



EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action

Call: H2020-ICT-2018-1

Call topic: ICT-29-2018 A multilingual Next generation Internet

Project start: 1 January 2019

Project duration: 36 months

D4.3: Initial cross-lingual news summarisation and visualisation technology (T4.2)

Executive summary

The task of cross-lingual news summarisation and visualisation (Task T4.2) addresses the problem of creating informative condensed textual and visual representations of a larger text or corpus. The proposed solutions must also address the difficulties of creating these summaries using cross-lingual content in the low-resourced languages of the EMBEDDIA project. In this deliverable the initial technology developed for creating cross-lingual textual summaries is presented along with an evaluation where it is shown that the developed technologies outperform the baselines results on the tested low resourced language summarisation tasks. Next, two visualisation techniques are presented: one based on summarizing the content of a corpus along temporal and topic axis; and a second which uses concordance visualisation and a novel metadata exploration tool to generate summaries of corpus partitions.

Partner in charge: UEDIN

Project co-funded by the European Commission within Horizon 2020
Dissemination Level

PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	—
RE	Restricted to a group specified by the Consortium (including the Commission Services)	—
CO	Confidential, only for members of the Consortium (including the Commission Services)	—



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

Deliverable Information

Document administrative information	
Project acronym:	EMBEDDIA
Project number:	825153
Deliverable number:	D4.3
Deliverable full title:	Initial cross-lingual news summarisation and visualisation technology
Deliverable short title:	Cross-lingual news summarisation and visualisation
Document identifier:	EMBEDDIA-D43-CrosslingualNewsSummarisationAndVisualisation-T42-submitted
Lead partner short name:	UEDIN
Report version:	submitted
Report submission date:	30/06/2020
Dissemination level:	PU
Nature:	R = Report
Lead author(s):	Shane Sheehan (UEDIN), Elvys Linhares Pontes (ULR)
Co-author(s):	Saturnino Luz (UEDIN), Senja Pollak (JSI)
Status:	_ draft, _ final, <u>x</u> submitted

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
15/05/2020	v0.1	Elvys Linhares Pontes (ULR)	First version of CLTS.
20/05/2020	v0.2	Shane Sheehan (UEDIN)	Visualisation section.
23/05/2020	v0.3	Shane Sheehan (UEDIN)	Initial draft for internal review.
25/05/2020	v0.4	S Luz (UEDIN)	Revised text; fixed some tex/maths issues.
26/05/2020	v0.5	Shane Sheehan (UEDIN)	Added associated outputs; Revised text.
26/05/2020	v1.0	Shane Sheehan (UEDIN)	Submitted for internal review.
02/06/2020	v1.1	Hannu Toivonen (UH)	Internal review.
05/06/2020	v1.2	Vid Podpečan (JSI)	Internal review.
06/06/2020	v1.3	Elvys Linhares Pontes (ULR)	Changes based on reviews.
07/06/2020	v1.4	Shane Sheehan (UEDIN)	Changes based on reviews.
10/06/2020	v2.0	Shane Sheehan (UEDIN)	Ready for quality control.
11/06/2020	v2.1	Nada Lavrač (JSI)	Quality control.
23/06/2020	v2.2	Elvys Linhares Pontes (ULR)	Implemented comments from quality control.
26/06/2020	v2.3	Shane Sheehan (UEDIN)	Final improvements.
27/06/2020	final	S Luz (UEDIN)	Final version of the report.
30/06/2020	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

1. Introduction.....	5
2. Cross-Language Text Summarisation.....	5
2.1 An overview of cross-language text summarisation.....	7
2.1.1 Machine Translation Quality.....	7
2.1.2 Joint Analysis of Source and Target Languages.....	9
2.2 Our cross-language text summarisation approach.....	11
2.2.1 Preprocessing.....	11
2.2.2 CoRank Method.....	12
2.2.3 Multi-Sentence Compression.....	12
2.2.4 Summary Generation.....	13
2.3 Experimental evaluation.....	13
2.3.1 Datasets.....	14
2.3.2 Evaluation.....	14
2.3.3 Results and analysis.....	14
3. Corpus Summary Visualisation.....	16
3.1 Metafacet.....	17
3.2 Corpus Comparison.....	21
4. Current visualization work partly related to EMBEDDIA.....	22
4.1 Temporal Mosaic Visualisation.....	22
4.1.1 Temporal Transcript Summarisation.....	24
4.1.2 Implementation.....	25
4.1.3 Contextual Linking.....	27
4.1.4 Temporal News Visualisation.....	29
4.2 Graph-based visualisation.....	30
5. Associated Outputs.....	31
6. Conclusions and Future Work.....	32
Appendix A: A Multilingual Study of Multi-Sentence Compression using Word Vertex-Labeled Graphs and Integer Linear Programming.....	35
Appendix B: Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge.....	55
Appendix C: Text Visualization for the Support of Lexicography-Based Scholarly Work.....	75
Appendix D: TeMoCo: A Visualization Tool for Temporal Analysis of Multi-Party Dialogues in Clin- ical Settings.....	107
Appendix E: TeMoCo-Doc: A visualization for supporting temporal and contextual analysis of dia- logues and associated documents.....	113

Appendix F: The NetViz terminology visualization tool and the use cases in karstology domain modeling	116
Appendix G: Communities of Related Terms in a Karst Terminology Co-occurrence Network.....	123
Appendix H: An Example of Cross-Language Text Summarization.....	140

List of abbreviations

CR	Compression Ratio
CLTS	Cross-Language Text Summarisation
ILP	Integer Linear Programming
MSC	Multi-Sentence Compression
NN	Neural Networks
PAS	Predicate-Argument Structures
TS	Text Summarisation
NER	Named Entity Recognition
NEL	Named Entity Linking

1 Introduction

This deliverable reports on the initial results achieved in cross-lingual news summarisation and visualisation performed in Task T4.2 of the EMBEDDIA project. This task, which began in M7, is described in the EMBEDDIA Description of Action (DoA) as follows:

This task will develop textual and visual language-independent multi-document news summarisation. The textual summaries will be created using contrastive approaches on multiple documents at once, focused on the core topics while maximizing the coverage in the number of topics in the sub-corpus. Visualisation will allow this generated summary to be investigated interactively. Sections of the summary will interactively display the sources which contributed to its generation and information related to the process of summary generation will be overloaded onto these source documents. Visual summaries of the document collections using the outputs of other tasks (Tasks T4.3 and T4.1) will provide additional, visual, analytical tools for investigating cross lingual trends.

In this M18 report, our research on cross-language text summarisation and visual summarisation is presented. Currently these two research directions have been pursued independently and in parallel. We plan to combine these outputs to produce a summarisation method based on combined textual and visual summarisation techniques.

The text summarisation work presented in Section 2 focuses on the generation of cross-lingual summaries using a novel method which groups similar sentences. The technique is designed to reduce redundancy and improve the informativeness of cross-lingual summaries. An evaluation of the technique was found to outperform baseline results for the tested low-resourced languages.

The visualisation work, presented in Section 3 makes use of our prior work on concordance visualisation to create summaries, which can be explored using a metadata based technique developed as part of the EMBEDDIA project.

In Section 4 we present two pieces of work initially developed for domains not related to the project but with obvious potential for EMBEDDIA. The first is a temporal visualisation where salient textual content related to a collection of topics is presented as a mosaic. This technique can be used to provide an overview of the content and patterns of change in a corpus or a document. The application of this technique to transcripts of dialogue audio and the linking of this visualisation with a textual summary are also presented. The second approach presented is a graph based visualisation of topics and related terminology which can serve as a visual summary of the topic structure of a corpus.

The work presented in this report resulted in five publications that are listed in Section 5 and included in Appendices A–E.

2 Cross-Language Text Summarisation

Technological advances have improved and increased the speed of world communication through the transmission of videos, images and audios. Nowadays, most books and newspapers have digital and/or audio versions while the popularisation of social networks (such as Facebook, Twitter, YouTube, among others) and news Web sites have enabled a great increase in the amount of data trafficked over the Internet about diverse subjects. Every day, a considerable amount of information is published in various sites, e.g., comments, photos, videos and audio in different languages. In this way, an event is quickly disseminated on the Web by different news medias from around the world and under various formats (audio, image, text and video).

The readers, besides not having the time to go through this amount of information, are not interested in all the proposed subjects and generally select the content of their interest. It is worth mentioning that much of the information is personal, such as comments on daily life, personal photos and videos posted on social networks and blogs. Thus, some of this information is not of interest to everybody. For this

reason, newspapers, movies, books, magazines, websites and blogs have headlines, summaries and/or synopses of the topics covered. From the headlines of a newspaper, the readers identify the subject of news articles and can choose which article to read in its entirety. This process is similar for books and movies with their synopses and descriptions on websites and blogs. In this way, the readers can quickly identify the subject of their interest and then continue the reading. These synopses, descriptions and headlines are different types of summaries that highlight the main information of books and articles at different levels of granularity.

In general, a summary is composed of the main idea presented in the original document in a short and objective way. The summary can be produced in different lengths depending on the desired purpose (Saggion & Lapalme, 2002; Moens, Angheluta, Mitra, & Jing, 2004). For example, news headlines from newspapers and sites have few words to catch the reader's attention and convey the key idea of the news. However, longer texts require a more comprehensive summary for the reader to understand the subject of the text as is the case of books, which require longer summaries to get their general idea conveyed. One way to measure the length of a summary is its number of words or characters. Another possible way is the compression ratio (CR), which is responsible for defining the size of the summary in relation to its original text. CR is defined by the length of the summary over the length of the document (Equation 1).

$$CR = \frac{|\text{summary}|}{|\text{document}|} \quad (1)$$

Summary can be considered as a kind of compression process that removes non-relevant content and maintains key text information. The lower the CR value, the shorter the summary of an analyzed text. This reduction, up to a certain level, improves the quality of a summary because it highlights the main information. However, the exaggerated reduction of a document causes the loss of relevant information and damages its comprehensibility.

An issue in news summarisation is the language of messages, given that a lot of news are available in languages that the readers do not know or have little knowledge of. The enormous amount of information prevents it from being summarized and translated by humans. Besides the problem of summarising all these documents, the translation of documents into several languages requires polyglot translators. This process requires a lot of time and resources when there are a huge amount of data to be analyzed.

Cross-Language Text Summarisation (CLTS) aims to generate a summary of a document where the summary language differs from the document language. More precisely, CLTS consists of analyzing a document in a language source to get its meaning and then generate a short, informative and correct summary of this document in a target language (Linhares Pontes, 2018). This process can be split into two main processes: text summarisation and text translation. Two simple possible procedures are: summarise the document and, then, translate the summary; or translate the document to the target language and, then, summarise the translated document.

The work presented in this deliverable proposes the initial version of our cross-language text summarisation approach to summarise news in low-resourced languages (Croatian, Estonian, Finnish, and Slovenian) to English. We analyse the similarity between sentences in the source and target languages to estimate the relevance of sentences. Then, we use the Linhares Pontes et al.'s (Linhares Pontes, Huet, Torres-Moreno, & Linhares, 2018, 2020) approach to group and compress similar sentences in order to reduce the redundancy and improve the informativeness of cross-lingual summaries. As this deliverable demonstrates, our proposition outperforms the baselines for all low-resourced languages.

The rest of the section is organized as follows: Section 2.1 makes a brief overview of the most recent and available CLTS state-of-the-art approaches. Section 2.2 details our approach to generate cross-lingual summaries by compressing similar sentences. Finally, the results achieved on the extended version of the MultiLing 2011 dataset and the analysis of cross-lingual summaries are reported in Section 2.3.

2.1 An overview of cross-language text summarisation

Cross-Language Text Summarization (CLTS) consists in analyzing a document in a source language to get its meaning and, then, generate a short, informative and correct summary of this document in a target language. Summary generation can be:

- Extractive methods estimate the relevance of sentences in a document to generate a summary by concatenating the most relevant sentences.
- Compressive methods compress sentences to reduce the length of sentences and to preserve only the main information. Then, they generate summaries by concatenating the most relevant sentences and compressions of a document.
- Finally, abstractive methods analyse a document and generate a summary with new sentences that contain the meaning of the source documents.

The first studies in cross-language document summarisation analyzed the information in only one language (Leuski et al., 2003; Orasan & Chiorean, 2008). Two typical CLTS schemes are the early and the late translations (Figure 1). The first scheme first translates the source documents into the target language, then it summarises the translated documents using only information of the translated sentences. The late translation scheme does the reverse: it first summarises the documents using abstractive, compressive or extractive methods, then it translates the summary into the target language.

Leuski et al. (2003) proposed an early translation method to generate English headlines for Hindi documents. Orasan and Chiorean (2008) implemented the late translation approach; they produced summaries with the maximal marginal relevance method from Romanian news articles and then automatically translated the summaries into English.

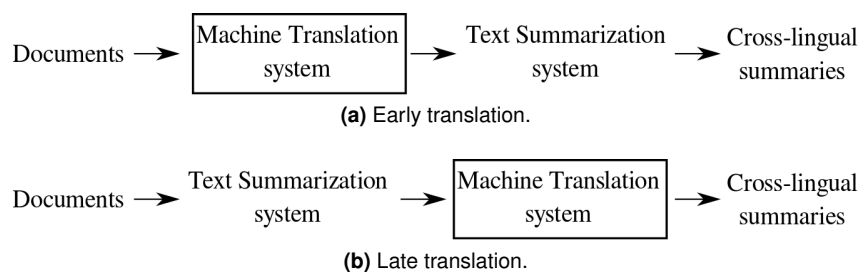


Figure 1: Early and late translations for CLTS.

Recent methods have improved the quality of cross-language summarisation using a translation quality score (Wan, Li, & Xiao, 2010; Boudin, Huet, & Torres-Moreno, 2011; Yao, Wan, & Xiao, 2015) and the information of the documents in the source and the target languages (Wan, 2011; Zhang, Zhou, & Zong, 2016). These two kinds of methods are described in the next two subsections.

2.1.1 Machine Translation Quality

Machine translation evaluation aims to assess the correctness and quality of the translation. Usually, a human reference translation is provided, and various methods and metrics have been developed for comparing the system-translated text and the human reference text.

Another possibility is the use of automatic methods to estimate translation quality (see for example the quality estimation shared task of the WMT conference (Bojar et al., 2017)). The translation quality of a sentence can be estimated at word-level, phrase-level and sentence-level. The estimation at word-level aims to detect errors for each token in machine translation outputs by deciding if a token is correct in the translation. An incorrect word can cause several errors in the translation, especially in its local context. The estimation at phrase-level is similar to the word-level, i.e. the estimation verifies if a phrase is correct

in the translation. Finally, the estimation of translation quality at sentence-level aims to generate scores for the translations according to post-editing effort, i.e. the percentage of needed edits, post-editing time, and so on.

Wan et al. (2010) trained a support vector machine regression method to predict the translation quality of a pair of English-Chinese sentences from basic features (such as sentence length, sub-sentence number, percentage of nouns and adjectives) and parse features (such as depth, number of noun phrases and verbal phrases in the parse tree) to generate English-to-Chinese CLTS. They used 1,736 pairs of English-Chinese sentences (English sentences were translated automatically by Google Translate) and computed translation quality scores in a range from 1 to 5 (1 means “very bad” and 5 corresponds to “excellent”). The translation quality and informativeness scores were linearly combined to select the English sentences with both a high translation quality and a high informativeness:

$$score(s_i) = (1 - \lambda) \cdot InfoScore(s_i) + \lambda \cdot TransScore(s_i) \quad (2)$$

where $InfoScore(s_i)$ and $TransScore(s_i)$ are the informativeness score and translation quality prediction of sentence s_i , respectively; and $\lambda \in [0, 1]$ is a parameter controlling the influence of the two factors. Finally, they translated the English summary to form the Chinese summary.

Similarly to (Wan et al., 2010), (Boudin et al., 2011) used an support vector regression to predict the translation quality score based on the automatic NIST metrics as an indicator of quality. They automatically translated English documents into French using Google Translate, then they analyzed some features (sentence length, number of punctuation marks, perplexities of source and target sentences using different language models, etc.) to estimate the translation quality of a sentence. They incorporated the translation quality score in the PageRank algorithm (Brin & Page, 1998) to calculate the relevance of sentences based on the similarity between the sentences and the translation quality scores to perform English-to-French cross-language summarization (Equations 3–5).

$$p(v_i) = (1 - d) + d \times \sum_{v_j \in pred(v_i)} \frac{score(s_i, s_j)}{\sum_{v_k \in succ(v_i)} score(s_k, s_i)} p(v_i) \quad (3)$$

$$score(s_i, s_j) = similarity(s_i, s_j) \times prediction(s_i) \quad (4)$$

$$similarity(s_i, s_j) = \frac{\sum_{w \in s_i, s_j} freq(w, s_i) + freq(w, s_j)}{\log(|s_i|) + \log(|s_j|)} \quad (5)$$

where d is the damping factor, $prediction(s)$ is the translation quality score of sentence s , $freq(w, s)$ is the frequency of the word w in sentence s , $pred(v_i)$ and $succ(v_i)$ are the predecessor and successor vertices of vertex v_i .

Inspired by the phrase-based translation models, (Yao et al., 2015) proposed a phrase-based model to simultaneously perform sentence scoring, extraction and compression. They designed a scoring scheme for the CLTS task based on a submodular term of compressed sentences and a bounded distortion penalty term to estimate the quality of the translation. Their summary scoring ($F(sum)$) measure was defined over a summary sum as:

$$F(sum) = \sum_{p \in sum} \sum_{i=1}^{count(p, sum)} d^{i-1} g(p) + \sum_{s \in sum} bg(s) + \eta \sum_{s \in sum} dist(pbd(s)) \quad (6)$$

where $g(p)$ is the score of phrase p (defined by the frequency of p in the document), $bg(s)$ is the bigram score of sentence s , $pbd(s)$ is the phrase-based derivation of sentence s and $dist(pbd(s))$ is the distortion penalty term based on the reordering probability of the phrase-based translation models. Finally, d is a constant damping factor to penalize repeated occurrences of the same phrases, $count(p, sum)$ is the number of occurrences of phrase p in the summary sum and η is the distortion parameter for penalizing the distance between neighboring phrases in the derivation.

2.1.2 Joint Analysis of Source and Target Languages

Wan (2011) proposed to leverage both the information in the source and in the target language for cross-language summarization. In particular, he introduced two graph-based summarization methods (SimFusion and CoRank) for using both the English-side and Chinese-side information in the task of English-to-Chinese cross-language summarization. The first method linearly fuses the English-side and Chinese-side similarities for measuring Chinese sentence similarity. In a nutshell, this method adapts the PageRank algorithm to calculate the relevance of sentences, where the weight arcs are obtained by the linear combination of the cosine similarity¹ of pairs of sentences for each language:

$$relevance(s_i^{cn}) = \mu \sum_{j \in D, j \neq i} relevance(s_j^{cn}) \cdot \tilde{C}_{ji}^{cn} + \frac{1 - \mu}{n} \quad (7)$$

$$C_{ij}^{cn} = \lambda \cdot sim_{cosine}(s_i^{cn}, s_j^{cn}) + (1 - \lambda) \cdot sim_{cosine}(s_i^{en}, s_j^{en}) \quad (8)$$

where s_i^{cn} and s_i^{en} represent the sentence i of a document D in Chinese and in English, respectively, μ is a damping factor, n is the number of sentences in the document and $\lambda \in [0, 1]$ is a parameter to control the relative contributions of the two similarity values. C^{cn} is normalized to \tilde{C}^{cn} to make the sum of each row equal to 1.

The CoRank method adopts a co-ranking algorithm to simultaneously rank both English and Chinese sentences by incorporating mutual influences between them (Figure 2). It considers a sentence as relevant if this sentence in both languages is heavily linked with other sentences in each language separately (source-source and target-target language similarities) and between languages (source-target language similarity) (Equations 9-13).

$$\mathbf{u} = \alpha \cdot (\tilde{\mathbf{M}}^{cn})^T \mathbf{u} + \beta \cdot (\tilde{\mathbf{M}}^{encn})^T \mathbf{v} \quad (9)$$

$$\mathbf{v} = \alpha \cdot (\tilde{\mathbf{M}}^{en})^T \mathbf{v} + \beta \cdot (\tilde{\mathbf{M}}^{encn})^T \mathbf{u} \quad (10)$$

$$M_{ij}^{en} = \begin{cases} \cosine(s_i^{en}, s_j^{en}), & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

$$M_{ij}^{cn} = \begin{cases} \cosine(s_i^{cn}, s_j^{cn}), & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

$$M_{ij}^{en,cn} = \sqrt{\cosine(s_i^{cn}, s_j^{cn}) \times \cosine(s_i^{en}, s_j^{en})} \quad (13)$$

where \mathbf{M}^{en} and \mathbf{M}^{cn} are normalized to $\tilde{\mathbf{M}}^{en}$ and $\tilde{\mathbf{M}}^{cn}$, respectively, to make the sum of each row equal to 1. \mathbf{u} and \mathbf{v} denote the saliency scores of the Chinese and English sentences, respectively; α and β specify the relative contributions to the final saliency scores from the information in the same language and the information in the other language, with $\alpha + \beta = 1$.

Recently, (Wan, Luo, Sun, Huang, & Yao, n.d.) carried out the cross-language document summarization task by extraction and compression through the ranking of multiple summaries in the target language. They analyzed many candidate summaries in order to produce a high-quality summary for every kind of documents. These candidate summaries were generated using multiple text summarization and machine translation methods, e.g., bilingual submodular function, multiple machine translations and multiple sentence compressions. Their method used a top-K ensemble ranking based on features at several levels and perspectives (word-level, sentence-level, summary-level, readability-related and source-side features) that characterized the quality of a candidate summary.

¹The cosine similarity between two vectors u and v associated with two sentences is defined by $\frac{u \cdot v}{||u|| ||v||}$ in the $[0, 1]$ range.

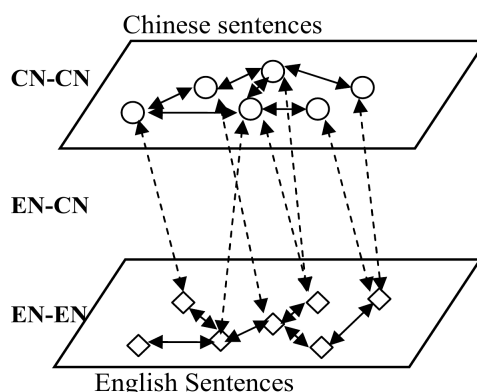


Figure 2: Sentence relationships of CoRank method.

By contrast with (Wan et al., n.d.) who generated extractive and compressive CLTS, (Zhang et al., 2016) analyzed Predicate-Argument Structures (PAS) to obtain an abstractive English-to-Chinese CLTS (Figure 3). They built a pool of bilingual concepts and facts represented by the bilingual elements of the source-side PAS and their target-side counterparts from the alignment between source texts and Google Translate translations. They used word alignment, lexical translation probability and 3-gram language model to measure the quality and the fluency of the Chinese translation, and the CoRank algorithm (Wan, 2011) to measure the relevance of the facts and concepts in both languages. Finally, summaries were produced by fusing bilingual PAS elements with an Integer Linear Programming (ILP) algorithm to maximize the saliency and the translation quality of the PAS elements. Their ILP model used the pool of bilingual concepts (facts) and their scores to generate summary sentences composed of a concept and at least one core fact.

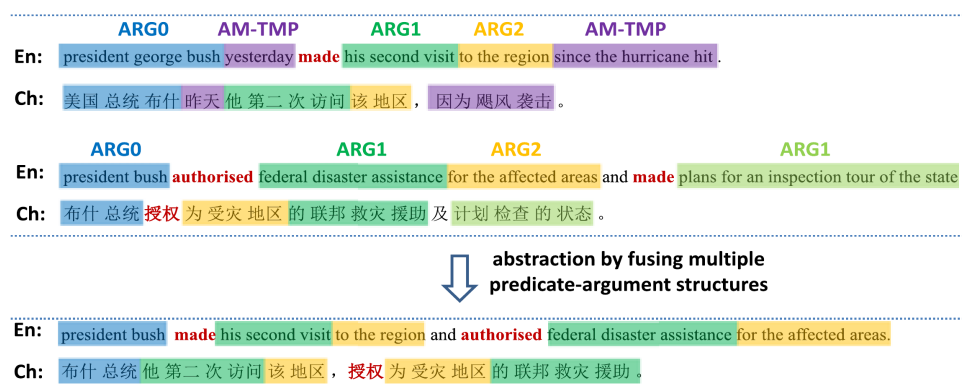


Figure 3: An example of CLTS based on PAS fusing.

Linhares Pontes et al. (Linhares Pontes, Huet, Torres-Moreno, & Linhares, 2018, 2020) developed a compressive approach to generate cross-lingual summaries. First, they analyse the similarity of sentences in the source and target languages to compute the relevance of sentences (CoRank method). Then, they use two sentence compression approaches to reduce the redundancy and improve the informativeness of cross-lingual summaries. Their first approach analyses clusters of similar sentences and compresses them by using a multi-sentence compression (MSC) method. The second approach is a Neural Network model that compress single sentences by deleting non-relevant words in the sentences.

