of poland)

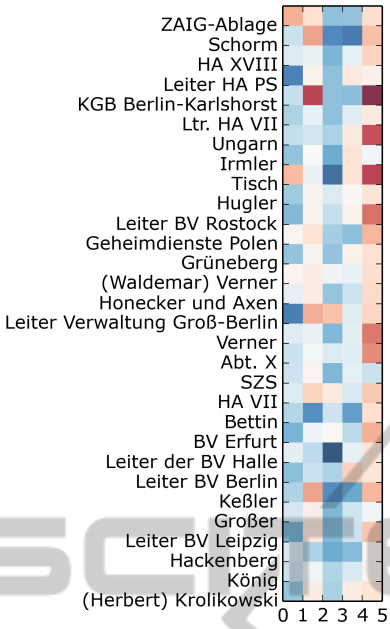

Figure 3: Visualization of a cluster of the addressee-matrix. Positive values are blue, negative values are red.

and ”Abt. X” (department for international alliances) also appear in that cluster, we may suppose this dimension describes the intensity of some kind of ”international involvement”. We believe these methods are a possibility to find latent structure in the data and to summarize the roles of recipients and document contents.

6 CONCLUSION

We believe that NLP methods yield the potential to play an ever important role in the edition, enrichment and publication of historical documents for scientific and public audiences. As shown in this publication, there are a lot of previously untapped possibilities to process the texts and metadata entries in order to create alternative perspectives and novel navigational means. With all that, our work is meant to complement existing digitization and edition projects. Several institutions are now beginning to evaluate and incorporate (in most cases rather basic but still very helpful) NLP methods to enrich their digital editions. One can for example look at the thematically matching portal ”DDR-Presse”[3], where a viewer for OCR-transformed newspaper articles from the GDR era is augmented with links to norm data, editorial articles and the like, when certain keywords appear in the text.

[3]http://zefys.staatsbibliothek-berlin.de/ddr-presse

Lastly, it becomes apparent, that through visual and interactive means, historians are enabled to verify the NLP methods’ quality while understanding both, the hidden structures in the data and the influences of methodological choices on the results. That leads to valuable feedback for the computer scientists and furthermore allows to define new (and alter existing) research questions while reviewing intermediate results. As an outlook, those visual interfaces also constitute an engaging mode of presentation for the general public since these techniques can remove entry barriers, provide navigational guidance and improve the overall user experience. We are looking forward to conjointly develop such NLP powered systems and to make tangible the nature of the ZAIG reports for a broad audience.

REFERENCES

Biemann, C. (2006). Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06*, New York, USA.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, volume 29, pages 1–12. ACM.

J. R. Finkel, T. G. and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.

Keim, D. A., Kohlhammer, J., Ellis, G., and Mansmann, F. (2010). *Mastering The Information Age-Solving Problems with Visual Analytics*. Florian Mansmann.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Münkel, D., editor (2009ff). *Die DDR im Blick der Stasi 1953–1989. Die geheimen Berichte an die SED-Führung*. www.ddr-im-blick.de.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*.

Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.