Among the existing approaches for summarization, abstractive text Text Summarisation (TS) methods have a greater capacity to generate summaries more similar to the human abstracts. However, this kind of summarization demands large datasets available in a language to train neural network (NN) models.

On the contrary, extractive summarization approaches do not require specific resources to generate summaries; nevertheless, these extracted sentences may contain redundant and/or non-relevant information, thus reducing the informativeness of summaries. Finally, some compressive methods only need a few resources in several languages to generate summaries. Therefore, our method implemented the (Linhares Pontes, Huet, Torres-Moreno, & Linhares, 2018, 2020)'s approach to generate cross-lingual summaries. More precisely, we use the CoRank method and their MSC approach optimized to CLTS, which generates compressions guided by keywords and the cohesion of words. This approach is easily adaptable to other languages and can still improve the informativeness of cross-lingual summaries.

2.2 Our cross-language text summarisation approach

We implemented the approach proposed by Linhares Pontes *et al.* (Linhares Pontes, Huet, Torres-Moreno, & Linhares, 2018, 2020) to combine the analysis of documents in the source and the target languages and the Multi-Sentence Compression (MSC) method (Linhares Pontes, Huet, Torres-Moreno, da Silva, & Linhares, 2020) to generate more informative summaries (Figure 4). The following subsections highlight the architecture of our CLTS approach.

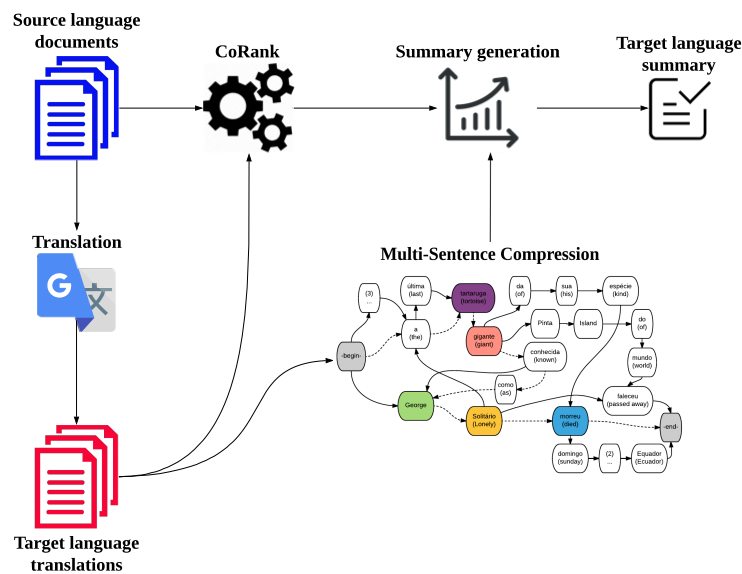


Figure 4: Architecture of the cross-language text summarisation model.

2.2.1 Preprocessing

Initially, source texts are translated into the target language with the Google Translate system², which was used in the majority of the state-of-the-art CLTS methods (Wan, 2011; Wan et al., n.d.; Linhares Pontes, Huet, Torres-Moreno, & Linhares, 2018).

Then, a chunk-level tokenization is performed on the target language side (Moirón & Tiedemann, 2006; de Caseli, Ramisch, das Graças Volpe Nunes, & Villavicencio, 2010). We applied a simple syntactic pattern ($< (ADJ)^*(NP|NC)^+ >$) to identify useful structures for English, where ADJ stands for adjective, NP for proper noun and NC for common noun. We also use the Stanford CoreNLP tool (Manning et al., 2014) for the English translations. This tool detects phrasal verbs, proper names, idioms and so

²<https://translate.google.com>

on. Unfortunately, we did not find a similar tool for the source languages; consequently, the chunk-level tokenization is limited to the target language (i.e. English).

2.2.2 CoRank Method

Sentences are scored based on their information in both languages using the CoRank method (Wan, 2011) which analyzes sentences in each language separately, but also between languages (Equations 14–18).

$$\mathbf{u} = \alpha \cdot (\tilde{\mathbf{M}}^{\text{sc}})^T \mathbf{u} + \beta \cdot (\tilde{\mathbf{M}}^{\text{tg,sc}})^T \mathbf{v} \quad (14)$$

$$\mathbf{v} = \alpha \cdot (\tilde{\mathbf{M}}^{\text{tg}})^T \mathbf{v} + \beta \cdot (\tilde{\mathbf{M}}^{\text{tg,sc}})^T \mathbf{u} \quad (15)$$

$$M_{ij}^{\text{tg}} = \begin{cases} \text{cosine}(s_i^{\text{tg}}, s_j^{\text{tg}}), & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

$$M_{ij}^{\text{sc}} = \begin{cases} \text{cosine}(s_i^{\text{sc}}, s_j^{\text{sc}}), & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

$$M_{ij}^{\text{tg,sc}} = \sqrt{\text{cosine}(s_i^{\text{sc}}, s_j^{\text{sc}}) \times \text{cosine}(s_i^{\text{tg}}, s_j^{\text{tg}})} \quad (18)$$

where M^{tg} and M^{sc} are normalized to \tilde{M}^{tg} and \tilde{M}^{sc} , respectively, to make the sum of each row equal to 1. \mathbf{u} and \mathbf{v} denote the relevance of the source and target language sentences, respectively. α and β specify the relative contributions to the final scores from the information in the source and the target languages, with $\alpha + \beta = 1$.

2.2.3 Multi-Sentence Compression

We aim to generate a single, short, and informative compression from clusters of similar sentences. Therefore, we grouped the similar sentences in the target language and then compressed them. Initially, similar sentences are grouped in clusters. Two sentences are considered similar if they have a similarity score higher than θ_1 . The similarity score of a pair of sentences i and j is defined by the cosine similarity in the target language.

As the majority of clusters are composed of few similar sentences (normally two or three sentences), Linhares Pontes *et al.*'s method (Linhares Pontes, Huet, Torres-Moreno, da Silva, & Linhares, 2020; Linhares Pontes, Huet, Torres-Moreno, & Linhares, 2020) processes MSC guided only by the cohesion of words and keywords.

Cohesion of words is defined by the frequency of words in the cluster:

$$w(i, j) = \frac{\text{cohesion}(i, j)}{\text{freq}(i) \times \text{freq}(j)}, \quad (19)$$

$$\text{cohesion}(i, j) = \frac{\text{freq}(i) + \text{freq}(j)}{\sum_{s \in C} \text{diff}(s, i, j)^{-1}}, \quad (20)$$

where $\text{freq}(i)$ is the chunk frequency mapped to vertex i and the function $\text{diff}(s, i, j)$ refers to the distance between the offset positions of chunks i and j in the sentences s of a cluster C containing these two chunks.

We also use Latent Dirichlet Allocation (LDA) to identify the keywords at the global (all texts of a topic) and local (cluster of similar sentences) levels to have the gist of a document and of a cluster of similar sentences.

The MSC method looks for a sentence that has a good cohesion and the maximum of keywords, inside a word graph built for each cluster of similar sentences according to the method devised by Filippova (Filippova, 2010). In this graph, arcs between two vertices representing words (or chunks) are weighted by a cohesion score that is defined by the frequency of these words inside the cluster. Vertices are labeled depending on whether they are or not a keyword identified by the Latent Dirichlet Allocation (LDA) method inside the cluster (see (Linhares Pontes, Huet, Torres-Moreno, & Linhares, 2018) for more details). From these scores and labels, the MSC problem is expressed as the following objective:

$$\text{minimize} \left(\sum_{(i,j) \in A} w(i,j) \cdot x_{i,j} - k \cdot \sum_{l \in L} b_l \right) \quad (21)$$

where x_{ij} indicates the existence of the arc (i,j) in the solution, $w(i,j)$ is the cohesion of the words i and j , L is the set of labels (each representing a keyword), b_l indicates the existence of a chunk with a keyword l in the solution, k is the keyword bonus of the graph³. Finally, we generate the 50 best solutions according to the objective (21) and we select the compression with the lowest normalized score (Equation 22) as the best compression:

$$\text{score}(c) = \frac{e^{\text{opt}(c)}}{||c||}, \quad (22)$$

where $\text{opt}(c)$ is the score of the compression c from Equation 21. We restrict the MSC method to the sentences in the target language in order to avoid errors generated by machine translation, which would be applied in a post-processing step on compressed sentences.

2.2.4 Summary Generation

A summary is composed of the most relevant sentences. Therefore, we select the most relevant sentences in the target language and then we replace these sentences by their compressions when they are available.

Finally, we add a sentence/compression to the summary only if it is sufficiently different (sentence similarity smaller than θ_2) from the sentences/compressions already in the summary.

2.3 Experimental evaluation

Following a similar experimental procedure to (Linhares Pontes, Huet, Torres-Moreno, & Linhares, 2018; Pontes, Huet, & Torres-Moreno, 2018), we estimate the performance of our approach in relation to the SimFusion and CoRank methods⁴. For the SimFusion approach, we considered 3 values for the λ : 0.0, 0.75 and 1.0. The MSC method selects the 10 most relevant keywords per topic and the 3 most relevant keywords per cluster of similar sentences to guide the compression generation. All systems generate summaries containing a maximum of 250 words with the best scored sentences without redundant sentences. We apply the cosine similarity measure with a threshold θ_2 of 0.4 to remove redundant sentences in the summary generation for all systems.

³The keyword bonus is defined by the geometric average of all weight arcs in the graph and aims at favoring compressions with several keywords.

⁴Unfortunately, the majority of state-of-the-art systems in CLTS are not available. Therefore, we only considered extractive systems in our analysis.

2.3.1 Datasets

We used the English language version of the MultiLing Pilot 2011 dataset (Giannakopoulos et al., 2011). This dataset contains 10 topics which have 10 source texts and 3 reference summaries per topic. These summaries are composed of 250 words. In order to analyse low-resourced languages, English source texts were automatically translated⁵ into the Croatian, Estonian, Finnish, and Slovenian languages. Specifically, we analyse source documents in Croatian, Estonian, Finnish, and Slovenian to generate cross-lingual summaries in English. Half of this dataset is dedicated to set the hyper-parameter (clustering similarity threshold θ_1) and the other half is to evaluate the performance of CLTS systems.

2.3.2 Evaluation

The informativeness is one of the most important features in CLTS. Informativeness measures how informative is the generated text. As references are assumed to contain the key information, we calculated informativeness scores counting the n -grams in common between the system output and the reference summaries. The ROUGE measure developed by (Lin, 2004) compares the differences between the distribution of words of the candidate summary and a set of reference summaries. The comparison is made splitting into n -grams both the candidate and the reference to calculate their intersection. Standard n -gram values for ROUGE are 1-gram and 2-gram, both expressed as:

$$\text{ROUGE-}n = \frac{\sum_{n\text{-grams} \in \{Sum_{can} \cap Sum_{ref}\}}}{\sum_{n\text{-grams} \in Sum_{ref}}, \quad (23)$$

where n is the n -gram order, Sum_{can} the candidate summary and Sum_{ref} the reference summary. A third common ROUGE variation is ROUGE-SU $_{\gamma}$. This ROUGE variation takes into account skip units (SU) $\leq \gamma$. For each ROUGE variation, we analyse its corresponding f-measure value.

2.3.3 Results and analysis

The clustering similarity threshold is a relevant parameter in our CLTS approach because our MSC method depends on the quality of clusters of similar sentences. While a lower clustering similarity threshold creates clusters of several sentences that may share a few similar contents, a high clustering similarity threshold creates clusters of similar sentences with few sentences. Our MSC approach performs best with large clusters of similar sentences. Therefore, we analysed our CLTS approach with different clustering similarity threshold values to estimate the best value for this threshold. Figure 5 shows the performance of our approach with the clustering similarity threshold θ_1 (cosine similarity) varies from [0.3 – 0.7] to group similar sentences in a same cluster. From the results on the four languages, our approach achieved the best results with the threshold = 0.5 and 0.6. To ensure that each cluster shares the same subject, we defined the similarity threshold θ_1 for the clustering process at 0.6.

Proceeding our analysis, all baselines generate summaries with the most relevant sentences and similarity threshold of 0.4 for removing redundant sentences in the summaries. ROUGE scores on the test datasets are listed in Table 1. Our system generated more informative summaries and achieved the best ROUGE scores. The analysis of repeated information in several documents helped our system to identify the most relevant information. Additionally, the MSC method reduced the redundancy and kept the key information of similar sentences.

The clustering threshold is a key factor in the MSC's performance in improving the quality of cross-lingual summaries. The higher the clustering threshold, the more similar the sentences are; however, a high threshold can create clusters with few similar sentences, reducing the performance of the MSC to combine these sentences and generate short informative compressions. The clustering threshold

⁵<https://translate.google.com>

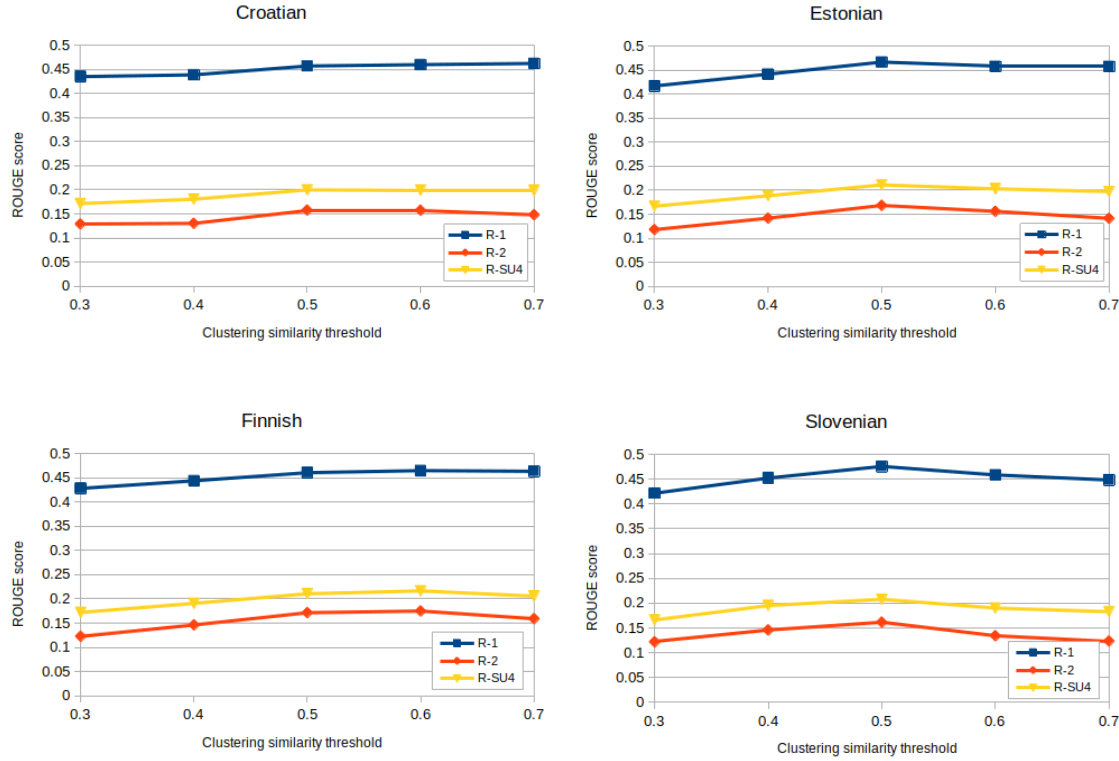


Figure 5: ROUGE scores of our approach on the validation datasets.

at 0.6 ensures that sentences in the same cluster share the same context/meaning; however, these clusters contain only two or three sentences, which limits the performance of the MSC method and, consequently, the quality of our cross-lingual summaries. Due to these factors, most cross-lingual summaries contain only one or two compressions, so our CLTS performs only slightly better than CoRank (Table 1).

Table 1: ROUGE scores on the test datasets.

Method	Croatian			Estonian			Finnish			Slovenian		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
SimFusion ($\lambda = 0.0$)	0.4084	0.1049	0.1532	0.4047	0.1061	0.1480	0.4198	0.1121	0.1601	0.3930	0.0900	0.1395
SimFusion ($\lambda = 1.0$)	0.3922	0.0814	0.1352	0.3922	0.0814	0.1352	0.3922	0.0814	0.1352	0.3922	0.0814	0.1352
SimFusion ($\lambda = 0.75$)	0.3968	0.0845	0.1389	0.3940	0.0844	0.1384	0.3939	0.0843	0.1381	0.3921	0.0837	0.1373
CoRank	0.4759	0.1506	0.1967	0.4638	0.1432	0.1907	0.4622	0.1401	0.1876	0.4646	0.1416	0.1877
Ours	0.4791	0.1567	0.2010	0.4676	0.1540	0.1972	0.4682	0.1512	0.1965	0.4656	0.1460	0.1891

The lower results of SimFusion method using $\lambda = 0.0$ and 1.0 with respect to other systems prove that the texts in each language provide complementary information. It also establishes that the analysis of sentences in the target language ($\lambda = 0.0$) plays a more important place to generate informative cross-lingual summaries. The CoRank method generates better results than all versions of the SimFusion because it considers the information in each language separately and together, while the SimFusion method using $\lambda = 0.75$ analyzes only the cross-lingual sentence similarity.

To sum up, the joint analysis of both languages improves the generation of cross-lingual summaries.

Our compressive approach achieved the best ROUGE results by reducing the redundant information and preserving the relevant information.

In the context of the EMBEDDIA project, our cross-lingual system can generate a short description (in English) composed of the most relevant sentences/compressions in a cluster of news documents in order to give the reader a general idea about these documents even if it is in a foreign language. Table 2 shows an example of Slovene-to-English cross-lingual summaries generated by a human and by our CLTS approach. More precisely, human annotators and our CLTS approach analysed source documents in Slovenian and generated summaries (up to 250 words) in English. All source documents and summaries of this example are listed in Appendix H.

Table 2: Example of summaries generated by a human and by our CLTS approach.

Manual summary

Pictures of extensive detainee abuse in Abu Ghraib were made public in early 2004. A year later, a United States Army court martial sentenced Army Spc. Graner to ten years in prison. A total of eleven American soldiers were found guilty in allegations stemming from the Abu Ghraib prison abuse scandal. Among them, two dog handler soldiers. During their testimony, they said they learned the interrogation techniques that included the use of dogs from a team of interrogators that was dispatched to Iraq from Guantanamo military base. On September 2005, a U.S. federal judge ordered the release of additional photos and video relating to the case. The Bush administration claimed that the ordered media could provoke terrorist attacks. Five months later, the Australian TV station SBS showed new photos claiming that some of them document previously unprosecuted abuse. According to a report published on June 2006 by Human Rights Watch, torture and other abuses against detainees in US custody in Iraq continued and were authorized. The Department of Defense denied any Pentagon approval for any abuse. On January 2007, the U.S. Army announced that U.S. Lt. Col. Steven Lee Jordan was going to be tried by a military court. Jordan is the only U.S. officer charged in the Abu Ghraib case. A year earlier, a top US commander, who supervised the detention and interrogation of detainees at Guantanamo Bay and Abu Ghraib facilities, had declined to testify in a court-martial proceeding, by invoking his right to not implicate himself.

Our CLTS approach

Convicted on the first of June by a military jury for participating in prisoner abuse at Abu Ghraib, former sergeant Santos Cardona, 32, a dog handler in the United States Army, was today sentenced to ninety days of hard labour and demoted to the rank of specialist. The Guardian cites an unnamed US defense official saying that the Army had reviewed the pictures posted by SBS and confirmed that they were among those that are subject to the Freedom of Information Act request made by the ACLU. Army charge sheets accuse Cardona and Smith with maltreating detainees from November 15, 2003, to January 15, 2004 by directing, encouraging, or permitting [their] unmuzzled military working dog[s] to bark and growl at detainees in order to unlawfully harass and threaten the detainees and in order to make the detainees urinate or defecate on themselves. A report is published by Human Rights Watch on treatment of prisoners in Iraq by US soldiers after the Abu Ghraib prison scandal. SBS alleged the photographs of the dead bodies were of people who had died at Abu Ghraib during interrogation. United States army officers stated that the Abu Ghraib prison will be closed within months, and its prisoners moved to other prisons and camps in Iraq. The two accused said in yesterday's testimony that Col. Thomas M. Pappas, the top military intelligence officer at Abu Ghraib, approved the use of the dogs. Charles Graner guilty of abusing prisoners at the Abu Ghraib prison .

The paper associated with this work is provided in Appendix A.

3 Corpus Summary Visualisation

This section presents our work in the area of visual textual summaries. This work ultimately aims to produce novel visual text summarisation techniques for use in visualising corpora of cross-lingual

texts and to enable exploration of generated textual summaries via their source texts. Here the idea of concordance visualisation as a form of topic based corpus summary is presented and a tool for exploring these visualisations through the lens of metadata is detailed.

Concordance analysis is a core technique in a number of scholarly disciplines where the analysis of corpora is performed. These disciplines include discourse analysis, translation studies and corpus linguistics, to name a few.

While concordance analysis has been used for the analysis of journalistic output (Hansen, 2016) it is not a technique which has found popularity amongst journalists. In the current age of digital journalism there have been calls for the use of concordance analysis, visualisation, and corpus comparison tools to enhance the journalistic method (Karlsson & Sjøvaag, 2018). By providing visual tools which support these techniques journalists may benefit from more efficient analysis of news corpora.

The concordance is a list of vertically aligned text fragments, where all occurrences of the keywords are displayed centrally along with a window of the left and right contexts. A concordance can be conceptualised as a summary of the contents of a corpus related to a keyword or a topic. A topic base concordance is simply a multi-keyword concordance where the keywords describe the topic.

In s previous collaboration with researchers from the Genealogies of Knowledge project⁶ we developed tools for the analyse of corpora. Development of these tools has continued and the metadata explorer *Metafacet* was added to the software. The tools were designed to be multilingual, initially providing explicit support English, Greek and Arabic while also working well for many other languages supported by the UTF8 character set.

In the following section the corpus visualisation tools which were built as plugins for the Modnlp toolkit (Luz, 2011, 2000), including the new *Metafacet* tool, are described using examples of how they could benefit in the summarisation and analysis of news corpora.

3.1 Metafacet

We developed the *Metafacet* visualisation to enhance exploratory interaction with corpora. The tool is designed to enable the interactive filtering of a corpus query using all available meta-data facets. The interfaces which currently can be filtered are the traditional concordance list and the concordance mosaic visualisation which we had previously developed (Luz & Sheehan, 2014). Both of these concordance interfaces provide a view of a keyword-in-context corpus search result. The *Metafacet* tool provides frequency distribution information of the returned concordance lines across the metadata attributes of the corpus.

The *Metafacte* interface uses a horizontal bar chart to display concordance line frequency per meta-data attribute. An attribute is a possible value that a meta-data facet can take. As an example “Plato” is an Attribute of the Facet “author”. A drop-down list is used to chose which facet is displayed and the bars are sortable by frequency or lexicographical order and the window can be zoomed using a sliding scale to view a smaller portion of the attributes. In Figure 6 the *Metafacte* interface is shown with the facet “Internet Outlet” displayed. The results shown are ordered by frequency displaying how many lines in some concordance are from each outlet.

In Figures 7 and 8 we see the *Metafacet* displayed along with a concordance list and concordance mosaic interface. A concordance for the keyword “news” from a corpus of political internet magazines is displayed. The concordance mosaic shows the frequency of words at four positions to the left and right of the keyword. In Figure 7 we can see the word “fox” is the second most frequent word occurring just before the keyword “news” in the concordance list. However, this concordance list and mosaic have been filtered using *Metafacet* to display only concordance lines from the years 2007 to 2015. In this interval we found the pattern “fox news” to be consistently the second most frequent concurrence pattern at the position for each individual year. In Figure 7 the red bars are attributes of the “Publication

⁶<http://genealogiesofknowledge.net/>

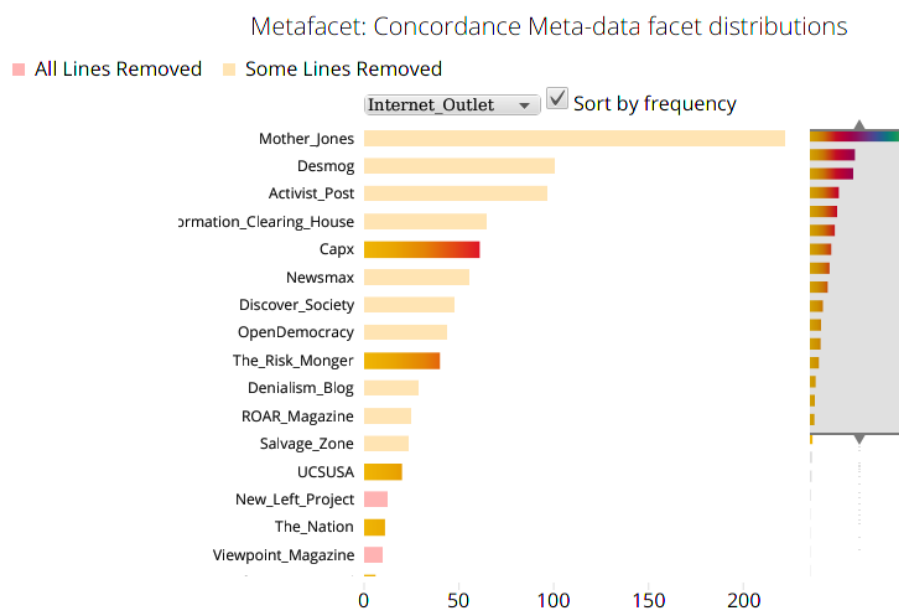


Figure 6: Metafacet tool showing attributes for the “Internet Outlet” facet. The interface shows partial and total line removal caused by a filter which is currently applied to another facet

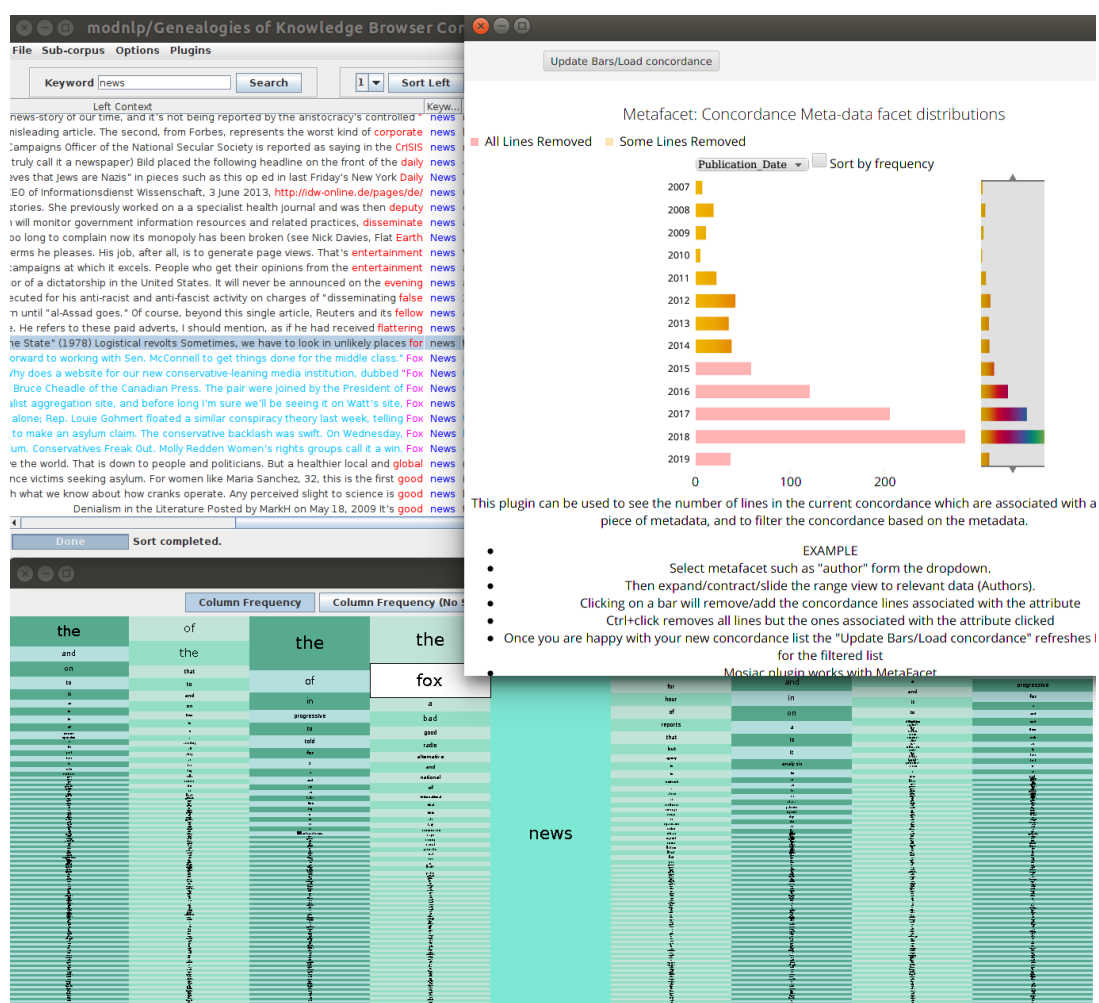


Figure 7: Metafacet, concordance lines and concordance mosaic views of the keyword “news” in a corpus of online magazines. Metafacet activated to remove concordance lines from the list and mosaic for any articles published after 2015

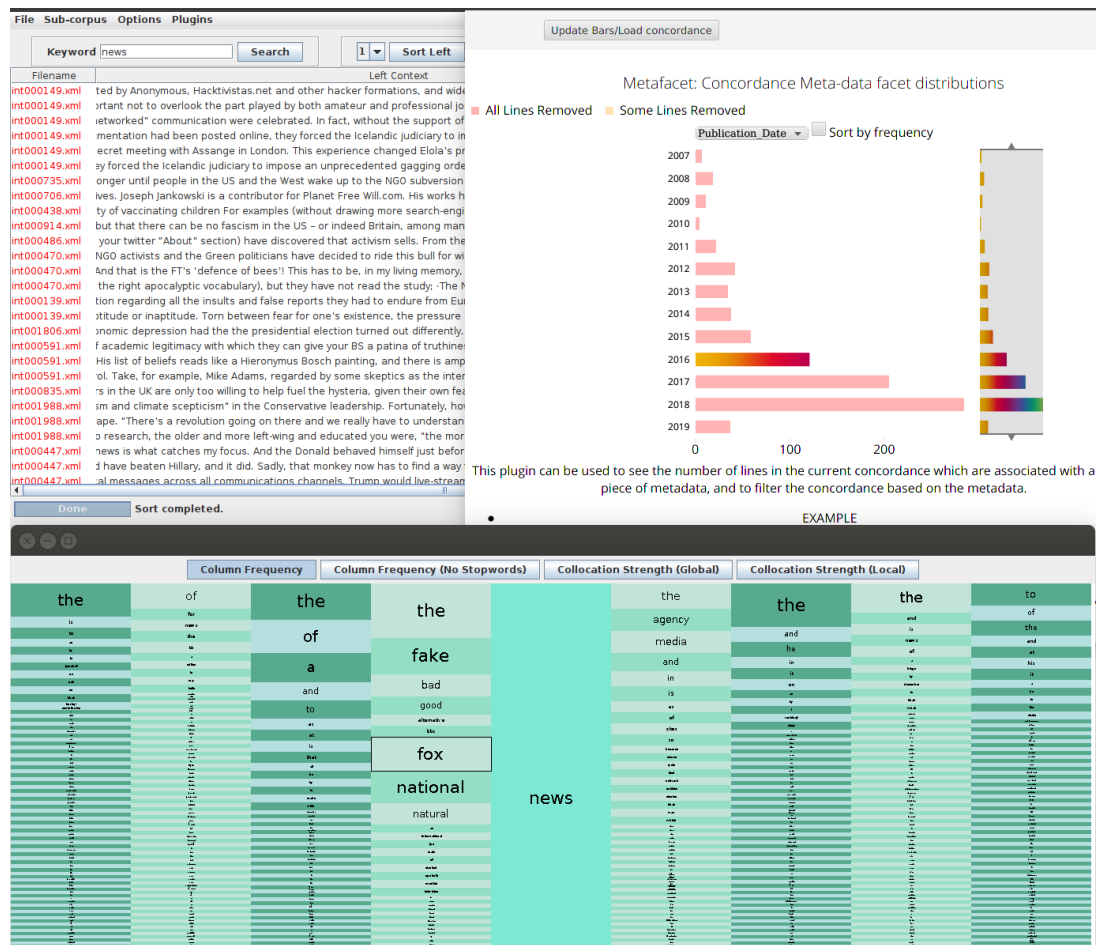


Figure 8: Metafacet, concordance lines and concordance mosaic views of the keyword “news” in a corpus of online magazines. Metafacet activated so that only concordance lines from the year 2016 are displayed

Date” facet which have been excluded from the current concordance displays. Looking to Figure 8, where only the year 2016 has been selected, we see the relative usage frequency of “fox” directly to the left of “news” has declined, while “fake” has emerged as a frequent collocate at this position. Looking at each subsequent year after 2016 the same collocation pattern of “fake news” is the second most frequent positional collocation while it was very rare in the years prior to 2016 in this corpus.

To conclude this example we look again to Figure 6. *Metafacet* is displaying the ‘Internet Outlet’ facet for the “news” concordance which has been filtered to remove all lines published before 2016. The yellow bars indicate that some of the lines associated with this attribute have been removed and that clicking the update button will reduce the length of these bars. The bars in red will be entirely removed, these internet outlets do not contain any texts in the corpus which contain the word “news” after the year 2016. The analysis could continue further by looking in more detail at both the concordance and metadata, perhaps exploring the concordance of the bigram “fake news” could help to explain the emergence of the term.

The concordance and metadata combined in a visual exploration tool offer a means of summarising the content of a corpus as it relates to a concept or keyword, but it does not offer high level summary of the corpus as a whole

3.2 Corpus Comparison

A visualisation tool which we had developed prior to this project can be used to compare visual summaries of the contents of two corpora. To use the tool for cross lingual summarisation both corpora must be translated into a single language.

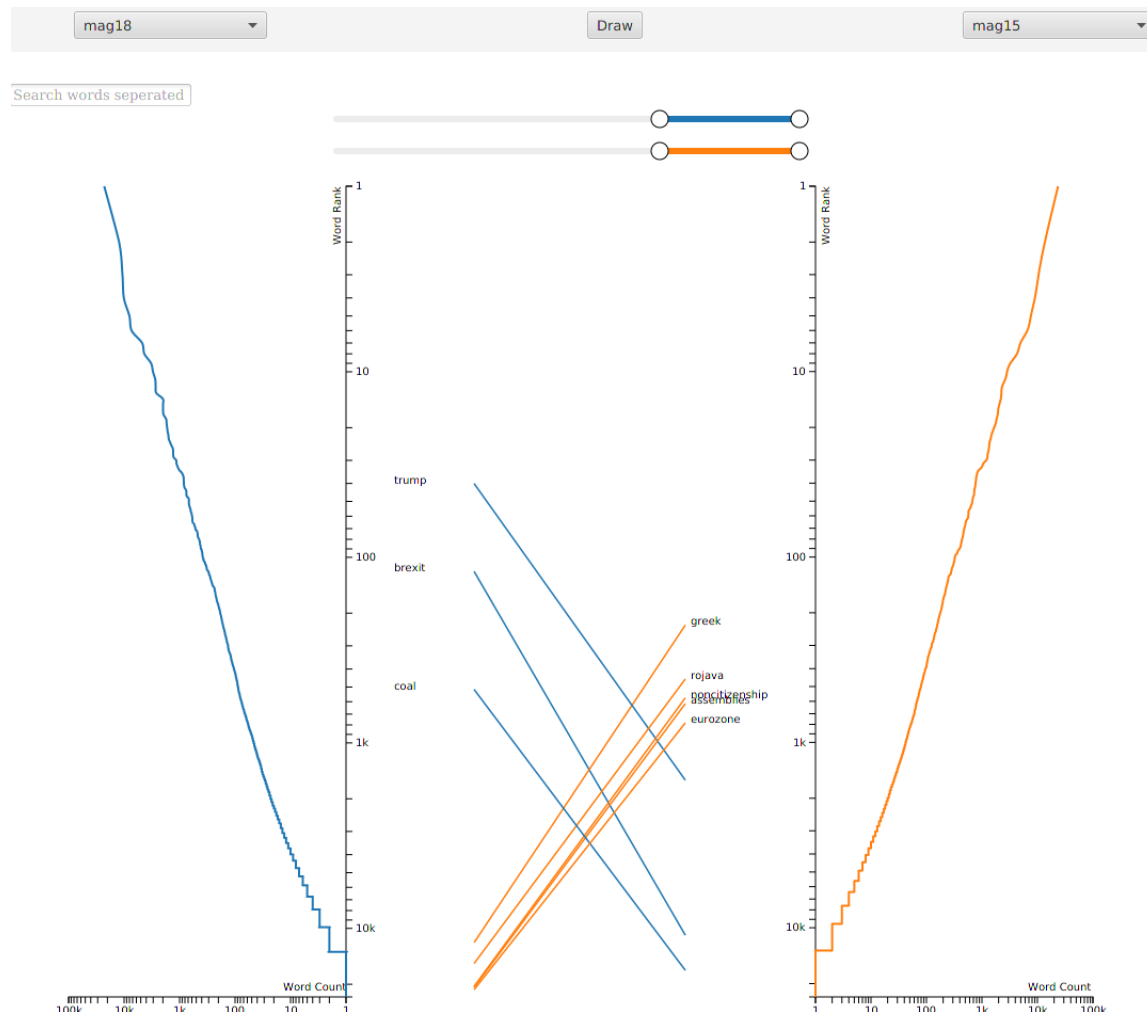


Figure 9: ComFre Visualization comparing the words with the largest change in distribution rank between magazine articles from the GoK internet corpus published in 2018 and 2015

The tool *ComFre* is a corpus comparison tool where frequency lists can be compared visually in a statistically valid manner (Sheehan, Masoodian, & Luz, 2018). It is currently available as a plugin for the Modnlp toolkit.

The Modnlp software has a sub-corpus selection interface which can be used to save named sub-corpora for later reuse. *ComFre* makes these named subcorpora available for comparison in dropdown lists. In Figure 9 “mag18” and “mag15” are selected for comparison, these subcorpora are magazine articles which were authored in 2018 and 2015 respectively, taken from the same corpus as was used in the *Metafacet* analysis of the keyword “news” in Section 3.1.

In *ComFre* both axis are log scaled which should yield a linear frequency diagram if the word frequencies follow a Zipfian distribution. Scaling both ranked lists to the same height and comparing a words position in the distributions lets us compare subcorpora of vastly different size. The change in height, between two corpora, of a word along the vertical axis is an a quantity linearly comparing position in the d rank

frequency distribution of two corpora. The higher a word appears on the vertical axis the greater its relative usage in the corpus.

In Figure 9 the majority of the words have been filtered out to reveal the words with the greatest frequency changes between the two corpora. We can see that the words “trump”, “brexit” and “coal” were used much more often in the 2018 corpus, while words such as “greek” and “eurozone” had much higher usage in 2015.

By comparing the differences between these temporal subcorpora it may be possible to identify reasons for the changes in context around the word “news” which were identified for the corpus in Section 3.1.

The two papers associated with this work are provided in Appendices B and D.

4 Current visualization work partly related to EMBEDDIA

The work described in this section was developed during the EMBEDDIA project but applied initially to other domains. This work is in the process of being adapted for the news data and has the potential to be an important contribution to the project.

Section 4.1 presents our work on temporal mosaics for the exploration and summarisation of temporal trends in corpora or single texts. This work includes techniques for exploring the links between a summary document and the source texts via these temporal mosaics. The visualisations were developed in the context of medical training and the design rationale is described. At the end of the section an early adaption to news data is described and future directions for the work are discussed

In Section 4.2 a graph based visual tool, named NetViz, for the exploration of topics and related terminology is presented. The tool was first applied to geological data but some early work applying it in a news context is also presented.

4.1 Temporal Mosaic Visualisation

When exploring a collection of texts it is helpful to be able to organise them and summarise their content. Visualisation can assist in this exploration through carefully designed data abstractions and visual encoding. Often documents have associated metadata and may be assigned to categories, either manually or using automated methods. In many domains including news media, the change in distribution and content of these categories is of interest. To visualise this change in content over time, keywords associated with a category or topic can be used to provide an overview and to identify patterns.

Figure 10 shows a mock up of a visualisation concept which makes use of the temporal mosaic idea (Luz & Masoodian, 2007) to summarise the content of a number of topics in a corpus over time. The mock-up shows horizontally aligned timeslots which contain stacked bars representing topics. Each colored block represents a topic, in this example the topics are European politicians and the time slots are years. Words which are strongly associated with the topic within the timeslot are displayed in the topics colored box. Topics which are significant within a timeslot are given equal vertical height. The visualisation allows the viewer to get an overview of the words associated with each topic and how they may have changed over time. The example shown in Figure 10 is a high fidelity prototype used for exploring the concept of visual summaries of temporally corpus content.



Figure 10: Temporal mosaic mock-up. Simulated change in keywords related to some European politicians over time



Figure 11: The *TeMoCo* prototype, with a speaker turn selected on the visualization (grayed out mosaic on the left), and the relevant parts of the transcripts highlighted (orange background text on the right).

4.1.1 Temporal Transcript Summarisation

As an extension and use case of the temporal mosaic visualisation we developed an interactive version, named *TeMoCo*, to visualize multi-party conversations. The prototype was designed to visualise the text related to the temporal content of a conversation where transcribed audio is available from multiple speakers. The prototype could be used to summarise audio and video news content such as interviews and debates where there are multiple speakers and transcribed audio is available. While in this example topics are mapped to speakers and linked to a transcript similar techniques could be used to connect temporally aligned corpora of news articles to a temporal mosaic based summary of viewpoints, named entities or topics. This technique currently uses simple frequency based measures of keyword salience, however embedding based techniques for describing the temporal topic slices will be explored in the future.

To develop the tool we used a corpus of recordings from an evaluation exercises performed by nursing students at a university hospital. Each interaction, or session, is a training exercise to assess clinical and communication skills of the student. In the dataset, each interaction of the corpus includes an anonymised time-aligned transcript. Figure 11 shows the interface of the *TeMoCo* prototype. As can be seen, the left-hand panel is the interactive visualization showing the temporal mosaic patterns of the conversation – along with the top keywords selected from each speaker turn – and the right-hand panel shows the transcript of the entire conversation session. In this conversation session there are five participants (Patient 1, Nurse 1, Doctor 1, Doctor 2, and Medical Registrar 1), who have been talking for 13 minutes and 30 seconds.

While the static view of *TeMoCo* is sufficient for seeing the patterns of conversation, and a summary of its main keyword points, the user can get a detail-on-demand view by clicking on a speaker turn mosaic on the visualization to access the relevant parts of the conversation on the transcript. Figure 11 shows a selected mosaic (in gray colour) on the left for participant D1, between 04:30 and 06:00. By selecting a mosaic, the prototype tool locates the start of the transcript text related to the selected time-slice (04:30–06:00), grays out the background of all the text for that time-slice, and then highlights the segments of the transcript text for the chosen speaker during that time-slice using the colour assigned to that speaker (the orange colour for D1 in Figure 11). This follows the well-known visual information seeking mantra (Shneiderman, 1996).

This choice of visual encoding is informed by Mackinlay's ranking of visual variables (Bertin, 1983). Vi-

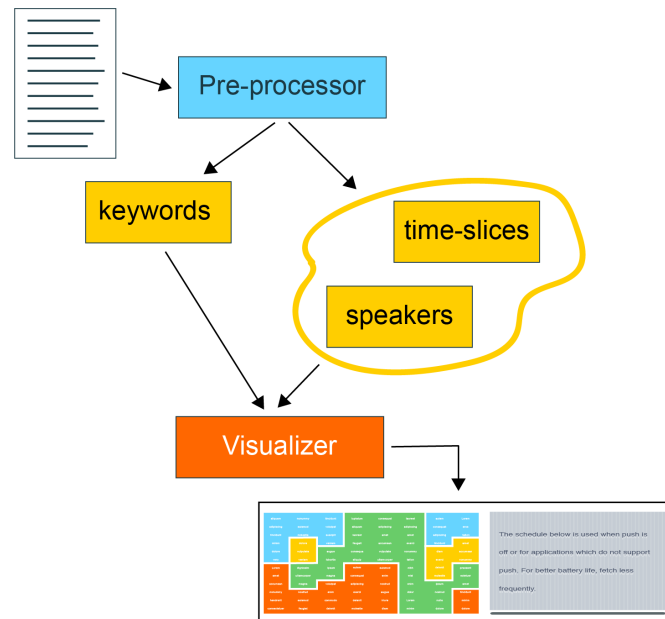


Figure 12: Architecture of the *TeMoCo* prototype.

visual variable theory describes how data attributes map to visual variables and the appropriate use cases for each variable in terms of data type. Ordinal data such as temporally aligned items are well suited to the use of the position variable, categorical data is also rendered effectively using position, but color is another strong choice for its encoding. In *TeMoCo* the order of the time-slots is mapped to horizontal position, and the categorical speaker information is mapped to color hue. The transcript is placed next to the temporal mosaic, and the vertically-ordered speech segments are placed in a scroll box again using position (vertical) to map the linear temporal data to a position attribute. In this visualisation a quantity displaying speaker/topic contribution in a time slot was not encoded, instead the topics of interest are given equal visual weight if they contribute in the time slot. This choice makes it easier to identify segments in which the same topics or speakers contributed, but does not provide information related to the strength of the contribution in each segment or across the entire dataset.

4.1.2 Implementation

Figure 12 shows the architecture of the *TeMoCo* prototype which has been implemented as a single-page web application using the *D3.js* framework (Bostock, Ogievetsky, & Heer, 2011). The current system creates the visualization using a transcript file made available to it on the server. The transcript text is time-stamped and tagged with the labels of the conversation participants or topics.

The system starts by pre-processing the transcript text to create two data streams. The first stream generates a data source containing relevant keywords, in which the keywords are selected for each time-slice and participant combination. Keyword salience is dependant on context and use case – measures such as word frequency or frequency in a domain specific reference corpus are an obvious starting point. In our tests we found raw frequency to be uninformative, and subsequently decided on the manual selection of seemingly salient words, this simulates the word selections that could be achieved automatically using reference corpus. Depending on the corpus and use case, any statistical measure of word salience or keyness could be injected to produce the keywords for a topic or speaker in a time-slice. The second stream of data is generated by extracting the time-slice and topic or speaker information. This information is then used for tagging the input transcript with HTML attributes. This enables dynamic manipulation of the raw transcript as a part of the system interface.

Once the two data streams have been processed, the system constructs the temporal mosaics of the



Figure 13: Sketch of the main components of the visualization: temporal mosaic of the audio dialogue (top left) and its transcript (top right), and the related text document (bottom).

TeMoCo visualization from the time-slices, speakers and keywords information. The visualization and transcript panels are then positioned in the same web page. Both views are linked via the data, allowing interactions between the two. Selection of a time-slice mosaic scrolls the transcript to the corresponding time-slice, as describe above.

4.1.3 Contextual Linking

Written reports of events or actions which have related recorded audio or video data are widespread across a variety of domains. For instance in the media domain, web-based news articles are often presented with video clips of the reported events. Similarly, medical contexts where recording and reporting is commonplace, audio and video recordings are often made during clinician-patient consultations, multidisciplinary medical team meetings and training, and so on.

Analysis of these reports are, however, often very time-consuming and not well supported by existing software or visualization tools. For instance, the link between an audio recording and a textual document is rarely made explicit, thus making it difficult to quickly switch between the textual document and the recording. To identify which spans of text in a report are linked to particular times in a recording requires either a full examination of both sources, or some other linking mechanism. Similarly, generated textual summaries are not often implicitly linked back to the source document content which was used to generate them.

We designed an interface which maps a textual document via contextual links to a temporal mosaic visualisation of the contents of a recording and individual speaker's contributions. In addition, both the document and temporal visualization are implicitly linked to a transcript of the recording (which could easily be replaced by a highlighted timeline with accompanying video or audio). This interface is an extension of the *TeMoCo* visualization and is named *TeMoCo-Doc*. While *TeMoCo* focused on identifying the temporal content, *TeMoCo-Doc* focuses on identifying the links between the content of a summary document and the temporal speech segments of recorded audio.

The temporal aspect of the visualization, seen in the top left panel of the interface sketch Figure 13 and prototype Figure 14, uses temporal mosaic visualizations (Luz & Masoodian, 2007) to render the speaker contribution per time-slot and displays the salient word for each speaker in the corresponding slot.

When used as an interactive visualization (Luz & Masoodian, 2005) each rectangular segment of a temporal mosaics visualization can be linked to the corresponding part of the data-stream it represents – thus supporting access to media content, both temporally as well as contextually. In Figure 14 one speaker has been selected in four time-slots. This causes the transcript to scroll to the beginning of the first selected time-slot, and the corresponding speech segments are highlighted. This combination of temporal mosaic and transcript is based on the *TeMoCo* visualization, speaker or topic contribution in each time segment is now calculated and used in the visualisation. We use the height of each colored box to encode this quantitative value of contribution withing a timeslot, making used of the second best visual variable for encoding quantitative information (Mackinlay, 1986). This encoding also now provides an overview of topic or speaker contribution in the entire window, the area of each color is proportional to the contribution.

Looking at the sketch of our visualization tool (13) we see three juxtaposed views (Javed & Elmqvist, 2012): the temporal mosaic (top left), the transcript (top right), and the view containing the report (bottom) with its contextual links made explicit by the red lines and outlines. The visualization prototype tool implements this sketch, and the red contextual linking via interactions with the report. Mini temporal mosaics (with the salient words removed) are rendered next to each section of the report. Spans of text in each section which were identified as being contextually linked to the transcript are colored according to the speaker they map to. By hovering over any of these contextual spans the user is able to see the related time-slots on the mini mosaic (as can be seen in the bottom most mosaic in 15), which gives a preview of the slots that will be selected by clicking on the span. Each of the document

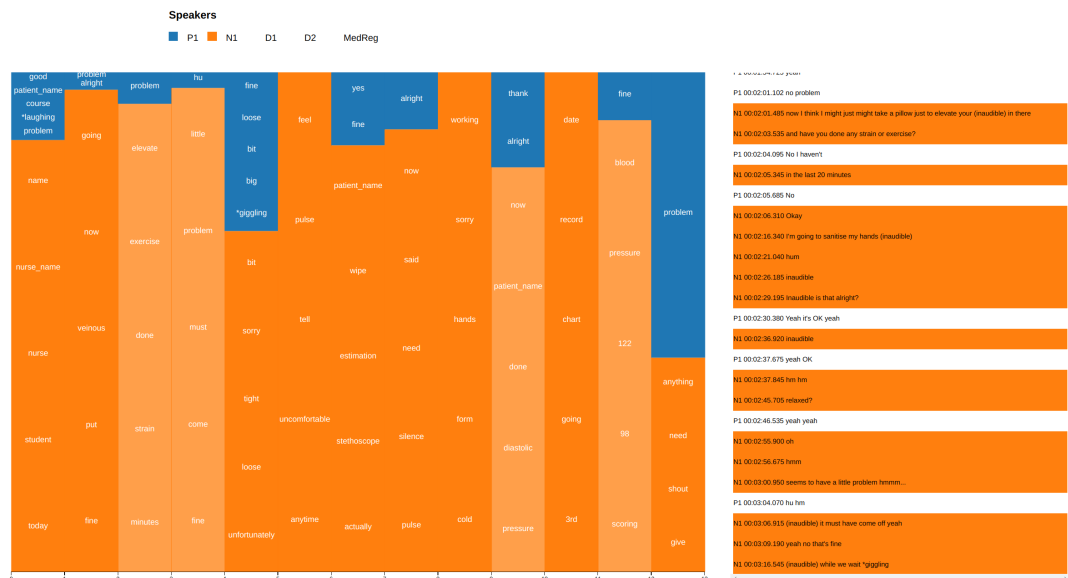


Figure 14: Temporal mosaic (left) and transcript (right) in the *TeMoCo-Doc* visualization. A contextual edge has been selected and the related speaker time segments have been highlighted. The related sections of the transcript have also been highlighted.

Reflection Document

Introduction

Reflection is defined by Dewey (1933) as cited by Bulman & Schutz (2008) as: Opening up one's practice for others to examine, and consequently requires courage and open-mindedness as well as a willingness to take on board and act on criticism (p.2). It's an essential skill that health professionals have to learn when working within a multidisciplinary team to help improve the overall standard of care. The Nursing and Midwifery Council of England (2016) as cited by Flaherty (2016, p.16) emphasise that all nurses and midwives as of April 2016 will need to prove they are able to practice safely and effectively. Therefore there is a need for nurses especially to be constantly analysing and reanalysing their actions and communication through reflection as they are at the frontline in practice. Nurses are vital in advocating patients' feelings hence their interpersonal skills need to be persistently clear, concise and easily understood to avoid miscommunication and putting the patient at risk (McCabe & Timmins 2006). In order for this to occur, nurses need to perfect the skill of reflection by using a reflection framework to improve their competence (Fildes et al. 2015). The framework I will be using is 'The Reflection Cycle' by Rolfe et al. (2010).

What happened?

The clinical skill I was told to perform was manual blood pressure. Upon receiving the skill I was slightly nervous. Manual blood pressure (BP) seems straightforward but if not conducted correctly it's very easy to get inaccurate results (Rafley 2009). I brought the equipment over to the patient then **introduced myself and explained the process. The patient expressed no feelings of discomfort** and consented for me to do their blood pressure. I tried to narrate all my actions in order to keep the patient informed (Rafley 2009). I sanitized my hands in preparation for the procedure. The patient stretched out their arm and I placed a pillow under it for support (Dougherty 2015). I sanitized again as I touched the patient's environment (World Health Organisation 2006). **I then measured the circumference of the upper arm with the cuff** folded at the bladder, making sure it was 80% the circumference of the upper arm, 40% the width and placed it 2-3 centimetres above the brachial

What was significant?

I felt by introducing myself and showing genuine concern about the patient's wellbeing, I reduced the 'white coat effect' (Rafley 2009). 'The white coat effect' is a temporary rise in blood pressure experienced by an individual, when a doctor or nurse is taking their BP. I was constantly narrating myself and **ensuring the patient was comfortable** and that neither the sleeve nor the cuff was too tight on their arm (Dougherty & Lister 2015). I admit I did get slightly bewildered when I discovered the valve on the sphygmomanometer was missing, thus I had to use a replacement sphygmomanometer with a cuff intended for obese people. In hindsight, **I should not have become so flustered as I then forgot to change the obesity cuff to a standard cuff perhaps leading to the high diastolic value I recorded.** It's been proven that an ill fitting cuff can give inaccurate results (Dougherty & Lister 2015). I just assumed that all the equipment was present and working, which is an assumption

Now what ?

Upon reviewing my performance, I'm quite pleased with my communication during the performance but feel I could have worked at a faster pace. I'm conscious that sometimes my perfectionist tendencies can affect my time management, which is an area I'll have to review in more detail. Perhaps upon returning to the clinical area, I will begin to time myself and set personal goals for completing tasks (Mirzaei et al. 2012). Henceforth, **I plan to check all equipment before beginning procedures**, give adequate patient response time and most importantly fasten my pace. 'The Reflective Cycle' by Rolfe et al. (2010) really helped me to identify areas of improvement in my clinical practice. Reflection is about evaluating your own mistakes and letting others examine them too as part of the learning process (Oliveira 2015).

Reference list: Beevers G., Lip G.Y.H. & O'Brien E. (2001) Blood pressure measurement. Part I Sphygmomanometry: factors common to all techniques. British Medical Journal 322(7292), 981-985. Bulman C. & Schutz S.

Figure 15: Textual document with contextual links in the *TeMoCo-Doc* visualization. A contextual link is selected and the related speaker time segments have been highlighted on the corresponding mini temporal mosaic. The related sections of the transcript have also been highlighted.

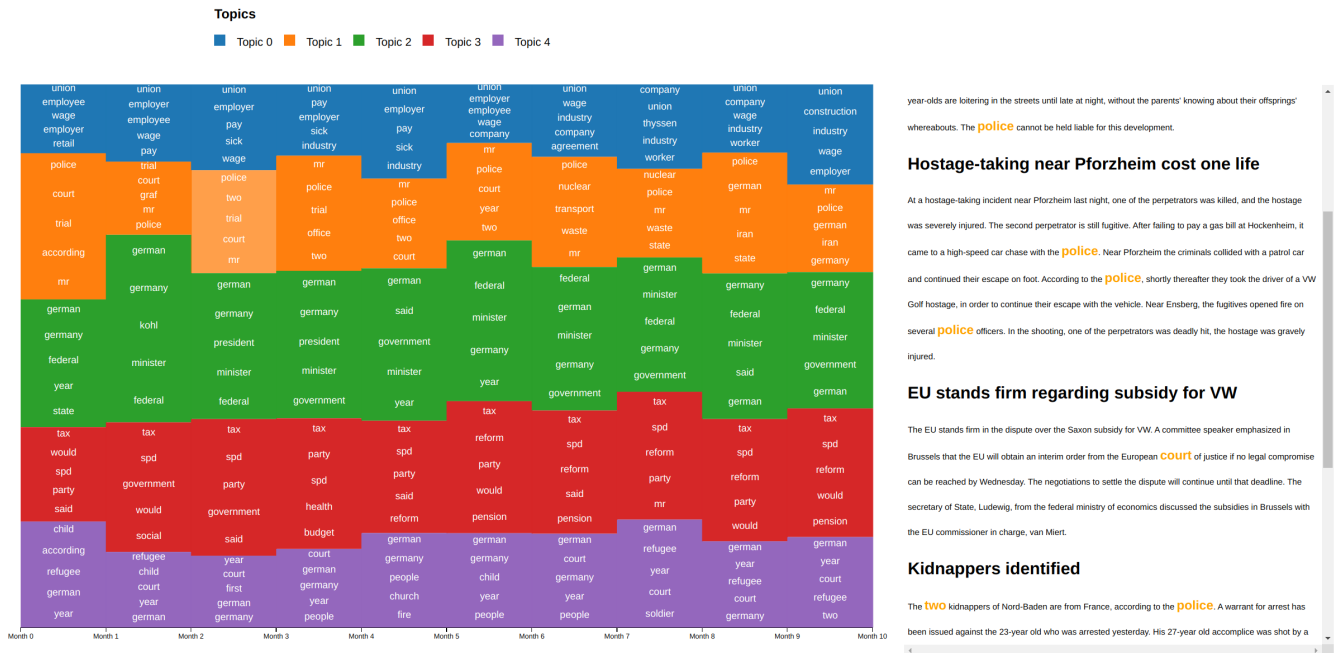


Figure 16: A version of TeCoMo which is applied to a dataset of short German news articles in the English language. The text window (right) displays the articles associated with the selected temporal topic segment in the visualisation (left)

sections is presented in a fixed height scroll-box, so that even large documents can be explored in a single screen.

The document example shown across Figure 14 and 15 represents a reflexive report based on a medical communication training exercise, which we will discuss further in the following section. In Figure 15 the user can see the span “I plan to check all equipment before beginning procedures” is clicked, outlining the related time-slots on the mini mosaic. Looking at the large mosaic Figure 14, the user can see the same time-slots are again selected and the transcript is highlighted accordingly. This allows the user to explore the transcript, and quickly see the time-slots and content in the conversation which relate to “I plan to check all check all equipment before beginning procedures”. This interface shows how visual and textual summarisation of content can be combined with a source document to provide summaries which cover several levels of overview abstraction and detail on demand.

4.1.4 Temporal News Visualisation

In a collaboration with task T4.1 a dataset of 2694 short news articles in English from German news outlets was prepared. The articles span a 10 month period from August 1996 to May 1997. Our collaborators automatically identified topics in the data and the topic weights associate with each topic for each month was calculated. From the articles associated with each month keywords were extracted. These keywords were linked to topics and timeslices through the article publication dates and topic score. Using this information we generated a version of the TeMoCo visualisation, seen in Figure 16, which shows the change in distribution and keywords associated with the top five topics in the dataset over the 10 month period. The visualisation helps with interpreting the unlabeled topics by showing the change in keywords for the topics over time, this gives an overview of the topics contents over the time period. To enhance the ability to investigate the temporal topics an interaction was added which displays the text of the articles associated with a topic time segment when the segment is clicked. In Figure 16 the orange segment for the third month is selected and the associated articles are displayed in the text

window (Right). The relevant keywords are also highlighted, in the segments color, in the text window to emphasise the link between the selected segment and the articles. This work is preliminary and has not yet been evaluated, one obvious future directions could be to display textual summaries of the articles related to a segment rather than the full collection of related articles.

In conclusion, the temporal mosaic provides an effective visual summary of corpora and single texts with a temporal component. For text corpora, publication date or similar attributes can be used to segment the corpus and single texts can use paragraph, chapter or section structure if no obvious temporal attributes exist as in the case of transcripts of audio. Connecting the temporal mosaic to source texts and enabling interactive exploration allows the user the freedom to quickly switch between high level overviews and the low level detail in the sources linked to the summary. One avenue for future work is to combine this interface with the technology produced in other tasks from the EMBEDDIA project. For example the viewpoint detection technology developed in other tasks could be used to segment a temporal corpus of news as an input to the visual system where the change in viewpoint content could be examined and explored in relation to the source articles. Similarly, by using cross-lingual embeddings temporal mosaics of multilingual content could be created in future work. Finally in future work, by using the contextual linking technique the temporal mosaic can be used to make explicit the links between generated textual summaries and the sentences which were used to generate them.

The two papers associated with this work are provided in Appendices C and E.

4.2 Graph-based visualisation

In (Miljkovic, Kralj, Stepišnik, & Pollak, 2019) we present an approach of graph-based topic modeling and visualisation. First, a network is constructed from co-occurring domain terms (keywords) and then the community detection algorithms grouped specialized terms into semantically related topics, which were also visually presented as coloured nodes in the graphs. In (Pollak, Podpečan, Miljkovic, Stepišnik, & Vintar, 2020), we developed the NetViz tool which allows for high-performance online network visualization where the user can upload the data in a simple CSV format, define the nodes (terms, categories), edges (relations) and their properties (by assigning different node colors). While originally, the NetViz visualisation tool was developed for non-news related domain (i.e. karstology domain modelling, see Appendix F) it can be useful in that context.

Initial experiments on news data: To exemplify the use of the tool in a news context of EMBEDDIA we constructed a network from the named entities in the SentiNews dataset (Bučar, 2017) (collaboration with task T4.3) covering news from several Slovenian news portals between 2007 and 2013. First, the dataset was processed by named entity recognition and named entity linking tools from WP2. Next, named entities that were not linked were removed. Then, named entities were filtered according to the manually specified min and max frequency. Finally, a network was constructed where a named entity is connected to another named entity if they co-occur in the same sentence. Singleton nodes are removed, while multiple co-occurrences increase the width of the edge which is finally normalized to the [1...10] interval.

In the NetViz visualisation, the following shapes are used to represent named entities categories: ellipse for location (LOC), box for organization (ORG), and circle for persons (PER). In addition, the color of the node codes the average sentiment (five-level Lickert scale (1 – very negative, 2 – negative, 3 – neutral, 4 – positive, and 5 – very positive)) of the sentences the entity appears in. Blue color denotes negative sentiment while red color denotes positive sentiment. The neutral sentiment is represented as white. While a detailed analysis was not yet performed, the graph in Figure 17 shows interesting information (for PER François Hollande and Vladimir Putin seem associated with positive sentiment, while Nicolas Sarkozy and several Slovenian politicians with a negative one. In terms of organizations (ORG), United nations and European Council are depicted as positive nodes, while Slovenian health institutions are coloured blue (negative), for LOC South Korea is associated with positive while North Korea with negative sentiment, and for example Haiti is very negative, where a hypothesis can be that the news refer to the 2010 Earthquake.

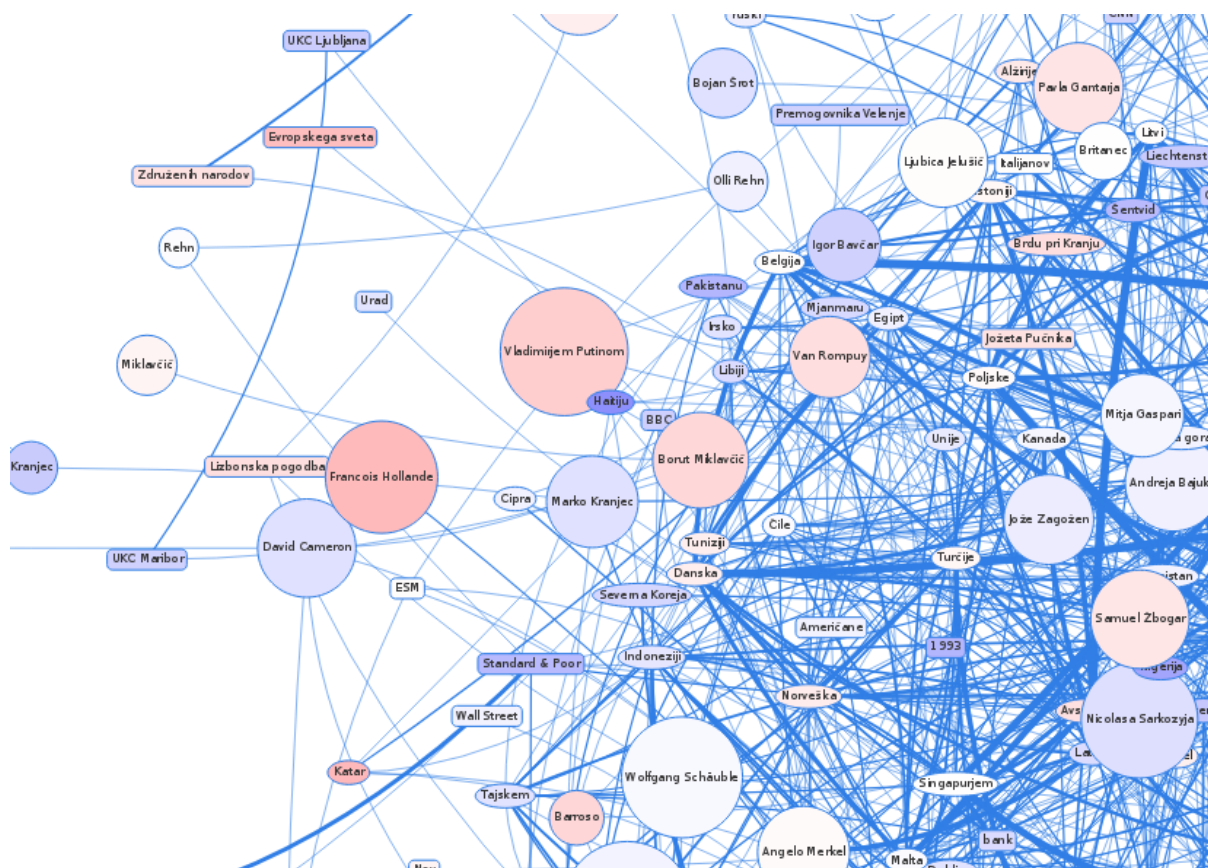


Figure 17: NetViz visualization of named entities' sentiment (red=positive, blue=negative, white=neutral).

In the future we could easily adapt this visualisation to be used for e.g. viewpoint analysis, where color codes would represent news sources, time periods, or news topics. The NetViz system could be integrated with our other visualisation systems. The NetViz graph representation could be linked to the temporal mosaic and/or concordance visualisations to enable filtering and interactive exploration of the source texts.

The two papers associated with this work are provided in Appendices F and G.

5 Associated Outputs

The work described in this deliverable has resulted in the following resources:

Description	URL	Availability
Code for CLTS	https://github.com/ElvysLPontes/Compressive-cross-language-text-summarisation	Private
Code for TeMoCo	https://github.com/sfermoy/TeMoCo	Public
Code for Metafacet	https://sourceforge.net/p/modnlp/plugins/code/ci/master/tree/Metafacet/	Public
Code for NetViz	https://github.com/vpodpecan/netviz	Public

Parts of this work are also described in detail in the following publications, which are attached to this deliverable as appendices:

Citation	Status	Appendix
Elvys Linhares Pontes, Stéphane Huet, Juan-Manuel Torres-Moreno, Thiago G. da Silva, and Andréa Carneiro Linhares (2020). "A Multilingual Study of Multi-Sentence Compression using Word Vertex-Labeled Graphs and Integer Linear Programming". In: <i>Computación y Sistemas</i> , 24(2).	To appear	Appendix A
Saturnino Luz and Shane Sheehan. "Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge". In: <i>Palgrave Communications</i> 6.1 (2020), pp. 1–20	Published	Appendix B
Shane Sheehan and Saturnino Luz. "Text Visualization for the Support of Lexicography-Based Scholarly Work". In: <i>Proceedings of the eLex 2019 conference on electronic lexicography in the 21st century</i> , Sintra, Portugal. 2019, pp. 694–725	Published	Appendix C

Parts of this work are also described in detail in the following publications (only partly related to EMBEDDIA).

Citation	Status	Appendix
Shane Sheehan, Pierre Albert, Masood Masoodian, and Saturnino Luz. "TeMoCo: A Visualization Tool for Temporal Analysis of Multi-party Dialogues in Clinical Settings". In: <i>2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)</i> . 2019, pp. 690–695	Published	Appendix D
Shane Sheehan, Masood Masoodian, Pierre Albert, and Saturnino Luz (2018). TeMoCo-Doc: A visualization for supporting temporal and contextual analysis of dialogues and associated documents. In: <i>Proceedings of the 2020 international conference on advanced visual interfaces</i>	Accepted	Appendix E
Senja Pollak, Vid Podpečan, Dragana Miljkovic, Uroš Stepišnik, and Špela Vintar (2020). The NetViz terminology visualization tool and the use cases in karstology domain modeling. In <i>Proceedings of the 6th International Workshop on Computational Terminology</i>	Published	Appendix F
Dragana Miljkovic, Jan Kralj, Uroš Stepišnik, and Senja Pollak (2019). Communities of related terms in Karst terminology co-occurrence network. In: <i>Proceedings of elex 2019</i>	Published	Appendix G

6 Conclusions and Future Work

In this report the work performed during the first year of Task 4.2 is presented. A novel method for the generation of cross-lingual summaries, from a corpus or single text, was created and evaluated. The technique was shown to outperform baseline results for low-resourced languages. A temporal visualisation which provides an overview of topic keywords was proposed and extended to visually link textual reports back to the temporal visualisation of the source material. A further concordance based visualisation approach to news summarisation was explored and a tool for interactive filtering of concordance views via metadata attributes was created. There are several steps planned for the future work. Combining the current textual and visual techniques developed in T4.2 is a priority. To achieve this, the generated textual summaries will be made interactive by linking the summary text with the documents or sections which most contributed to their generation via the temporal visualisation. The similar sentences which are used to generate a span of text will also be presented as a concordance thus allowing the frequency patterns and metadata distribution to be examined. We plan to apply these techniques to more of the EMBEDDIA datasets and languages. We will seek collaboration with our media partners to help evaluate and improve on the current technologies. For the textual summarisation subtask, we plan to improve the similarity measure by using a siamese neural network model (Linhares Pontes, Huet, Linhares, & Torres-Moreno, 2018) instead of the cosine similarity. We also will integrate developments from Tasks 4.1 (Real-time multilingual news linking) and 4.3 (Cross-lingual identification of viewpoints and sentiment in news reporting) into our textual and visual summaries.

References

- Bertin, J. (1983). *Semiology of Graphics*. University of Wisconsin Press.
- Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., ... Turchi, M. (2017, September). Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the second conference on machine translation, volume 2: Shared task papers* (pp. 169–214). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from <http://www.aclweb.org/anthology/W17-4717>
- Bostock, M., Ogievetsky, V., & Heer, J. (2011, December). D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2301–2309.
- Boudin, F., Huet, S., & Torres-Moreno, J. (2011). A Graph-based Approach to Cross-Language Multi-Document Summarization. *Polibits*, 43, 113–118.
- Brin, S., & Page, L. (1998, April). The Anatomy of a Large-scale Hypertextual Web Search Engine. *Comput. Netw. ISDN Syst.*, 30(1-7), 107–117.
- Bučar, J. (2017). *Manually sentiment annotated slovenian news corpus SentiNews 1.0*. Retrieved from <http://hdl.handle.net/11356/1110> (Slovenian language resource repository CLARIN.SI)
- de Caseli, H. M., Ramisch, C., das Graças Volpe Nunes, M., & Villavicencio, A. (2010, Apr 01). Alignment-based Extraction of Multiword Expressions. *Language Resources and Evaluation*, 44(1), 59–77. Retrieved from <https://doi.org/10.1007/s10579-009-9097-9> doi: 10.1007/s10579-009-9097-9
- Filippova, K. (2010). Multi-Sentence Compression: Finding Shortest Paths in Word Graphs. In *Coling* (p. 322-330).
- Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J., & Varma, V. (2011). TAC2011 MultiLing Pilot Overview. In *4th Text Analysis Conference TAC*.
- Hansen, K. R. (2016). News from the future: A corpus linguistic analysis of future-oriented, unreal and counterfactual news discourse. *Discourse & Communication*, 10(2), 115-136. Retrieved from <https://doi.org/10.1177/1750481315611240> doi: 10.1177/1750481315611240
- Javed, W., & Elmqvist, N. (2012). Exploring the design space of composite visualization. In *Pacific Visualization Symposium (PacificVis), 2012 IEEE* (pp. 1–8). doi: 10.1109/PacificVis.2012.6183556
- Karlsson, M., & Sjøvaag, H. (2018). *Rethinking research methods in an age of digital journalism*. Routledge.
- Leuski, A., Lin, C.-Y., Zhou, L., Germann, U., Och, F. J., & Hovy, E. (2003, September). Cross-lingual C*ST*RD: English Access to Hindi Information.. Retrieved from <https://www.microsoft.com/en-us/research/publication/cross-lingual-cstrd-english-access-to-hindi-information/>
- Lin, C.-Y. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)* (p. 74-81).
- Linhares Pontes, E. (2018). *Compressive Cross-Language Text Summarization* (Doctoral dissertation). Retrieved from <http://www.theses.fr/2018AVIG0232>
- Linhares Pontes, E., Huet, S., Linhares, A. C., & Torres-Moreno, J.-M. (2018). Predicting the Semantic Textual Similarity with Siamese CNN and LSTM. In *25e Conférence sur le Traitement Automatique des Langues Naturelles*.
- Linhares Pontes, E., Huet, S., Torres-Moreno, J.-M., da Silva, T. G., & Linhares, A. C. (2020). A multilingual study of multi-sentence compression using word vertex-labeled graphs and integer linear programming. *Computación y Sistemas*, 24(2).
- Linhares Pontes, E., Huet, S., Torres-Moreno, J.-M., & Linhares, A. C. (2018). Cross-Language Text Summarization Using Sentence and Multi-Sentence Compression. In M. Silberstein, F. Atigui,

- E. Kornysheva, E. Métais, & F. Meziane (Eds.), *Natural Language Processing and Information Systems* (pp. 467–479). Cham: Springer International Publishing.
- Linhares Pontes, E., Huet, S., Torres-Moreno, J.-M., & Linhares, A. C. (2020). Compressive approaches for cross-language multi-document summarization. *Data & Knowledge Engineering*, 125, 101763. doi: <https://doi.org/10.1016/j.datak.2019.101763>
- Luz, S. (2000, May). A Software Toolkit for Sharing and Accessing Corpora Over the Internet. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Stainhauer (Eds.), *Proceedings of the second international conference on language resources and evaluation: LREC-2000* (pp. 1749–1754).
- Luz, S. (2011). Web-based corpus software. In A. Kruger, K. Wallmach, & J. Munday (Eds.), *Corpus-based Translation Studies – Research and Applications* (pp. 124–149). Continuum.
- Luz, S., & Masoodian, M. (2005, Jan). A Model for Meeting Content Storage and Retrieval. In *Proceedings of the 11th international multimedia modelling conference* (pp. 392–398). doi: 10.1109/MMMC.2005.12
- Luz, S., & Masoodian, M. (2007, July). Visualisation of Parallel Data Streams with Temporal Mosaics. In *Proceedings of the 11th International Conference Information Visualization* (pp. 197–202). doi: 10.1109/IV.2007.127
- Luz, S., & Sheehan, S. (2014). A Graph Based Abstraction of Textual Concordances and Two Renderings for their Interactive Visualisation. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 293–296). New York, NY, USA: ACM. doi: 10.1145/2598153.2598187
- Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2), 110–141. doi: <http://doi.acm.org/10.1145/22949.22950>
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations* (pp. 55–60).
- Miljkovic, D., Kralj, J., Stepišnik, U., & Pollak, S. (2019). Communities of related terms in Karst terminology co-occurrence network. In *Proceedings of eLex 2019* (pp. 357–373).
- Moens, M.-F., Angheluta, R., Mitra, R., & Jing, X. (2004). *K.U.Leuven summarization system at DUC 2004*. Retrieved from <https://lirias.kuleuven.be/handle/123456789/135440> (Document Understanding Conference, Boston, 2004)
- Moirón, B. V., & Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. In *Eacl 2006 workshop on multiword expressions in a multilingual context*.
- Orasan, C., & Chiorean, O. A. (2008). Evaluation of a Cross-lingual Romanian-English Multi-document Summariser. In *6th international conference on language resources and evaluation (lrec)*.
- Pollak, S., Podpečan, V., Miljkovic, D., Stepišnik, U., & Vintar, Š. (2020, May). The NetViz terminology visualization tool and the use cases in karstology domain modeling. In *Proceedings of the 6th international workshop on computational terminology* (pp. 55–61). Marseille, France: European Language Resources Association. Retrieved from <https://www.aclweb.org/anthology/2020.computerm-1.8>
- Pontes, E. L., Huet, S., & Torres-Moreno, J. (2018). A Multilingual Study of Compressive Cross-Language Text Summarization. In *Advances in Computational Intelligence - 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22-27, 2018, Proceedings, Part II* (pp. 109–118). doi: 10.1007/978-3-030-04497-8_9
- Saggion, H., & Lapalme, G. (2002, December). Generating Indicative-informative Summaries with sumUM. *Comput. Linguist.*, 28(4), 497–526. Retrieved from <http://dx.doi.org/10.1162/089120102762671963> doi: 10.1162/089120102762671963
- Sheehan, S., Masoodian, M., & Luz, S. (2018). COMFRE: A Visualization for Comparing Word Frequencies in Linguistic Tasks. In *Proceedings of the 2018 International Conference on Advanced Visual*

Interfaces (pp. 42:1–42:5). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/3206505.3206547> doi: 10.1145/3206505.3206547

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings the IEEE Symposium on Visual Languages* (pp. 336–343). doi: 10.1109/VL.1996.545307

Wan, X. (2011). Using Bilingual Information for Cross-Language Document Summarization. In *ACL* (pp. 1546–1555). The Association for Computer Linguistics.

Wan, X., Li, H., & Xiao, J. (2010). Cross-Language Document Summarization Based on Machine Translation Quality Prediction. In *Acl*.

Wan, X., Luo, F., Sun, X., Huang, S., & Yao, J.-g. (n.d.). Cross-Language Document Summarization via Extraction and Ranking of Multiple Summaries. *Knowledge and Information Systems*.

Yao, J., Wan, X., & Xiao, J. (2015). Phrase-based Compressive Cross-Language Summarization. In *EMNLP* (pp. 118–127).

Zhang, J., Zhou, Y., & Zong, C. (2016). Abstractive Cross-Language Summarization via Translation Model Enhanced Predicate Argument Structure Fusing. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(10), 1842–1853.

Appendix A: A Multilingual Study of Multi-Sentence Compression using Word Vertex-Labeled Graphs and Integer Linear Programming

A Multilingual Study of Multi-Sentence Compression using Word Vertex-Labeled Graphs and Integer Linear Programming

Elvys Linhares Pontes^{1,2,3}, Stéphane Huet², Juan-Manuel Torres-Moreno^{2,3},
Thiago G. da Silva^{4,5}, Andréa Carneiro Linhares⁶

¹ L3i, University of La Rochelle, La Rochelle, France

² LIA, University of Avignon, Avignon, France

³ Polytechnique Montréal, Montréal, Canada

⁴ Inst. Federal de Educação, Ciência e Tecnologia da Paraíba, PB, Brazil

⁵ Instituto de Computação Univ. Federal Fluminense, RJ, Brazil

⁶ Universidade Federal do Ceará, Sobral, Brazil

elvys.linhares.pontes@univ-lr.fr

Abstract. Multi-Sentence Compression (MSC) aims to generate a short sentence with the key information from a cluster of similar sentences. MSC enables summarization and question-answering systems to generate outputs combining fully formed sentences from one or several documents. This paper describes an Integer Linear Programming method for MSC using a vertex-labeled graph to select different keywords, with the goal of generating more informative sentences while maintaining their grammaticality. Our system is of good quality and outperforms the state of the art for evaluations led on news datasets in three languages: French, Portuguese and Spanish. We led both automatic and manual evaluations to determine the informativeness and the grammaticality of compressions for each dataset. In additional tests, which take advantage of the fact that the length of compressions can be modulated, we still improve ROUGE scores with shorter output sentences.

Keywords. Multi-Sentence Compression, Integer Linear Programming, Word Graph.

1 Introduction

A considerable amount of information is published in various sites every day, e.g. comments, photos, videos and audio in different languages. The increased number of electronic devices (smartphones, tablets, etc.) have made access to these information easier and faster. Moreover,

websites such as Wikipedia or news aggregators can provide detailed data on various issues but texts may be long and convey a lot of information. Readers, besides not having the time to go through this amount of information, are not interested in all the proposed subjects and generally select the content of their interest. One solution to this problem is the generation of summaries containing only the key information.

Among the various applications of Natural Language Processing (NLP), Automatic Text Summarization (ATS) aims to automatically identify the relevant data inside one or more documents, and create a condensed text with the main information [21]. At the same time, summaries should be short with as little redundant information as possible. Summarization systems usually rely on statistical, morphological and syntactic analysis approaches [37]. Some of them use Multi-Sentence Compression (MSC) in order to produce from a set of similar sentences a small-sized sentence which is both grammatically correct and informative [1, 10, 21]. Although compression is a challenging task, it is appropriate to generate summaries that are more informative than the state-of-the-art extractive methods for ATS.

The contributions of this article are two-fold. (i) We improved the model for MSC [16] that extends

the common approach based on Graph Theory, using vertex-labeled graphs and Integer Linear Programming (ILP) to select the best compression. The vertex-labeled graphs¹ are used to model a cluster of similar sentences with keywords. (ii) Whereas previous work usually limited the experimental study on one or two datasets, we tested our model on three corpora, each in a different language. Evaluations led with both automatic metrics and human evaluations show that our ILP model consistently generate more informative sentences than two state-of-the-art systems while maintaining their grammaticality. Interestingly, our approach is able to choose the amount of information to keep in the compression output, through the definition of the maximum compression length.

This paper is organized as follows: we describe and survey the MSC problem in Section 2. Next, we detail our approach in Section 3. The experiments and the results are discussed in Sections 4 and 5. Lastly, conclusions and some final comments are set out in Section 6.

2 Related Work

Sentence Compression (SC) aims at producing a reduced grammatically correct sentence. Compressions may have different Compression Ratio (CR) levels,² whereby the lower the CR level, the higher the reduction of the information is. SC can be employed in the contexts of the summarization of documents, the generation of article titles or the simplification of complex sentences, using diverse methods such as optimization [7, 8], syntactic analysis, deletion of words [11] or generation of sentences [25, 32]. Recently, many SC approaches using Neural Network (NN) have been developed [25, 32]. These methods may generate good results for a single sentence because they combine many complex structures such as recurrent neural networks (based on Gated Recurrent Units and Long Short Term Memory),

the sequence-to-sequence paradigm and condition mechanisms (e.g., attention). However, these composite neural networks need huge corpora to learn how to generate compressions (e.g., Rush et al. used the Gigaword corpus that contains around 9.5 million news) and take a lot of time to accomplish the learning process.

Multi-Sentence Compression (MSC), also coined as Multi-Sentence Fusion, is a variation of SC. Unlike SC, MSC combines the information of a cluster of similar sentences to generate a new sentence, hopefully grammatically correct, which compresses the most relevant data of this cluster. The idea of MSC was introduced by Barzilay and McKeown [3], who developed a multi-document summarizer which represents each sentence as a dependency tree; their approach aligns and combines these trees to fusion sentences. Filippova and Strube [12] also used dependency trees to align each cluster of related sentences and generated a new tree, this time with ILP, to compress the information. In 2010, Filippova presented a new model for MSC, simple but effective, which is based on Graph Theory and a list of stopwords. She used a Word Graph (WG) to represent and to compress a cluster of related sentences; the details of this model, which is extended by the work of this paper, can be found in Section 2.1.

Inspired by the good results of the Filippova's method, many studies have used it in a first step to generate a list of the N shortest paths, then have relied on different reranking strategies to analyze the candidates and select the best compression [1, 5, 23, 38]. Boudin and Morin [5] developed a reranking method measuring the relevance of a candidate compression using *key phrases*³, obtained with the TextRank algorithm [26], and the length of the sentence. Another reranking strategy was proposed by Luong et al. [23]. Their method ranks the sentences from the counts of unigrams⁴ occurring in every source sentence. ShafieiBavani et al. [34] also used a WG model; their approach consists of three main components: (i) a merging

¹A vertex-labeled graph means a graph where each node has a label. In this work, a label is represented by a color and different nodes can have the same label.

²The CR is the length of the compression divided by the average length of all source sentences

³*key phrases* are words that capture the main topics of a document.

⁴An n -gram is a contiguous sequence of n items from a given text.

stage based on Multiword Expressions (MWE), (ii) a mapping strategy based on synonymy between words and (iii) a reranking step to identify the best compression candidates generated using a Part-of-Speech-based language model (POS-LM). Tzouridis et al. [38] proposed a structured learning-based approach. Instead of applying heuristics as Filippova [10], they adapted the decoding process to the data by parameterizing a shortest path algorithm. They devised a structural support vector machine to learn the shortest path in possibly high dimensional joint feature spaces and proposed a generalized loss-augmented decoding algorithm that is solved exactly by ILP in polynomial time.

Linhares Pontes et al. [16] also presented an ILP approach that models a set of similar sentences as vertex-labeled word graphs. Their approach selects keywords and relevant 3-grams to generate more informative compressions while maintaining their grammaticality as possible. They have studied the quality of compressions by analyzing different amounts of keywords in order to manage both the length and the informativeness of compressions.

We found two other studies that applied ILP to combine and compress several sentences. Banerjee et al. [1] developed a multi-document ATS system that generated summaries after compressing similar sentences. They used Filippova's method to generate 200 random compressed sentences. Then they created an ILP model to select the most informative and grammatically correct compression. Thadani and McKeown [36] proposed another ILP model using an inference approach for sentence fusion. Their ILP formulation relies on n-gram factorization and aims at avoiding cycles and disconnected structures.

In the ATS task, Shang et al. [35] adapted the Boudin and Morin's approach [5] to take into account the grammaticality for the reranking of compressions. Instead of the TextRank algorithm, they analyze the spreading influence in WG to generate more informative and grammatical compressions and to improve the quality of summaries. Nayeem et al. [28] designed a paraphrastic sentence fusion model which jointly performs sentence fusion and paraphrasing using

skip-gram word embedding model at the sentence level.

Recently, Zhao et al. [39] presented an unsupervised rewriter to improve the grammaticality of MSC outputs while introducing new words. They used the WG approach to produce coarse-grained compressions, from which they substitute words with their shorter synonyms to yield paraphrased sentence. Then, their neural rewriter proposes paraphrases for these compressions in order to improve grammaticality and encourage more novel words.

Another related task is the sentence aggregation that combines a group of sentences, not necessarily with a similar semantic content, to generate a single sentence (e.g., "*The car is here.*" and "*It is blue.*" can be aggregated into "*The blue car is here.*"). This aggregation can be at semantic and syntactic levels [31]. The aggregation rules can be acquired automatically from a corpus [2]. However, this process is not possible for all situations and the sentence aggregation depends on the sentence planning to combine the sentences.

Following previous studies for MSC that rely on Graph Theory with good results, this work presents a new ILP framework that takes into account keywords for MSC. We compare our learning approach to the graph-based sentence compression techniques proposed by Filippova [10] and Boudin and Morin [5], considered as state-of-the-art methods for MSC. We intend to apply our method on various languages and not to be dependent on linguistic resources or tools specific to languages. This led us to put aside systems which, despite being competitive, rely on resources like WordNet or Multiword expression detectors [34]. Since we borrowed concepts and ideas from Filippova's method, we detail her approach in the next section.

2.1 Filippova's Method

Filippova [10] modeled a document D containing n similar sentences $\{s_1, s_2, \dots, s_n\}$, as a directed word graph $G = (V, A)$. V is the set of vertices (words) and A is the set of arcs (adjacency relationship). Figure 1 illustrates the word graph of the following Portuguese sentences:

1. *George Solitário, a última tartaruga gigante Pinta Island do mundo, faleceu.* (Lonesome George, the world's last Pinta Island giant tortoise, has passed away.)
2. *A tartaruga gigante conhecida como George Solitário morreu domingo no Parque Nacional de Galapagos, Equador.* (The giant tortoise known as Lonesome George died Sunday at the Galapagos National Park in Ecuador.)
3. *Ele tinha apenas cem anos de vida, mas a última tartaruga gigante Pinta conhecida, George Solitário, faleceu.* (He was only about a hundred years old, but the last known giant Pinta tortoise, Lonesome George, has passed away.)
4. *George Solitário, a última tartaruga gigante da sua espécie, morreu.* (Lonesome George, a giant tortoise believed to be the last of his kind, has died.)

The initial graph G is composed of the first sentence (1) and the vertices –begin– and –end–. For a new sentence, a new vertex is created when a word/POS pair cannot be matched to an existing vertex of G once lowercased. Besides, at most one occurrence of a given word/POS inside a sentence can be associated with a given vertex.

Sentences are individually analyzed and added to G . Each sentence represents a simple path between the –begin– and –end– vertices and its words are inserted in the following order:

1. Non-stopwords for which no candidate exists in the graph or for which an unambiguous mapping is possible;
2. Non-stopwords for which there are several possible candidates in the graph that may occur more than once in the sentence;
3. Stopwords.

In cases 2 and 3, the word mapping is ambiguous because there is more than one vertex in the graph that references the same word/POS. In this case, we analyze the immediate context (the preceding and following words/POSS in the sentence and the neighboring nodes in the graph)

or the frequency (i.e., the number of words that were mapped to the considered vertex) to select the best candidate node.

Once vertices have been added, arcs are valued by weights which represent the levels of cohesion between two words in the graph (Equation 1). Cohesion is calculated from the frequency and the position of these words in sentences, according to Equation 2:

$$w(i, j) = \frac{\text{cohesion}(i, j)}{\text{freq}(i) \times \text{freq}(j)}, \quad (1)$$

$$\text{cohesion}(i, j) = \frac{\text{freq}(i) + \text{freq}(j)}{\sum_{s \in D} \text{diff}(s, i, j)^{-1}}, \quad (2)$$

where $\text{freq}(i)$ is the word frequency mapped to the vertex i and the function $\text{diff}(s, i, j)$ refers to the distance between the offset positions of words i and j in the sentences s of D containing these two words.

From the graph G , the system calculates the 50 shortest paths that are longer than eight words and have at least one verb. Finally, the system reranks the paths by normalizing the total path weight over their length and selects the path with the lowest score as the best MSC.

3 Our Approach

Filippova's method chooses the path with the lowest score taking into account the level of cohesion between two adjacent words in the document. However, two words with a strong cohesion do not necessarily have a good informativeness because the cohesion only measures the distance and the frequency of words in the sentences. In this work, we propose a method to concurrently analyze cohesion and keywords in order to generate a more informative and comprehensible compression.

Our method calculates the shortest path from the cohesion of words and grants bonuses to the paths that have different keywords. For this purpose, our approach is based on Filippova's method (Section 2.1) to model a document D as a graph and to calculate the cohesion of words. In addition, we analyze the keywords of the document to favor hypotheses with meaningful information.

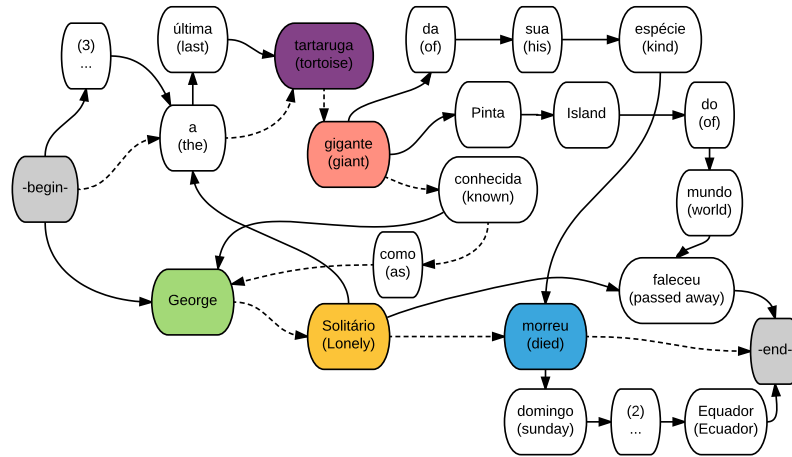


Fig. 1. WG generated from the sentences (1) to (4) (without the punctuation and Part-of-Speech (POS) for easy readability). The dotted path represents the best compression for this WG and the colored vertices represent the keywords of the document.

3.1 Keyword Extraction

Introducing keywords in the graph helps the system to generate more informative compressions because it takes into account the words that are representative of the cluster to calculate the best path in the graph, and not only the cohesion and frequency of words. Keywords can be identified for each cluster with various extraction methods and we study three widely used techniques: Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA) and TextRank. Despite the small number of sentences per cluster, these methods generate good results because clusters are composed of similar sentences with a high level of redundancy. LSI uses Singular-Value Decomposition (SVD), a technique closely related to eigenvector decomposition and factor analysis, to model the associative relationships [9]. LDA is a topic model that generates topics based on word frequency from a set of documents [4]. Finally, TextRank algorithm analyzes the words in texts using WGs and estimates their relevance [26]. For LDA whose modeling is based on the concept of topics, we consider that the document D describes only one topic since it is composed of semantically close sentences related to a specific news item. A

same word or keyword can be represented by one or several nodes in WGs (see Section 2.1). In order to prioritize the sentence generation containing multiple keywords and to reduce the redundancy, we add a bonus to the compression score when the compression contains different keywords.

3.2 Vertex-Labeled Graph

A vertex-labeled graph is a graph $G = (V, A)$ with a label on the vertices $K = \{0, \dots, |K|\}$, where $|K|$ is the number of different labels. This graph type has been employed in several domains such as biology [40] or NLP [6]. In this last study, the correction of Wikipedia inter-language links was modeled as a Colorful Components problem. Given a vertex-colored graph, the Colorful Components problem aims at finding the minimum-size edge sets that are connected and do not have two vertices with the same color.

In the context of MSC, we want to generate a short informative compression where keyword may be represented by several nodes in the word graph. Labels enable us to represent keywords in vertex-labeled graphs and generate a compression without repeated keywords while preserving the informativeness. In this framework, we grant

bonuses only once for nodes with the same label to prioritize new information in the compression (Figure 1). To make our model coherent, we added a base label (label 0) for all non-keywords in the word graph. The following section describes our ILP model to select sentences including labeled keywords inside WGs.

3.3 ILP Modeling

There are several algorithms with a polynomial complexity to find the shortest path in a graph. However, the restriction on the minimum number P_{\min} of vertices (i.e., the minimum number of words in the compression) makes the problem NP-hard. Indeed, let v_0 be the –begin– vertex. If P_{\min} equals $|V|$ and if we add an auxiliary arc from –end– vertex to v_0 , our problem is similar to the Traveling Salesman Problem (TSP), which is NP-hard.

For this work we use the formulation known as Miller-Tucker-Zemlin (MTZ) to solve our problem [30, 36]. This formulation uses a set of auxiliary variables, one for each vertex in order to prevent a vertex from being visited more than once in the cycle and a set of arc restrictions.

The problem of production of a compression that favors informativeness and grammaticality is expressed as Equation 3. In other words, we look for a path (sentence) that has a good cohesion and contains a maximum of labels (keywords).

$$\text{Minimize } \left(\sum_{(i,j) \in A} w(i,j) \cdot x_{i,j} - c \cdot \sum_{k \in K} b_k \right) \quad (3)$$

where x_{ij} indicates the existence of the arc (i,j) in the solution, $w(i,j)$ is the cohesion of the words i and j (Equation 1), K is the set of labels (each representing a keyword), b_k indicates the existence of a word with label (keyword) k in the solution and c is the keyword bonus of the graph.⁵

⁵The keyword bonus allows the generation of longer compressions that may be more informative.

3.4 Structural Constraints

We describe the structural constraints for the problem of consistency in compressions and define the bounds of the variables. First, we consider the problem of consistency which requires an inner and an outer arc active for every word used in the solution, where y_v indicates the existence of the vertex v in the solution.

$$\sum_{i \in \delta^+(v)} x_{vi} = y_v \quad \forall v \in V, \quad (4)$$

$$\sum_{i \in \delta^-(v)} x_{iv} = y_v \quad \forall v \in V. \quad (5)$$

The constraints (6) and (7) control the minimum and the maximum number of vertices (P_{\min} and P_{\max}) used in the solution respectively, i.e., the minimum and the maximum number of words in the final compression.

$$\sum_{v \in V} y_v \geq P_{\min}, \quad (6)$$

$$\sum_{v \in V} y_v \leq P_{\max}. \quad (7)$$

The set of constraints (8) matches label variables (keywords) with vertices (words), where $V(k)$ is the set of all vertices with label k .

$$\sum_{v \in V(k)} y_v \geq b_k, \quad \forall k \in K. \quad (8)$$

Equality (9) sets the vertex v_0 in the solution.

$$y_0 = 1. \quad (9)$$

The restrictions (10) and (11) are responsible for the elimination of sub-cycles, where u_v ($\forall v \in V$) are auxiliary variables for the elimination of sub-cycles and M is a large number (e.g., $M = |V|$).

$$u_0 = 1, \quad (10)$$

$$u_i - u_j + 1 \leq M - M \cdot x_{ij} \quad \forall (i, j) \in A, j \neq 0. \quad (11)$$

Finally, equations (12) – (14) define the field of variables.

$$x_{ij} \in \{0, 1\}, \quad \forall (i, j) \in A, \quad (12)$$

$$y_v \in \{0, 1\}, \quad \forall v \in V, \quad (13)$$

$$u_v \in \{1, 2, \dots, |V|\}, \quad \forall v \in V. \quad (14)$$

We calculate the 50 best solutions according to the objective (3) having at least eight words and at least one verb. Specifically, we find the best solution, then we add a constraint in the model to avoid this solution and repeat this process 50 times to find the other solutions.

The optimized score (Equation 3) explicitly takes into account the size of the generated sentence. Contrary to Filippova's method, sentences may have a negative score because we subtract from the cohesion value of the path the introduced scores for keywords. Therefore, we use the exponential function to ensure a score greater than zero. Finally, we select the sentence with the lowest final score (Equation 15) as the best compression.

$$\text{score}_{\text{norm}}(s) = \frac{e^{\text{score}_{\text{opt}}(s)}}{\|s\|}, \quad (15)$$

where $\text{score}_{\text{opt}}(s)$ is the score of the sentence s from Equation 3.

4 Experimental Setup

Algorithms were implemented using the Python programming language with the `takahe`⁶ and `gensim`⁷ libraries. The mathematical model was implemented in C++ with the `Concert` library and we used the solver `CPLEX 12.6`⁸.

⁶<http://www.florianboudin.org/publications.html>

⁷<https://radimrehurek.com/gensim/models/ldamodel.html>

⁸<https://www.ibm.com/products/ilog-cplex-optimization-studio>

The objective function (see Equation 3) involves a keyword bonus. Since each WG can have weight arcs of different values, fixing this bonus is decisive to allow the generation of slightly longer compressions. We tested several metrics (fixed values, the arithmetic average, the median, and the geometric average of the weights arcs of WG) to define the keyword bonus of the WG and empirically found that geometric mean outperformed others.

4.1 Evaluation Datasets

Various corpora have been developed for MSC and are composed of clusters of similar sentences from different source news in English, French, Portuguese, Spanish or Vietnamese languages. Whereas the data built by McKeown et al. [24] and Luong et al. [23] have clusters limited to pairs of sentences, the corpora made by Filippova [10], Boudin and Morin [5], and Linhares Pontes et al. [22] contain clusters of at least 7 similar sentences. McKeown et al. [24] collected 300 English sentence pairs taken from newswire clusters using Amazon's Mechanical Turk, while the corpus introduced in Luong et al. [23] is made of 250 Vietnamese sentences divided into 115 groups of similar sentences with 2 sentences by group. McKeown et al. [24], Luong et al. [23], Boudin and Morin [5], and Linhares Pontes et al. [22] made their corpora publicly available, but only the data associated with these last two articles are more suited to the multi-document summarization or question-answering tasks because the documents to analyze are usually composed of many similar sentences. Therefore, we use these two corpora made of French, Portuguese and Spanish sentences.

Table 1 summarizes the statistics of this set of data having 40 clusters of sentences for each language. The Type-Token Ratio (TTR) indicates the reuse of tokens in a cluster and is defined by the number of unique tokens divided by the number of tokens in each cluster; the lower the TTR, the greater the reuse of tokens in the cluster. The

sentence similarity represents the average cosine similarity of the sentences in a cluster.⁹

The French corpus has 3 sentences compressed by native speakers for each cluster, references having a compression rate (CR) of 60%. Like the French corpus, the Portuguese and Spanish corpora are composed of the first sentences of the articles found in Google News. Each cluster is composed of related sentences and was chosen among the first sentence from different articles about Science, Sport, Economy, Health, Business, Technology, Accidents/Catastrophes, General Information and other subjects. A cluster has at least 10 similar sentences by topic and 2 reference compressions made by different native speakers. The average CRs are 54% and 61% for the Portuguese and the Spanish corpora, respectively.

The three languages derive from Latin and are closely related languages. However, they differ in many details of their grammar and lexicon. Moreover, the datasets produced for the three languages are unlike according to several features. First, the corpus made by Linhares Pontes et al. [22] contains a smaller (Portuguese corpus) and a larger (Spanish corpus) dataset in terms of sentences than the French corpus. Besides, the compression rates of the three datasets indicate that the Portuguese source sentences have more irrelevant tokens. The sentence similarity (Table 1, second last line) describes the variability of sentences in the source sentences and in the references, and reflects here that the sentences are slightly more diverse for the French corpus. This translates into a higher TTR observed for the French part (38.8%) than for the two other languages (33.7% and 35.2%).

4.2 Automatic and Manual Evaluations

The most important features of MSC are informativeness and grammaticality. Informativeness measures how informational is the generated text. As references are assumed to contain the key information, we calculated informativeness scores

⁹The cosine similarity between two vectors u and v associated with two sentences is defined by $\frac{u \cdot v}{\|u\| \|v\|}$ in the $[0,1]$ range.

counting the n-grams in common between the system output and the reference compressions using ROUGE [14]. In particular, we used the metrics ROUGE-1 and ROUGE-2, F-measure being preferred to recall for a fair comparison of various lengths of compressed sentences. Like in [5], ROUGE metrics are calculated with stopwords removal and stemming¹⁰.

Due to limitations of the ROUGE systems that only analyze unigrams and bigrams, we also led a manual evaluation with four native speakers for French, Portuguese and Spanish. The native speakers of each language evaluated the compression in two aspects: informativeness and grammaticality. In the same way as Filippova [10] as well as Boudin and Morin [5], the native speakers evaluated the grammaticality in a 3-point scale: 2 points for a correct sentence; 1 point if the sentence has minor mistakes; 0 point if it is none of the above. Like grammaticality, informativeness is evaluated in the same range: 2 points if the compression contains the main information; 1 point if the compression misses some relevant information; 0 point if the compression is not related to the main topic.

5 Experimental Assessment

Compression rates are strongly correlated with human judgments of meaning and grammaticality [27]. On the one hand, too short compressions may compromise sentence structure, reducing the informativeness and grammaticality. On the other hand, longer compressions may be more interesting for ATS when informativeness and grammaticality are decisive features. Consequently, we analyze compression with multiple maximum compression lengths (50%, 60%, 70%, 80%, 90% and ∞ , the last value meaning that no constraint is fixed on the output size).

Following the idea proposed by ShafieiBavani et al. [34] and already implemented with success in other domains such as speech recognition (e.g., [13]), we tested the use of a POS-based Language Model (POS-LM) as a post-processing stage in order to improve the grammaticality of

¹⁰<http://snowball.tartarus.org/>

Table 1. Statistics of the corpora.

Characteristics	French		Portuguese		Spanish	
	Source	References	Source	References	Source	References
#tokens	20,224	2,362	17,998	1,425	30,588	3,694
#vocabulary (tokens)	2,867	636	2,438	533	4,390	881
#sentences	618	120	544	80	800	160
avg. sentence length (tokens)	33.0	19.7	33.1	17.8	38.2	23.1
type-token ratio (TTR)	39%	50%	34%	68%	35%	43%
sentence similarity	0.46	0.67	0.51	0.59	0.47	0.64
compression rate	—	60%	—	54%	—	61%

compressions. Specifically, for each cluster, the ten best compressions according to our optimized score are reranked by a 7-gram POS-LM trained with the SRILM toolkit¹¹ on the French, Portuguese and Spanish parts of the Europarl dataset,¹² tagged with TreeTagger [33].

5.1 Results

Since our method strongly depends on the set of keywords to generate informative compressions, we investigate the performance of the three keyword methods (LDA, LSI and TextRank), selecting the 5 or 10 most relevant words. We verified the percentage of keywords generated by these methods that are included in the reference compression (Table 2). A significantly higher rate of keywords in the references is observed when using LDA or LSI instead of TextRank. In order to obtain the most relevant words in a cluster with different sizes, we used LDA in our final MSC system to identify 10 keywords for each cluster.

Tables 3, 4 and 5 describe the ROUGE recall scores measured for Filippova's [10] method (named F10), Boudin and Morin's [5] method (named BM13) and our method with multiple maximum compression lengths. As for each CR setup the size of the outputs to evaluate are comparable, the recall scores are preferred in this case to measure the information retained in compressions. First, let us note that CRs

Table 2. Percentage of keywords included in the reference compression for French, Portuguese and Spanish corpora.

Methods	fr	pt	es
LDA: 5 kws	91%	88%	85%
LSI: 5 kws	90%	87%	81%
TextRank: 5 kws	69%	55%	58%
LDA: 10 kws	84%	70%	76%
LSI: 10 kws	84%	69%	73%
TextRank: 10 kws	56%	44%	50%

effectively observed may differ from the fixed value of P_{max} . For example, a 50% threshold leads to real CRs of 38% to 40% for all languages, while an 80% level creates new sentences with real CRs between 53% and 60%. Interestingly, our system obtained better ROUGE recall scores than both baselines in all languages for comparable compression lengths. If we prioritize meaning, our method with no explicit constraint on the maximum compression length (ILP: ∞) improved the compression quality with a small increase of the compression length (compression ratio between 55.4% and 65.9%). Instead, we can limit the length and generate compressions that are shorter and have still better ROUGE scores than the baselines.

Based on these results, a further analysis was done for the 80% and ∞ configurations.

¹¹<http://www.speech.sri.com/projects/srilm/>

¹²<http://www.statmt.org/europarl/>

Table 3. ROUGE recall scores for multiple maximum compression lengths using the French corpus.

French			
Methods	ROUGE-1	ROUGE-2	CR
F10	0.5971	0.4072	51.3%
BM13	0.6740	0.4695	59.8%
ILP:50%	0.4763	0.3039	39.1%
ILP:60%	0.5990	0.4101	47.4%
ILP:70%	0.6420	0.4206	53.5%
ILP:80%	0.6783	0.4573	60.0%
ILP:90%	0.6981	0.4758	61.8%
ILP: ∞	0.7010	0.4751	62.6%

Table 4. ROUGE recall scores for multiple maximum compression lengths using the Portuguese corpus.

Portuguese			
Methods	ROUGE-1	ROUGE-2	CR
F10	0.5354	0.2935	52.2%
BM13	0.6304	0.3493	69.1%
ILP:50%	0.4689	0.2521	40.0%
ILP:60%	0.5369	0.2967	48.1%
ILP:70%	0.5652	0.3088	54.0%
ILP:80%	0.6056	0.3321	59.0%
ILP:90%	0.6341	0.3492	64.6%
ILP: ∞	0.6407	0.3546	65.9%

Table 6¹³ describes the results for the French, Portuguese and Spanish corpora using ROUGE F-measure scores. The first two columns display the evaluation of the two baseline systems; the ROUGE scores measured with our method using either 80% or ∞ maximum compression lengths are shown in the next two columns and the last two columns respectively. The outputs produced by all of these systems for two sample clusters in Spanish and Portuguese

¹³Although we used the same system and data as Boudin and Morin [5] for the French corpus, we were not able to exactly reproduce their results. The ROUGE F-measure scores given in their article are close to ours for their system: 0.6568 (ROUGE-1), 0.4414 (ROUGE-2) and 0.4344 (ROUGE-SU4), but using F10 we measured higher scores than them: 0.5744 (ROUGE-1), 0.3921 (ROUGE-2) and 0.3700 (ROUGE-SU4).

Table 5. ROUGE recall scores for multiple maximum compression lengths using the Spanish corpus.

Spanish			
Methods	ROUGE-1	ROUGE-2	CR
F10	0.4437	0.2631	43.2%
BM13	0.5167	0.2981	61.2%
ILP:50%	0.3814	0.1990	38.7%
ILP:60%	0.4594	0.2651	45.3%
ILP:70%	0.5050	0.2922	50.2%
ILP:80%	0.5191	0.2982	53.2%
ILP:90%	0.5242	0.2982	54.4%
ILP: ∞	0.5305	0.3036	55.4%

can be found in the Appendix. Globally, all versions of our ILP method outperform both baselines according to ROUGE F-measures for the Portuguese and Spanish corpora, and our ILP systems (ILP:80% and ILP: ∞) obtained similar results to BM13 for the French corpus. The POS-LM post-processing further improved the ROUGE scores for Portuguese and Spanish, but unfortunately not for the French corpus.

Table 7 displays the average length, the compression ratio and the average number of keywords that are kept in the final compression. F10 generated the shortest compressions for all corpora, our approach producing outputs of an intermediate length with respect to BM13, except for the French corpus for which ILP: ∞ generated slightly longer compressions. The keyword bonus and the POS-LM score act differently on the selection of words. On the one hand, the keyword bonus promotes the integration of keywords from difference sentences. On the other hand, the POS-LM favors grammaticality and longer subsequences of the original sentences, which reduces the mix of sentences and, consequently, the number of keywords in the compressions.

We also led a manual evaluation to study the informativeness and grammaticality of compressions. We measured the inter-rater agreement on the judgments we collected, obtaining values of Fleiss' kappa of 0.423, 0.289 and 0.344 for French, Portuguese and Spanish respectively. These results show that human evaluation is rather subjective. Questioning evaluators on how they

Table 6. ROUGE F-measure results on the French, Portuguese and Spanish corpora. The best ROUGE results are in bold.

Metrics	Methods					
	F10	BM13	ILP:80%	ILP:80%+LM	ILP:∞	ILP:∞+LM
French						
ROUGE-1	0.6384	0.6674	0.6630	0.6418	0.6730	0.6460
ROUGE-2	0.4423	0.4672	0.4487	0.4187	0.4567	0.4179
ROUGE-SU4	0.4297	0.4602	0.4410	0.4152	0.4511	0.4136
Portuguese						
ROUGE-1	0.5388	0.5532	0.5668	0.5763	0.5700	0.5811
ROUGE-2	0.2971	0.3029	0.3105	0.3112	0.3132	0.3249
ROUGE-SU4	0.2938	0.2868	0.3060	0.3149	0.3057	0.3210
Spanish						
ROUGE-1	0.5004	0.5140	0.5422	0.5500	0.5425	0.5442
ROUGE-2	0.2983	0.2960	0.3128	0.3195	0.3109	0.3194
ROUGE-SU4	0.2847	0.2801	0.2973	0.3052	0.2963	0.3047

proceed to rate sentences reveals that they often made their choice by comparing outputs for a given cluster.

Table 8 shows the manual analysis that ratifies the good results of our system. Informativeness scores are consistently improved by the ILP method, whereas grammaticality results measured on the three systems are similar. Besides, statistical tests show that this enhancement regarding informativeness and grammaticality is significant for Spanish corpus. For the Portuguese and Spanish corpora, our method obtained the best results for informativeness and grammaticality with shorter compressions. For the French corpus, F10 obtained the highest value for grammatical quality, while BM13 generated more informative compressions. Finally, the reranking method proposed by BM13 based on the analysis of *key phrases* of candidate compression improves informativeness, but not to the same degree as our ILP model. This more moderate enhancement can be related to the limitation of this reranking method to candidate sentences generated by F10.

5.2 Discussion

Short compressed sentences are appropriate to summarize documents; however, they may remove

key information and prejudice the informativeness of the compression. For instance, for the sentences that would be associated with a higher relevant score by the ATS system, producing longer sentences would be more appropriate. Generating longer sentences makes easier to keep informativeness but often increases difficulties to have a good grammatical quality while combining different parts of sentences. Depending on the kind of cluster short compressions can be generated or not with good informativeness scores. In that respect, the system has to adapt its analysis to generate long or short sentences.

F10 produced the shortest compressions for all corpora but its outputs have the worst informativeness score. BM13 improved these results; however, their compressions are longer than F10 (for all corpora) and our system (for the Portuguese and the Spanish corpora). For Spanish, the informativeness scores of all versions of our method are statistically better than F10, and the version ILP:∞+LM is statistically better than both baselines for this corpus. Given the small difference of informativeness between BM13 and our ILP approach for the Portuguese and the French corpora, we analyzed the relation between informativeness and CR to define which method

Table 7. Compression length (#words), standard deviation and number of used keywords computed on the French, Portuguese and Spanish corpora.

Metrics	Methods					
	F10	BM13	ILP:80%	ILP:80%+LM	ILP:∞	ILP:∞+LM
French						
Avg. Length	16.9 ± 5.1	19.7 ± 6.9	19.8 ± 4.8	19.5 ± 4.9	20.6 ± 5.5	20.8 ± 5.8
Comp. Ratio. (%)	51.3	59.8	59.9	59.2	62.6	63.1
Keywords	6.8	7.7	8.3	7.9	8.5	8.1
Portuguese						
Avg. Length	17.3 ± 5.3	22.9 ± 6.3	19.5 ± 4.0	19.4 ± 4.4	21.8 ± 5.5	20.5 ± 5.0
Comp. Ratio. (%)	52.2	69.1	59.0	58.7	65.9	62.2
Keywords	7.0	8.5	8.2	8.0	8.9	8.3
Spanish						
Avg. Length	16.5 ± 6.4	23.4 ± 8.4	20.3 ± 5.9	20.9 ± 5.2	21.1 ± 7.0	23.4 ± 7.3
Comp. Ratio. (%)	43.2	61.2	53.2	54.7	55.4	61.2
Keywords	5.8	6.9	7.7	7.6	7.9	7.9

obtains the best results. For Portuguese, BM13 and all versions of our system achieved similar informativeness scores, whereas our method generated significantly shorter compressions with an absolute decrease in the range 3.0–10.1 points. For the French corpus, it is complicated to define the best system because the second baseline, ILP:80% and ILP:∞ have similar informativeness scores for similar CRs. An inspection of the compressions generated by all systems highlighted that the low performance of our approach for the French dataset is partly related to the structure of negative sentences in French. In this language, these sentences must usually be composed of the tokens “ne” and “pas” to be correct, like in the following example: “La France n’a pas remporté le championnat du monde de volley-ball” (France did not win the world volleyball championship). In the studied dataset, the French corpus contains 27 negative source sentences divided into 13 clusters. Our approach often missed one of these tokens in its output compressions with the negative structure, which reduced the scores for informativeness and grammaticality. A post-processing of compressions could check if these two tokens are presented in the compression and correct this error.

Tables 7 and 8 show that the informativeness scores and keywords are related, i.e., the higher the number of keywords the higher the informativeness score. According to its type (with respect to the size and the amount of information), a cluster can have a different number of real keywords (more or less than 10 keywords). The number of keywords and informativeness scores are related, except for BM13 on the French corpus that used fewer keywords than our method and still generated more informative compressions.

The POS-LM post-processing does not improve significantly the compression quality of our method. This post-processing maintain or enhance grammaticality for all corpora, except for the ILP:∞+LM for Portuguese corpus, and informativeness for the Portuguese and the Spanish corpora. The biggest difference between these two versions of all methods is on the Spanish corpus (differences of 0.1 and 0.14 are observed for informativeness and grammaticality, respectively), for which the POS-LM version generated a longer version (CR is increased by 5.8 points), which justifies the improvement of informativeness.

Table 8. Manual evaluation of compression (ratings are expressed on a scale of 0 to 2). The best results are in bold (* and ** indicate significance at the 0.01 and the 0.001 level using ANOVA's test related to F10, respectively; † and †† indicate significance at the 0.01 and the 0.001 level using ANOVA's test related to BM13, respectively).

Metrics	Methods					
	F10	BM13	ILP:80%	ILP:80%+LM	ILP:∞	ILP:∞+LM
French						
Informativeness						
Score 0	20%	10%	14%	16%	14%	14%
Score 1	36%	31%	32%	35%	27%	34%
Score 2	44%	59%	54%	49%	59%	52%
Avg.	1.25 ± 0.8	1.48 ± 0.7	1.40 ± 0.7	1.33 ± 0.7	1.45 ± 0.7	1.39 ± 0.7
Grammaticality						
Score 0	6%	7%	12%	8%	10%	10%
Score 1	23%	29%	36%	29%	35%	36%
Score 2	71%	64%	52%	63%	55%	54%
Avg.	1.65 ± 0.6	1.56 ± 0.6	1.44 ± 0.7	1.55 ± 0.6	1.45 ± 0.7	1.44 ± 0.7
Portuguese						
Informativeness						
Score 0	9%	7%	8%	5%	7%	8%
Score 1	30%	16%	18%	22%	12%	13%
Score 2	61%	77%	74%	73%	81%	79%
Avg.	1.51 ± 0.7	1.70 ± 0.6	1.66 ± 0.6	1.68 ± 0.6	1.74 ± 0.6	1.71 ± 0.6
Grammaticality						
Score 0	9%	8%	6%	5%	4%	7%
Score 1	21%	18%	18%	21%	15%	17%
Score 2	70%	74%	76%	74%	81%	76%
Avg.	1.61 ± 0.6	1.66 ± 0.6	1.71 ± 0.6	1.69 ± 0.6	1.76 ± 0.5	1.68 ± 0.6
Spanish						
Informativeness						
Score 0	24%	26%	12%	11%	10%	10%
Score 1	49%	31%	39%	36%	39%	29%
Score 2	27%	43%	49%	53%	51%	61%
Avg.	1.02 ± 0.7	1.16 ± 0.8	1.36 ± 0.7 **	1.41 ± 0.7 **	1.40 ± 0.7 **	1.50 ± 0.7 ***†
Grammaticality						
Score 0	11%	18%	12%	8%	10%	6%
Score 1	26%	33%	35%	36%	35%	29%
Score 2	63%	49%	53%	56%	55%	65%
Avg.	1.51 ± 0.7	1.30 ± 0.8	1.40 ± 0.7	1.48 ± 0.6	1.45 ± 0.7	1.59 ± 0.6 †

5.3 Applications

Most of previous MSC approaches have been applied on the Text Summarization problem and its variations. Among these works, several versions of our ILP method on different types of documents and in multiple languages have been successfully tested.

In the first application, the ILP approach was applied to the problem of microblog contextualization [17, 19]. Given a microblog about a festival, Linhares Pontes et al.'s [17, 19] system was able to generate a summary (maximum of 120 words) in four languages (English, French, Portuguese

and Spanish) of Wikipedia's pages describing this microblog. In order to get more information about these festivals, they used Wikipedia to find information about these festivals and adapt the MSC method to extract relevant information related to the festival and generate a summary.

Linhares Pontes et al. [15] also investigated the generation of cross-lingual speech summaries of news documents. The goal was to analyze an audio file in French and generate a text summary in English. Contrary to the text document, the transcription of audio files must use Automatic Speech Recognition (ASR), which complicates and reduces the quality of the summary generation.

They adapted the MSC method to analyze sentences, both in their original and translated forms, and generate informative compressions in English using the relevance of French and English sentences. Their MSC method also analyzed 3 grams to add grammatically correct sequences of words into the compressions. This feature allowed their method to generate compressions with a good grammaticality, even when there are erroneous transcribed sentences.

Finally, Linhares Pontes et al. [18, 20, 21] also dealt with the issue of Cross-Language Text Summarization to generate English and French summaries from clusters of news documents in French, Portuguese and Spanish languages. Their MSC approach was applied on similar sentences among the documents to summarize. Despite the variety of these sentences (short, long, verbal and non-verbal sentences) and the introduction of errors by the used machine translation engine, experiments showed that the system usually generated correct compressions that are shorter and more informative than their source sentences.

6 Conclusion

Multi-Sentence Compression aims to generate a short informative text summary from several sentences with related and redundant information. Previous works built word graphs weighted by cohesion scores from the input sentences, then selected the best path to select words of the output sentence. We introduced in this study a model for MSC with two novel features. Firstly, we extended the work done by Boudin and Morin [5] that introduced keywords to post-process lists of N-best compressions. We proposed to represent keywords as labels directly on the vertices of word graphs to ensure the use of different keywords in the selected paths. Secondly, we devised an ILP modeling to take into account these new features with the cohesion scores, while selecting the best sentence. The compression ratio can be modulated with this modeling, by selecting for example a higher number of keywords for the sentences considered essential for a summary.

Our methodology was evaluated on three corpora built from Google news: a first one

in French which had been built and used in [5], a second and a third one in Portuguese and in Spanish [22]. Automatic measures with the ROUGE package were supplemented with a manual evaluation carried out by human judges in terms of informativeness and grammaticality. We showed that keywords are important features to produce valuable compressed sentences. The paths selected with these features generate results consistently improved in terms of informativeness while keeping up their grammaticality.

There are several potential avenues of work. We can use other kinds of language models based on Neural Networks [29] as an additional score to the optimization criterion to improve grammaticality. Another objective can be to manage polysemy through the use of the same label for the synonyms of each keyword inside the word graph. Finally, MSC can be jointly employed with the classical methods of Automatic Text Summarization by extraction in order to generate better summaries.

7 Appendix

Two examples in Spanish and Portuguese are provided in this section to illustrate the differences observed between the tested methods.

7.1 Spanish

The Spanish cluster (Table 9) is composed of 20 similar sentences. The vocabulary of this cluster is composed of 880 tokens and this cluster has a TTR of 33.3%. F10 generated the shortest compression; however, the sentence has missing information. The second baseline system and our method without post-processing generated incorrect compressions. Our method without post-processing generated a sentence with relevant keywords but it is not correct. The post-processing selected a more grammatical compression without reducing informativeness. The top 10 keywords selected by LDA were : *vuelo, cuba, fort, lauderdale, unidos, primer, jetblue, comercial, clara* and *florida*.

Table 9. Example in Spanish showing the first 3 sentences among 20 source sentences and 1 of 3 available references.

Source document	
<p>El vuelo 387 de la aerolínea estadounidense JetBlue inauguró una nueva era en el transporte entre ambos países, al partir desde Fort Lauderdale (Florida, sureste) cerca de las 10:00 locales (14H00 GMT), y llegar a Santa Clara, 280 Km al este de La Habana, a las 10:57. (<i>Flight 387 of the US airline JetBlue inaugurated a new era in transport between the two countries, departing from Fort Lauderdale (Florida, southeast) at around 10:00 local time (14H00 GMT), and arriving in Santa Clara, 280 km east of Havana, at 10:57.</i>)</p> <p>Un avión de pasajeros de la línea aérea JetBlue despegó este miércoles a Cuba desde el aeropuerto Internacional de Fort Lauderdale en lo que viene a ser el primer vuelo regular entre Estados Unidos y la isla caribeña desde 1961, en un nuevo hito en la nueva fase de relaciones entre Washington y La Habana. (<i>A JetBlue airliner took off for Cuba on Wednesday from Fort Lauderdale International Airport, thus becoming the first regular flight between the United States and the Caribbean island since 1961, as a new milestone in the new phase of relations between Washington and Havana.</i>)</p> <p>La aerolínea JetBlue inaugurará los vuelos directos comerciales el 31 de agosto con un viaje entre Fort Lauderdale, Florida, hasta el aeropuerto de Santa Clara, a unos 270 kilómetros al este de La Habana, reportó la compañía estadounidense. (<i>JetBlue will inaugurate direct commercial flights on Aug. 31 with a trip from Fort Lauderdale, Florida, to Santa Clara airport, some 270 kilometers east of Havana, the U.S. company reported.</i>)</p>	
Reference	
<p>La aerolínea JetBlue Airways Corp inauguró el 31 de agosto los vuelos directos entre Estados Unidos y Cuba tras 50 años de suspensión . (<i>The airline JetBlue Airways Corp opened on August 31 direct flights between the United States and Cuba after 50 years of suspension .</i>)</p>	
Compressions	
F10:	la aerolínea <u>jetblue</u> inauguró este miércoles a <u>cuba</u> el primer <u>vuelo</u> inaugural . (<i>the airline jetblue opened the inaugural first flight to cuba this wednesday .</i>)
BM13:	el aeropuerto de <u>fort lauderdale</u> , <u>florida</u> , sureste de estados <u>unidos</u> y <u>cuba</u> desde 1961 partió este miércoles el primer <u>vuelo</u> inaugural . (<i>the airport of fort lauderdale , florida , southeastern united states and cuba since 1961 departed this Wednesday on the inaugural first flight .</i>)
ILP:80%	el aeropuerto de <u>fort lauderdale</u> , <u>florida</u> , sureste de estados <u>unidos</u> y <u>cuba</u> desde 1961 partió este miércoles el primer <u>vuelo</u> inaugural . (<i>the airport of fort lauderdale , florida , southeastern united states and cuba since 1961 departed this Wednesday on the inaugural first flight .</i>)
ILP:80%+LM	la aerolínea <u>jetblue</u> inauguró este miércoles el primer <u>vuelo</u> desde <u>fort lauderdale</u> , <u>florida</u> , sureste de estados <u>unidos</u> a <u>cuba</u> desde 1961 . (<i>the airline jetblue opened Wednesday the first flight from fort lauderdale , florida , southeastern united states to cuba since 1961.</i>)
ILP:∞	el aeropuerto de <u>fort lauderdale</u> , <u>florida</u> , sureste de estados <u>unidos</u> y <u>cuba</u> desde 1961 partió este miércoles el primer <u>vuelo</u> inaugural . (<i>the airport of fort lauderdale , florida , southeastern united states and cuba since 1961 departed this Wednesday on the inaugural first flight .</i>)
ILP:∞+LM	la aerolínea <u>jetblue</u> inauguró este miércoles el primer <u>vuelo</u> desde <u>fort lauderdale</u> , <u>florida</u> , sureste de estados <u>unidos</u> a <u>cuba</u> desde 1961 . (<i>jetblue airlines inaugurated this wednesday the first flight from fort lauderdale, florida , southeastern united states to cuba since 1961 .</i>)

7.2 Portuguese

Table 10 displays a cluster composed of 11 Portuguese sentences with a TTR of 37% and a

vocabulary of 351 tokens. In this case, F10 did not generate the shortest compression and has

incorrect information. The second baseline, which post-processes the outputs of the first one, was not able to correct the errors. Almost all versions of our method generated the shortest and the most informative compressions related to the text. Our method without post-processing generated the best compression. The post-processing selected a more grammatically correct sentence, while its information is incorrect. The top 10 keywords selected by LDA were : *tesla, solarcity, milhões, 2,6, solar, empresa, carros, fabricante, dólares* and *motors*.

Acknowledgments

This work was partially financed by the European Project CHISTERA-AMIS ANR-15-CHR2-0001 and the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

References

1. Banerjee, S., Mitra, P., & Sugiyama, K. (2015). Multi-document abstractive summarization using ilp based multi-sentence compression. *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, AAAI Press, pp. 1208–1214.
2. Barzilay, R. & Lapata, M. (2006). Aggregation via set partitioning for natural language generation. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 359–366.
3. Barzilay, R. & McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Comput. Linguist.*, Vol. 31, No. 3, pp. 297–328.
4. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022.
5. Boudin, F. & Morin, E. (2013). Keyphrase extraction for n-best reranking in multi-sentence compression. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, HLT-NAACL, The Association for Computational Linguistics, pp. 298–305.
6. Bruckner, S., Hüffner, F., Komusiewicz, C., & Niedermeier, R. (2013). *Evaluation of ILP-Based Approaches for Partitioning into Colorful Components*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 176–187.
7. Clarke, J. & Lapata, M. (2007). Modelling compression with discourse constraints. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning (EMNLP-CoNLL-2007)*, Prague, Czech Republic, pp. 1–11.
8. Clarke, J. & Lapata, M. (2008). Global inference for sentence compression an integer linear programming approach. *J. Artif. Int. Res.*, Vol. 31, No. 1, pp. 399–429.
9. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for Information Science*, Vol. 41, No. 6, pp. 391–407.
10. Filippova, K. (2010). Multi-sentence compression: Finding shortest paths in word graphs. *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 322–330.
11. Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., & Vinyals, O. (2015). Sentence compression by deletion with lstms. Márquez, L., Callison-Burch, C., Su, J., Pighin, D., & Marton, Y., editors, *EMNLP*, The Association for Computational Linguistics, pp. 360–368.
12. Filippova, K. & Strube, M. (2008). Sentence fusion via dependency graph compression. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 177–185.
13. Huet, S., Gravier, G., & Sébillot, P. (2010). Morpho-syntactic post-processing of n-best lists for improved french automatic speech recognition. *Computer Speech and Language*, Vol. 24, No. 4, pp. 663–684.
14. Lin, C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*.
15. Linhares Pontes, E., González-Gallardo, C.-E., Torres-Moreno, J.-M., & Huet, S. (2019). Cross-lingual speech-to-text summarization. Choroś, K., Kopel, M., Kukla, E., & Siemiński, A., editors, *Mul-*

Table 10. Example in Portuguese showing the first 3 sentences among 11 source sentences and 1 of 2 available references.

Source document	
<p>A Tesla fez uma oferta de compra à empresa de serviços de energia solar SolarCity por mais de 2300 milhões de dólares (<i>Tesla made an offer to purchase the SolarCity solar energy services company for over 2,300 million dollars.</i>).</p> <p>A Tesla Motors, fabricante de carros elétricos, anunciou aquisição da SolarCity por US\$ 2,6 bilhões (<i>Tesla Motors, a manufacturer of electric cars, announced the purchase of SolarCity for \$2.6 billion.</i>).</p> <p>A fabricante de carros elétricos e baterias Tesla Motors disse nesta segunda-feira (1) que chegou a um acordo com a SolarCity para comprar a instaladora de painéis solares por US\$ 2,6 bilhões, em um grande passo do bilionário Elon Musk para oferecer aos consumidores um negócio totalmente especializado em energia limpa, informou a Reuters (<i>Electric car and battery manufacturer Tesla Motors said on Monday (1) that it reached an agreement with SolarCity to buy the solar panel installer for \$2.6 billion, in a big step took by billionaire Elon Musk to offer consumers a fully specialized clean energy business, Reuters reported.</i>).</p>	
Reference	
<p>A Tesla Motors anunciou acordo para comprar a SolarCity por US\$ 2,6 bilhões. (<i>Tesla Motors has announced an agreement to buy SolarCity for US\$ 2.6 billion.</i>)</p>	
Compressions	
F10	a solarcity para comprar a instaladora de painéis solares por us\$ 2,6 bilhões (<i>solarcity to buy the solar panel installer for us\$ 2.6 billions .</i>)
BM13	a solarcity para comprar a instaladora de painéis solares por us\$ 2,6 mil milhões de dólares (<i>solarcity to buy the solar panel installer for us\$ 2.6 billion dollars.</i>).
ILP:80%	a tesla vai comprar a solar solarcity por 2,6 mil milhões de dólares (<i>tesla will buy the solar solarcity for 2.6 billion dollars.</i>)
ILP:80%+LM	a solarcity para comprar a instaladora de painéis solares por 2,6 mil milhões de dólares (<i>solarcity to buy the solar panel installer for 2.6 billion dollars.</i>)
ILP:∞	a tesla vai comprar a solar solarcity por 2,6 mil milhões de dólares (<i>tesla will buy the solar solarcity for 2.6 billion dollars.</i>)
ILP:∞+LM	a solarcity para comprar a instaladora de painéis solares por 2,6 mil milhões de dólares (<i>solarcity to buy the solar panel installer for 2.6 billion dollars.</i>)

timedia and Network Information Systems, Springer International Publishing, Cham, pp. 385–395.

16. Linhares Pontes, E., Huet, S., Linhares, A. C., & Torres-Moreno, J.-M. (2018). Multi-sentence compression with word vertex-labeled graphs and integer linear programming. *Proceedings of the 12th Workshop on Graph-Based Natural Language Processing (TextGraphs)*, Association for Computational Linguistics.
17. Linhares Pontes, E., Huet, S., & Torres-Moreno, J.-M. (2018). Microblog contextualization: Advantages and limitations of a multi-sentence compression approach. Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J. Y., Soulier, L., SanJuan, E., Cappellato, L., & Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality,*

and Interaction, Springer International Publishing, Cham, pp. 181–190.

18. Linhares Pontes, E., Huet, S., & Torres-Moreno, J.-M. (2018). A multilingual study of compressive cross-language text summarization. *Proceedings of the 17th Mexican International Conference on Artificial Intelligence (MICAI)*, Springer, Guadalajara, Mexico.
19. Linhares Pontes, E., Huet, S., Torres-Moreno, J.-M., & Linhares, A. C. (2017). Microblog contextualization using continuous space vectors: Multi-sentence compression of cultural documents. *Working Notes of the CLEF Lab on Microblog Cultural Contextualization*, volume 1866, CEUR-WS.org.
20. Linhares Pontes, E., Huet, S., Torres-Moreno,

- J.-M., & Linhares, A. C. (2018).** Cross-language text summarization using sentence and multi-sentence compression. **Silberstein, M., Atigui, F., Kornysheva, E., Métails, E., & Meziane, F.**, editors, *Natural Language Processing and Information Systems*, Springer International Publishing, Cham, pp. 467–479.
21. **Linhares Pontes, E., Huet, S., Torres-Moreno, J.-M., & Linhares, A. C. (2020).** Compressive approaches for cross-language multi-document summarization. *Data & Knowledge Engineering*, Vol. 125, 101763.
 22. **Linhares Pontes, E., Torres-Moreno, J.-M., Huet, S., & Linhares, A. C. (2018).** A new annotated portuguese/spanish corpus for the multi-sentence compression task. *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*.
 23. **Luong, A., Tran, N., Ung, V., & Nghiem, M. (2015).** Word graph-based multi-sentence compression: Re-ranking candidates using frequent words. **Merialdo, B., Nguyen, M. L., Le, D., Duong, D. A., & Tojo, S.**, editors, *2015 Seventh International Conference on Knowledge and Systems Engineering, KSE 2015, Ho Chi Minh City, Vietnam, October 8-10, 2015*, IEEE, pp. 55–60.
 24. **McKeown, K., Rosenthal, S., Thadani, K., & Moore, C. (2010).** Time-efficient creation of an accurate sentence fusion corpus. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 317–320.
 25. **Miao, Y. & Blunsom, P. (2016).** Language as a latent variable: Discrete generative models for sentence compression. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 319–328.
 26. **Mihalcea, R. & Tarau, P. (2004).** TextRank: Bringing order into texts. **Lin, D. & Wu, D.**, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, Association for Computational Linguistics, Barcelona, Spain, pp. 404–411.
 27. **Napoles, C., Van Durme, B., & Callison-Burch, C. (2011).** Evaluating sentence compression: Pitfalls and suggested remedies. *Proceedings of the Workshop on Monolingual Text-To-Text Generation, MTTG '11*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 91–97.
 28. **Nayeem, M. T., Fuad, T. A., & Chali, Y. (2018).** Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 1191–1204.
 29. **Niu, J., Chen, H., Zhao, Q., Su, L., & Atiquz-zaman, M. (2017).** Multi-document abstractive summarization using chunk-graph and recurrent neural network. *Proceedings of the 2017 IEEE International Conference on Communications (ICC)*, pp. 1–6.
 30. **Öncan, T., Altinel, İ. K., & Laporte, G. (2009).** A comparative analysis of several asymmetric traveling salesman problem formulations. *Computers & Operations Research*, Vol. 36, No. 3, pp. 637–654.
 31. **Reape, M. & Mellish, C. (1999).** Just what is aggregation anyway? **Dizier, P. S.**, editor, *Proceedings of the 7th European Workshop on Natural Language Generation*, Toulouse, pp. 20–29.
 32. **Rush, A. M., Chopra, S., & Weston, J. (2015).** A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, pp. 379–389.
 33. **Schmid, H. (1995).** Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIGDAT Workshop*, pp. 47–50.
 34. **ShafieiBavani, E., Ebrahimi, M., Wong, R. K., & Chen, F. (2016).** An efficient approach for multi-sentence compression. **Durrant, R. J. & Kim, K.-E.**, editors, *Proceedings of The 8th Asian Conference on Machine Learning*, volume 63 of *Proceedings of Machine Learning Research*, PMLR, The University of Waikato, Hamilton, New Zealand, pp. 414–429.
 35. **Shang, G., Ding, W., Zhang, Z., Tixier, A., Meladianos, P., Vazirgiannis, M., & Lorré, J.-P. (2018).** Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, pp. 664–674.
 36. **Thadani, K. & McKeown, K. (2013).** Supervised sentence fusion with single-stage inference. *Pro-*

ceedings of the Sixth International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, pp. 1410–1418.

37. **Torres-Moreno, J.-M. (2014).** *Automatic Text Summarization*. John Wiley & Sons.
38. **Tzouridis, E., Nasir, J. A., & Brefeld, U. (2014).** Learning to summarise related sentences. *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, ACL, pp. 1636–1647.
39. **Zhao, Y., Shen, X., Bi, W., & Aizawa, A. (2019).**

Unsupervised rewriter for multi-sentence compression. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 2235–2240.

40. **Zheng, C., Swenson, K., Lyons, E., & Sankoff, D. (2011).** OMG! orthologs in multiple genomes — competing graph-theoretical formulations. **Przytycka, T. M. & Sagot, M.-F.**, editors, *WABI*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 364–375.

Appendix B: Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge

palgrave
communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS




ARTICLE



<https://doi.org/10.1057/s41599-020-0423-6>

OPEN

Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge

Saturnino Luz ¹✉ & Shane Sheehan¹

ABSTRACT An approach to establishing requirements and developing visualization tools for scholarly work is presented which involves, iteratively: reviewing published methodology, in situ observation of scholars at work, software prototyping, analysis of scholarly output produced with the support of text visualization software, and interviews with users. This approach is embodied by the software co-designed by researchers working on the Genealogies of Knowledge project. This paper describes our co-design methodology and the resulting software, presenting case studies demonstrating its use in test analyses, and discussing methodological implications in the context of the Genealogies of Knowledge corpus-based approach to the study of medical, scientific, and political concepts.

¹Usher Institute, Edinburgh Medical School, The University of Edinburgh, Edinburgh, UK. ✉email: s.luz@ed.ac.uk

ARTICLE

PALGRAVE COMMUNICATIONS | <https://doi.org/10.1057/s41599-020-0423-6>

Introduction

The analysis of corpora has always been of central importance in the humanities. More recently, the spread of computing technology and the consolidation of the field of digital humanities has transformed the way corpus analysis is done. The use of computational tools has a relatively long tradition in the disciplines of lexicography and corpus linguistics (Svartvik, 2011), and the roots of the widely used Keyword-In-Context (KWIC) technique can be traced back at least to the 1950s, starting with the work of Luhn on concordance indexing (Luhn, 1960). This has been an extremely productive relationship, influencing many other areas of investigation in the humanities (Frank et al., 2018), including corpus-based translation studies (Baker, 1993a; Bernardini and Kenny, 2020).

As explained in the introduction to this special issue, the focus of the Genealogies of Knowledge (GoK) project is on exploring the role of translation and other sites of mediation in shaping the historical evolution of scientific and political concepts. A noteworthy aspect of the project is that it explores these issues through the methodological lens of concordance and collocation analysis, an approach that is strongly influenced by the work of the British linguists J.R. Firth, John Sinclair and Michael Halliday (Léon, 2007; Sinclair, 1991), and shaped by the use of computational tools. Scholarly work in this field of study traditionally proceeds in a “bottom-up” manner. Selected texts are read and analysed by scholars, and synthesis often relies on the investigator’s memory and powers of abstraction, as well as their theoretical framework. The use of corpus-based methods can radically change this mode of work. Corpus analysis suggests a “top-down” approach where one usually starts by obtaining an overview of the data and exploring a much larger volume of text than would be practical to do by means of eye and hand alone. This leads to an iterative process in which the investigator switches between overview and detail towards analysis and generalization. Visualization tools can aid this process by providing effective overviews and drawing the researcher’s attention to patterns that might otherwise go unnoticed, as well as serving as vehicle for visual explanations (Tufte, 1990).

All modern corpus-based studies of text in the Firthian tradition involve, minimally, computational support for term indexing, search, retrieval and display. However, despite these commonalities, different fields and studies often need to adapt old methods and develop new ones to suit their particular analytical needs. While this work of adapting and developing becomes part of the study’s methodology and will affect the research outcomes, this process is rarely documented or discussed at a theoretical level. Taking a broader methodological view of tool development by regarding this activity as a part of the conceptual framework of the corpus-based studies tradition that informs GoK is particularly important, as the project adapts a well established linguistics methodology to the study of spatiotemporal evolution of medical, scientific and political concepts.

Our goal in this paper is to document and discuss the ongoing process of co-design and development of text visualization tools to support the corpus-based investigations conducted as part of the GoK project. In doing so, we hope to establish general methods for the development of such tools in interdisciplinary contexts. We envision this as a first step towards the more ambitious goal of creating the basis for a truly interdisciplinary methodology for scholarly work that breaks the barriers between “developers” and “users” of tools, and that welcomes equally the contributions of interactive systems designers, corpus researchers and humanities scholars.

We start by presenting the development methodology and the design rationale for the software tools developed for the GoK project, covering the steps of methodology review, requirements

elicitation, observation of scholarly work, and prototyping activities. We then describe the GoK tools proper, as they exist today, present case studies illustrating the tools in use, and discuss the results and implications of this methodological approach.

Related work

The history of digital humanities dates back to the 1940s when Roberto Busa began work on *Index Thomisticus*, the first tool for text search in a very large corpus. Visualization tools in digital humanities mostly focused on close reading techniques for investigating individual texts until the explosion of interest in distance reading techniques triggered by Moretti’s “Graphs, maps, trees” in 2005 (Moretti, 2005).

While there are many visualization techniques and systems developed for digital humanities (Jänicke et al., 2015) concordance-based visualization is quite rare, that is to say you do not often find visualizations which encode lexical and grammatical patterns of co-occurrence around a keyword in digital humanities literature. These co-occurrence patterns are used extensively in related fields such as translation studies and corpus linguistics upon which the foundations of the GoK project stand.

The practitioners of corpus linguistics range across many diverse disciplines in the study of language. For example, McEnery and Wilson (2001) introduce corpus linguistics by covering topics such as: lexical studies, grammar, semantics pragmatics and discourse analysis, sociolinguistics, stylistics and text linguistics, historical linguistics, dialectology and variation studies and psycholinguistics, teaching of languages and linguistics, cultural studies and social psychology. While the exact methodology differs in each case the use of computer generated quantitative information, the investigation of lexical or grammatical patterns in the corpus and the qualitative discussion of the quantitative and textual information are strong components of the techniques.

Corpus linguistics quantitative techniques employ frequency and statistical analysis to collect evidence of language structure or usage. Many of the methods can be thought of as empirical linguist techniques. However, a common misconception is that corpus-based approaches are entirely quantitative and do not require any qualitative input (Baker, 2006). Biber et al. (1998) present a collection of quantitative corpus-based methods, in each case “... a great deal of space is devoted to explanation, exemplification, and interpretation of the patterns found in quantitative analyses. The goal of corpus-based investigations is not simply to report quantitative findings but to explore the importance of these findings for learning about the patterns of language use”. These qualitative interpretations are important as “... a crucial part of the corpus-based approach is going beyond the quantitative patterns to propose functional [qualitative] interpretations explaining why the patterns exist”. In a linguistic study, before the application of quantitative techniques the formulation of hypothesis and research questions is often informed by qualitative analysis and/or prior knowledge of the texts under investigation. A good quantitative study must be preceded by a qualitative approach if anything beyond a simple description of the statistical properties of the corpus is to be achieved (Schmied, 1993).

Translation studies is a field where corpus-based methods have grown in popularity. Baker’s early advocacy for the use of corpus-based methods in the study of translation (Baker, 1993b) has led to its adoption in various sub-fields of translation (Baker, 1995; Olohan, 2002; Rabadán et al., 2009; Zanettin, 2001, 2013). The re-emergence of corpus-based methods had a transformative effect, and the corpus-based methodology has been described as one of

the most important gate-openers to progress in translation studies (Hareide and Hofland, 2012).

Concordance analysis is a core activity of scholars in a number of humanities disciplines, including corpus linguistics, classical studies, and translation studies, to name a few. Through the advent of technology and the ever increasing availability of textual data this type of structured analysis of text has grown in importance (Bonelli, 2010; Sinclair, 1991). Some of the most popular tools which have concordance browsing at their core include *WordSmith Tools* (Scott, 2008), *SketchEngine* (Kilgariff et al., 2004) and *AntConc* (Anthony, 2004). While there is some variation in advanced features across the range of concordance browsers each provides a windowed concordance which can be explored (via scrolling or multiple pages) and is usually sortable at word positions. This simple feature set is the key to supporting the traditional corpus linguistic methodology of concordance analysis.

As computational tools and methods for concordance and collocation analysis are central to the GoK research programme of extending the methodology of corpus-based translation studies, we focus on the tasks to be supported in this domain. We will return to a review of tools for corpus analysis and visualization in section “Analysis of existing visualizations”, once we have defined and conceptualized the tasks involved in greater detail.

Iterative co-design for corpus-based scholarly work

The process of development for the GoK visualization tools involved various steps, including an analysis of the published methods in corpus linguistics on which the GoK project draws, low-fidelity level prototyping and requirements elicitation, high-fidelity prototyping and formative evaluation of these prototypes in use. These steps were not usually performed sequentially, but rather iteratively: progress made by means of one activity often informed our approach at other stages in the process, whilst simultaneously reflecting knowledge and techniques gained during other iteration cycles. In what follows, however, these stages must be shown sequentially, as cohesive blocks, for presentation purposes. Cross connections are indicated as needed, and the reader is warned that some of these are forward references.

Analysis of published methodology. The work of Sinclair in corpus linguistics (Sinclair, 1991, 2003) and Baker and others in translation studies (Baker, 1993a) are the main theoretical influences guiding the GoK methodology. Published methodology in this area is therefore a natural starting point for the identification of aspects of analytical work that can be supported by computational tools. In the case of corpus-based analysis, we were fortunate to be able to rely on the work of John Sinclair, who not only developed the foundations of a method for linguistic analysis which has subsequently influenced a number of research programmes, including GoK, but who also described this method in a detailed tutorial form (Sinclair, 1991, 2003).

In his book *Reading Concordances*, Sinclair (2003) presents 18 tasks. Each task guides the reader through an analysis, describing both the mechanics and analysis required to complete the task. Although a computational element is implied, the tasks described in the book are based on given sets of printed and pre-formatted concordances. For each of these tasks we performed a hierarchical task analysis (Annett, 2003; Newman and Lamming, 1995) by combining or splitting the steps into a series of actions and sub-actions. The goal of such an analysis is to identify the low level actions (for example, sorting a list of words) required to complete higher level tasks (e.g. finding a significant collocate), and to order them in terms of importance or frequency. Using the completed task analysis to detect those actions which are not well

supported by current tools or techniques can lead to rational design choices.

Each task was tagged to assist with classification and counting of the actions and sub-actions. We add tags to each task in an iterative process, the first stage of which involved the tagging of each analysis step with potential action tags. On completion of this initial pass, the tags are reviewed for relevance and redundancy, and are collapsed into more generalized actions where relevant. A similar review cycle is performed at the level of the eighteen tasks presented by Sinclair to homogenize the actions across tasks.

As an example let us look at the tags which were applied to the first instruction in the second of Sinclair’s tasks. This task focuses on regularity and variation in phrase usage. The first step in the task description asks the reader to use a supplied concordance to “Make a list of the repeated words that occur immediately to the left of gamut. Sort them in frequency order. Then make a similar list of the words immediately to the right of gamut. Ignore single occurrences at present”. As this step involves listing all words in frequency order at a position the tags *frequency*, *word position* are applied. We choose not to apply the *estimate frequency* tag as that would apply if the reader were asked to find the most frequent word at a position where the actual counts would not matter. We are also not looking for *frequent patterns* as we only look at a single position for a single keyword.

This tagging procedure can allow a visualization researcher with limited knowledge of the problem domain to extract meaningful actions. However, this is a subjective process and additional efforts from other perspectives could yield interesting differences or similarities in the action hierarchy and counts.

While most of the tags we used to analyse the 18 tasks represent actions, a few additional tags were chosen to help clarify and add information about the actions and sub-actions. The purely clarifying tags are omitted from the analysis of tag frequency. Examples of these are the tags *expert knowledge*, *combinations* and others are not themselves actions, but are useful in clarifying the objective or operation of the sub-actions. These clarifying tags always appear with other primary tags. The part of speech (POS) tag is both a primary action tag and a clarifying tag. The POS primary action is to determine the POS of a word occurrence. The POS clarifying tag represents the use of POS information in another action.

We recorded the distribution of the tags according to the number of tasks in which each appeared and the total number of actions which received the tag as shown in Table 1. At a high level, this table tells us that reading concordance lines (CLs) and

Table 1 Action counts from task analysis.

Tag	No. of tasks in which an action appears	Total action appearances
estimate frequency	16	34
read context	16	31
frequent patterns	15	21
frequency	14	18
word position	13	24
POS: Part of speech	11	23
filter	11	18
sense	10	19
group	7	9
significant collocate	5	7
usage	5	6
phrase	5	6

Number of tasks which feature the action, out of 18 tasks, and total numbers of actions found in these 18 tasks.

ARTICLE

PALGRAVE COMMUNICATIONS | <https://doi.org/10.1057/s41599-020-0423-6>

actions which require positional statistics are necessary for the style of concordance analysis outlined by Sinclair (2003). Looking in more depth we see that reading the context of a concordance and having an expert classify some linguistic property is a very common action. Estimating word frequency and calculating exact frequencies are also very common tasks which often combine with the position tag.

The actions and sub-action tags generalize the descriptive analysis steps into operations which are common to many of the tasks. In this way, we try to generalize the actions required to analyse a concordance. The results of this process are discussed below.

At the top level of the hierarchy Fig. 1, the highest level actions are defined. These actions are distinct pieces of analysis which form part of the analysis tasks presented by Sinclair (2003). These high level actions are use cases of concordance analysis, such as attempts to investigate meaning, identify phrases or explore word usage properties. We identified these actions during the tagging

process by investigating the clusters and order of the tags attached to each task.

Under each of the top level actions we identify sub-actions which contribute to the completion of the top level goals. For example, frequent patterns must be identified before they can be classified as phrasal or non-phrasal usage (Sinclair, 1991). These sub-actions are given in the order they were observed for the top level action. The sub-actions may need to be completed in the order they were observed to successfully complete the top level task; this is often the case with the first sub-action, such as those listed for *Identify Frequent Patterns* and *Identify Phrases*. The second level tasks for the top level *Investigate Usage* action, on the other hand, do not need to be performed in order as they represent the sub-actions required to analyse separate forms of usage. It should also be noted that there is a lot of repetition across the second level sub-actions for each top level task. *Identifying frequent patterns* is an example of a sub-action which appears under every high level action. The type of pattern under

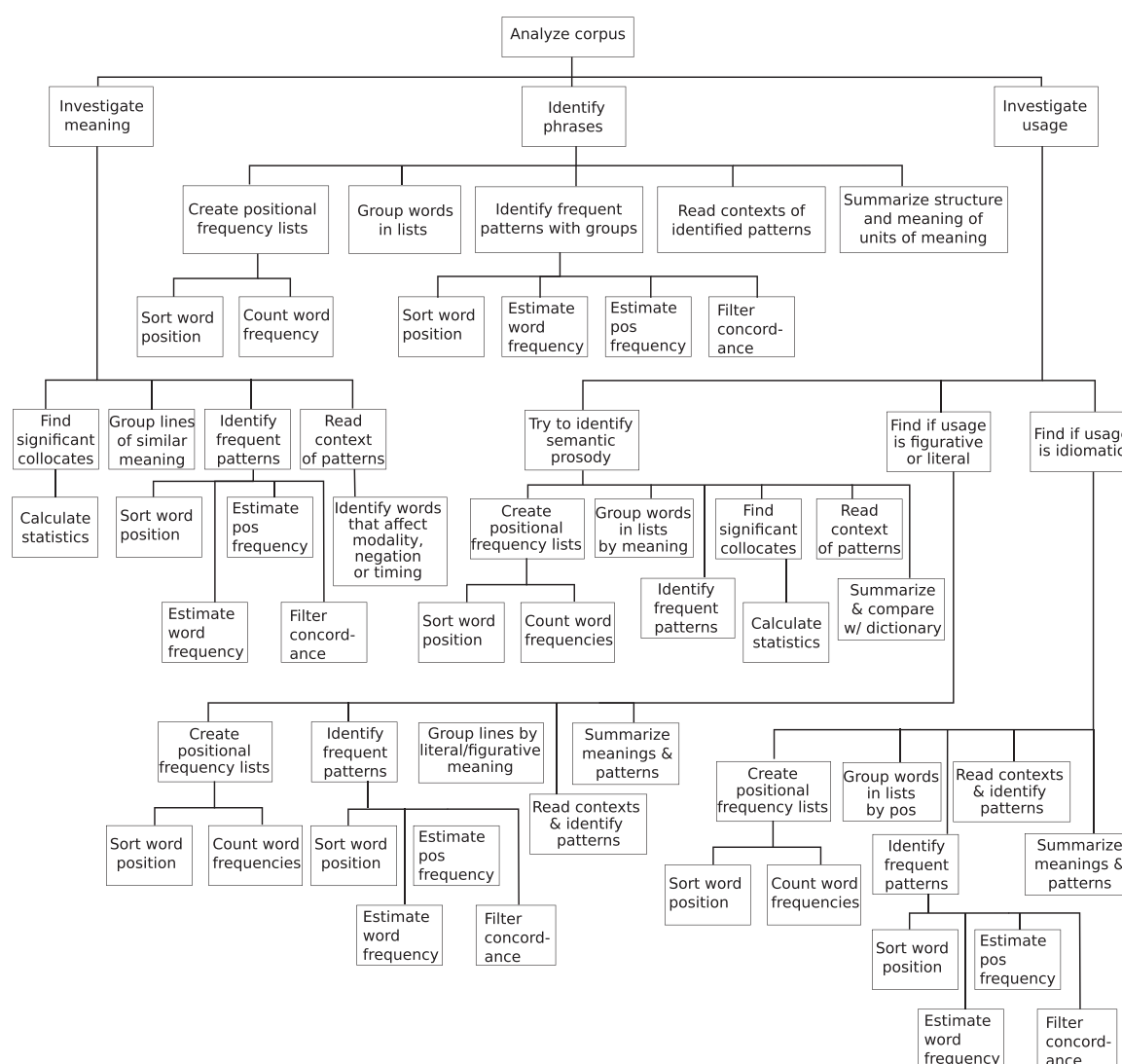


Fig. 1 Hierarchical task diagram. Tasks and action hierarchy for concordance-based corpus analysis.

investigation changes with the task but the mechanics remain relatively unchanged.

As we progress further into the hierarchy we find increasingly lower level sub-actions. These sub-actions describe the mechanics of the analysis. *Sort Word Position*, *Count Word Frequencies* and *Filter Concordance* are very specific low level sub-actions that are easy to identify and perform. The *Read Contexts* mechanics are simple but the purpose of the reading is usually to gain understanding or insight. The mechanics are not enough, an analyst or expert is required for the interpretation of what is read.

All of these actions and sub-actions should be considered when developing tools for research in this domain. If possible, they should all be supported and made more efficient and user friendly or even automated to reduce the workload for the corpus analyst. However, the tag and task analysis presented here would suggest most pressing that, in order to support the method outlined by Sinclair (2003) tools should allow close examination of individual CLs, while also providing support for analyzing positional frequencies and collocation patterns.

Conceptual data model of KWIC. To formalize the data structures, attributes and relationships inherent to the concordance list as revealed by the analysis above, a conceptual model of the KWIC concordance list has been created so that visualizations can be evaluated and designed for in terms of their effectiveness at representing the model. The design of the model seeks to structure the data entities in a manner which best supports the actions described in the task and action hierarchy Fig. 1. Our data model is simple and a natural extension of the task analysis, it is an abstraction of the concordance list which identifies the data attributes required for the identified tasks and actions. Creation of data abstractions by qualitatively analyzing the output of domain characterization effort is typical of good visualization design methods such as the Munzner's Nested Model (Munzner, 2009).

The traditional rendering of a KWIC concordance list evokes a conceptual model consisting of a list of aligned sentence fragments (CLs). In this model each CL has an attribute representing its position in the list (concordance lists are usually presented in alphabetical order) and contains an ordered set of word objects (WO) which make up the string that represents the CL. The WO represent an individual occurrence of a word in a CL and contain its string representation (nominal data), its position relative to the keyword and any other nominal variables (meta-data) available e.g. POS tags.

Many of the actions identified in the task analysis require reading of the CLs ("read context"). In order to read the context the text fragments must be available. Since the linear structure (sentence structure) of the CLs is emphasised in the CL model it is included in the KWIC conceptual model to facilitate this read context action.

Since the WO in the CL model are representative of a single occurrence of a word they do not have as an attribute a quantitative variable such as word frequency. The frequency values are available by counting similar WO within all CLs but the frequencies are not attributes of any entities in the model. We would like our KWIC model to contain these quantitative variables as attributes of some entity since "Estimate Word Frequency", "Count Word Frequencies" and "Estimate POS frequency" are needed in each of the three identified core tasks. We also found that word position was often required in conjunction with these frequency orientated actions. For instance, estimating word frequency at a position relative to the keyword is a common action required for analysis showing up in 13 of Sinclair's 18 tasks.

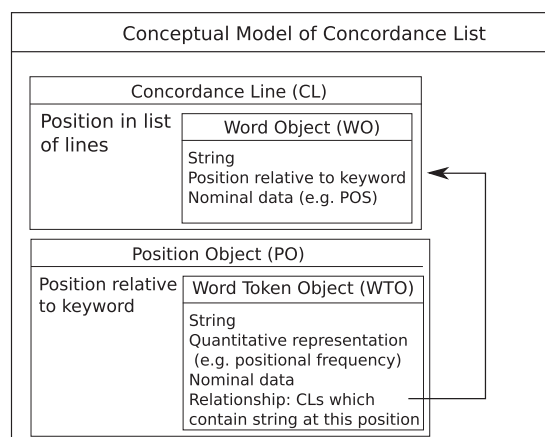


Fig. 2 Diagram representing the conceptual model of concordance lists. Conceptual data model of concordance lists.

With this in mind we now conceptualize the concordance lists as an ordered set of position objects (PO). Within each position object there is an attribute for the position relative to the keyword and a set of word token objects (WTO). These WTO differ from the WO in the CL model in several ways. The most important way they differ is that these objects represent all occurrences of a string (or string and nominal variable) at the position in which they reside. That is to say, there will be at most one object with a particular string and nominal variable (meta-data) combination. For example, if POS tags are available there will be one object representing the noun "date" and one representing the verb "date". Each WTO inherits its position as an attribute. Quantitative attributes which represent positional count, frequency or other statistics of the WTO in the KWIC are included after calculation. Finally, an attribute (relationship) which maps each WTO to the CLs in which it occurs is also available. This attribute and the position attribute of the CL WO provide a link between the models unifying the KWIC conceptual model as seen in Fig. 2. This linking of the conceptual models is especially useful for the frequent patterns action, where word combination frequencies between/across positions are required.

Analysis of existing visualizations. Text visualization encompasses many different visual methods. We are mainly interested in comparing visualizations which have been designed for keyword search results represented as a concordance list. However, we also wish to investigate techniques which, while not designed with the concordance in mind, could potentially have applications in concordance analysis.

Both the task analysis and conceptual model described above have identified a split in KWIC tool requirements, with each task requiring some combination of qualitative and quantitative actions. Qualitative actions most often operate on the concept of CLs where readability and the linear structure of the lines are emphasised. While the quantitative actions require positional statistics, they do not often require readability of the individual CLs.

Looking at the conceptual model, task analysis and action counts, we conclude that for qualitative primary actions the most important attribute to present is CL word order, so that the fragments are readable. By choosing an appropriate encoding of CL word order visualizations can aid qualitative concordance analysis.

ARTICLE

PALGRAVE COMMUNICATIONS | <https://doi.org/10.1057/s41599-020-0423-6>

The quantitative actions often require estimation of frequencies/statistics, or finding exact word frequencies/statistics. These quantitative variables are then explored using positional analysis/estimation of these statistics to identify linguistic patterns. Clearly then, one should prioritize the attributes PO Position (in which position relative to the keyword the word occurs), the qualitative attribute of WTO (how often the word occurs at each position) and to a lesser extent the PO WTO position attribute (which represents the quantitative ordering of the words at a position; i.e. the order of a frequency list).

One approach to evaluating the effectiveness of visual encoding is a qualitative discussion of images or video of a visualization system. In what follows, we present a qualitative discussion of visualization systems related to concordance analysis in a semi-structured manner. Paraphrasing Munzner (2009), while visualization experts may draw the same conclusions from the inspection of a system, the validation is strongest when there is an explicit discussion pointing out the desirable properties in the results.

A ranking of visual variables (Bertin, 1983) per data type was proposed by Mackinlay (1986). This ranking of variables is in agreement with a ranking proposed for quantitative data by Cleveland and McGill (1985). Mackinlay's ranking of visual variables for the three data categories (ordinal, nominal and quantitative) are useful to help guide variable choice. If a variable is chosen for a data attribute instead of a higher ranked variable, justification should be given. Often visual encodings are presented without such justification leaving the reader to guess at the authors reasoning (Green, 1998).

We go about evaluating related text visualization methods for concordance analysis by mapping each attribute of our conceptual model (section "Establishing initial requirements") to the visual variables used. We make note of the number of visual variables mapped to the attribute and the ranking of the visual variables for the data type it is representing. The attributes of the KWIC conceptual model are further expanded by categorizing them as nominal, ordinal or quantitative data types. This helps us to apply Mackinlay's rankings which rely on these three data categories.

To begin with, we will evaluate the most widely used concordance visualization, the traditional KWIC visualization (or concordance list). CL position is represented by the vertical position of its enclosed strings. These enclosed strings are rendered horizontally, left to right, in the order they appear in the text fragments. Both CL and WO are mapped to the best visual variables (*position*) for their ordinal position data types. Because of this, we expect it will be easy to identify individual CLs and find where they rank in the chosen ordering scheme (usually alphabetically by a selected or default word position). Similarly, identifying the WO in the order they appear in the text fragments will not be difficult. The concept of a position object can be loosely applied in this visualization; the word positions across CLs can be identified using horizontal position, even though this is made difficult due to the variable length of words. A visual variable that is associative, such as *color hue*, can be encoded to identify words at the same position.

Since the KWIC display is designed with the readability of CLs in mind, the inability to gain an overview of a large concordance is a necessary trade-off. In this rendering the detail is presented at all times, an overview of the entire concordance list is only available for concordances which fit within the screen at a readable font size. Windowing the concordance and scrolling is the usual solution, this works well for viewing individual CLs, but to get an overview of the positional frequencies and patterns a higher level view would be better. Larger screen sizes and higher resolutions can improve the situation but as more data becomes available the scale required becomes impractical.

As this visualization contains no explicit representation of the WTO as such, we expect visual assessment of exact or estimated word frequency to be difficult. Nevertheless, this visualization is the most commonly used tool in concordance analysis where, as we have shown, positional frequencies are regularly required. From our observational study (section "Observational research") and task analysis we found that counting the strings is the usual way to calculate these positional word frequencies. The standard tools used by scholars in the field do not offer positional frequency lists or other tools to make position word frequencies easier to work with. While this visualization is very effective for reading CLs it would seem to be of limited use for quantitative concordance actions.

Several tree-based visualizations (Culy and Lyding, 2010; Luz and Sheehan, 2014; Wattenberg and Viégas, 2008) have been proposed to attempt to bridge the gap between qualitative and quantitative analysis. *Word Tree* (Wattenberg and Viégas, 2008) displays the keyword and either the left or right context, taking the familiar form of a tree structure, in which the keyword is displayed as the root and additional word vertices are connected in text order to each other. The main benefits of this visualization are that the linear structure and readability of the CLs is maintained through the combination of the visual variables connection and horizontal position. Connection defines the word position by the number of edges from the root, and horizontal position per branch provides partial positional groups (ordered positions in a sub-tree). These positional groups allow the frequencies at a position along a branch to be estimated since the words are rendered proportionally in size of their sub-tree. While frequency in a sub-tree is easy to estimate, the frequency at a word position is less clear. Looking at word positions as they move away from the root (keyword) positional frequencies become increasingly difficult to estimate. This is because each branching point can contain words which occur in other branches, leading to the possibility of multiple occurrences of a word object at a position. So at a depth of one from the root each rendered word represents a positional WTO, but deeper into the tree each rendered word is a partial WTO which only represents each occurrence of a token at that position in the sub-tree. Although sorting by frequency is supported at the first position from the keyword, combinatorial explosion causes the estimation of frequency to become more difficult as we view positions deeper into the tree.

An additional problem with estimation of frequency (or other word statistics) using this visualization is that variation in word length causes the variable representing the quantitative information to be inconsistent. Font size, which can be equated with visual variable *area* is used to represent frequency in a branch. The square root scale used by *Word Tree* and other visualizations should make word area roughly proportional to frequency if not for word length variations. While it seems natural to include quantitative information about a word by scaling the font representing that word, it is worth noting that the visual variable *area* ranks fifth for the display of quantitative information under Mackinlay's ranking scheme and, additionally, variations of word length complicate the interpretation of the quantitative values. The other tree based visualizations when viewed through the lens of our conceptual model suffer from similar problems of positional branching and frequency representation. They do however solve a readability problem by displaying both left and right contexts simultaneously and connecting them so that a full CL can be read. *Double tree* is designed with different tasks and users in mind, ("linguist's task of exploratory search using linguistic information"), so it is not surprising that it does not map well to our model. In this representation, word position is strongly encoded by using an integral (visual variables which are

perceived together) combination of connection and horizontal position. This contrasts with the loose horizontal positioning in word three where only connection can be used to reliably derive word positions. Double Trees encode frequency using *color saturation*, the eighth ranking variable for quantitative data, combined with the previously discussed branching issues quantitative information is not strongly encoded in this design.

Other visualizations do not place great emphasis on the readability of the CLs. One such visualization is *interHist* (Lyding et al., 2014), a complementary visualization which is used to display quantitative information about its accompanying KWIC view. In this case, the visualization is rendered as stacked bars (rectangles) where height is used to display quantitative information (*length* is the second best variable for quantitative information). While this interface is designed for POS information and does not represent individual tokens or WTOs, it is not difficult to imagine a variant with these rectangles representing WTOs with no change to the visual representation. In *interHist* vertical positioning of the rectangles is not encoded with meaning, making it more difficult to perceive quantitative differences between the rectangles. However, for word frequency all bars will be the same height as each CL will contain the same number of tokens, and since there will be many more words than POS tags rendering the words using a color and a legend could become impractical.

Corpus Clouds (Culy and Lyding, 2011) is a frequency focused corpus exploration tool which consists of composite views of a corpus query. The main display is a word cloud, based on the tag cloud visualization (Viégas and Wattenberg, 2008), where the absolute frequencies of all words returned by a corpus query are displayed. These word clouds map this quantity to *area* using font-size, the limitations of which were previously discussed. This visual encoding does not translate well to our conceptual model since positional concordance frequencies are the quantity of interest, not global frequency lists. Another view in the interface presents a modified KWIC display. The modification is the addition of a small vertical bar, similar to a sparkline, beside each word token in the KWIC view. This makes use of the variable *length* for frequency information, but the effectiveness of the variable is reduced for several reasons. The main limitation is that the size of the bars is restricted, causing only large differences in frequency to be perceived easily. Additionally, comparisons between lines take place in both planes, vertically across CLs and horizontally within lines, again making it difficult to perceive small variations in frequency. The number of KWIC lines which can be displayed per screen is also practically limited if readability is to be maintained.

Structured Parallel Coordinates (Culy et al., 2011) is an application of the parallel coordinates visualization technique to different types of structured language data, one of which is a KWIC plus frequency visualization. This visualization places WTOs, rendered as text labels, on the parallel axis which represent ordered word positions. The CL structure is maintained using connecting lines between WTOs. Statistical information is then placed on additional axes and the connection between the position axes and the quantitative axes are used to express the desired quantitative attributes of the WTOs. An individual quantitative axis is required for each quantity and word position pair. As with all parallel coordinate visualizations the choice of axis orderings is important. In this case the choice was to order the word positions in concordance list order and create the statistical axes to the right of the collection of position axes. This positioning makes it difficult to follow connections from a word position to its related quantitative axis if there are many other axes in between. Additionally, comparing two word positions is perceptually difficult, as the user needs information from four

axes for a comparison across two word positions for a single statistic, such as frequency. In Structured Parallel Coordinates the linear order of the sentences is partially maintained through connection. However, since the connected nodes are WTOs the actual sentences are lost and only the preceding and next connections are meaningful.

TagSpheres (Jänicke and Scheuermann, 2017) have a rendering which encodes keyword based co-occurrences as position aware word cloud. Word position is encoded using an integral combination of color and radial position from the central keyword. This creates a strong positional encoding relative to the keyword. The linear structure of the concordance is abandoned in this rendering. The layout attempts to render in close proximity the same token at different word positions, this can help with identifying patterns of frequent co-occurrence for a single word. However, multi-word co-occurrence patterns do not have a clear mechanism for their identification. The layout uses font size, *area*, to represent quantitative information, comparison of quantitative information is made somewhat difficult due to comparison of area between words which may not be positioned on the same horizontal or vertical axis, as well as the issues of encoding font size with non-uniform word lengths. Another visualization which aims at providing an overview of concordances is Fingerprint Matrices (Oelke et al., 2013) where words of interest can be represented as rows and columns of an adjacency matrix, with glyphs representing co-occurrences across a document. While this allows high density of POs to be encoded, the matrix representation does not map well to the overall concordance list model.

It is also worth considering visualizations which were not explicitly designed with the concordance list in mind. Text visualizations which focus on summarizing or exploration texts can also be applied to a concordance list by viewing the concordance list as a document, instead of as a selection of fragments from source documents. We now discuss some of these visualizations and suggest possible modifications to better fit them to the KWIC model and action requirements. This discussion is not exhaustive, as its goal is to further illustrate visual tools for the text analyst and to suggest possible starting points for new research into concordance visualization, rather than general text exploration or summarization.

TextArc (Paley, 2002) is a radial layout of the sentences in a document, within which WTOs are placed at a location which represents the average position of that word in the document. Font size is often used to represent a quantitative value such as word frequency. In terms of visual variables area in the form of font size is used for quantitative information. Word positions are not identified but perhaps color could be used and reading of the lines is still possible as they are rendered as part of the visualization. While this visualization does not seem to be a good fit for the tasks we have identified, it could give visual insight into which documents a collocated word is mostly contained in. Alternatively by ordering the CLs from multiple documents according to the position in the documents spatial patterns of co-occurrence across documents could be observed.

It is easy to imagine extending *Phrase Nets* (van Ham et al., 2009) to be more positionally aware and using them to look at collocations within a concordance list instead of a document. For example instead of the typical use of PhraseNet to examine a pattern such as “X and Y” a PhraseNet could be built for the pattern “happy one position to left of keyword, Y in window 5 words”, this visualization would give information about the collocations of the word happy within a five word window of the concordance list. Limitations of this visualization from a concordance analysis perspective include quantitative information representation by font size and, as given, lack of positional information.

ARTICLE

PALGRAVE COMMUNICATIONS | <https://doi.org/10.1057/s41599-020-0423-6>

Another word cloud based visualization *TagPies* (Jänicke et al., 2018) visualizes keywords and their co-occurrences in a radial layout. While not explicitly mentioned, this interface could be used to visualize the comparative co-occurrences of the same keyword per position by having a separate slice of the pie for each word position (PO). However this visualization would encode the WTO positions as slice shaped word clouds, with word position relative to the centre of a slice and the pie centre having some quantitative interpretation. This visualization, while interesting, does not map well onto our conceptual model.

Sketch Engine, the most popular commercial concordance software, also contains visualization capabilities. These visualizations generate a radial layout of similar words in a corpus. Currently no positional concordance-based visualization is supported (Kocincová et al., 2015).

Similarly, a popular text analysis framework in digital humanities, *Voyant tools* (Miller, 2018; Voyant, 2020) contains a number of visualization components. These components, again, do not address the issue of positional word frequencies relative to a keyword, mostly focusing on keyword frequency and distribution. The built in concordance view is not split by word position only displaying textual columns for the left context, keyword and right context.

Regarding the ranking of visual variables, there are visual variables that are more effective at representing quantity than area (which seems to be used quite often), namely: length, position, angle and slope. It must be noted, however, that domain or task specific requirements ultimately drive one's choice of variables; in this case it just happens that quantitative information segmented by word position is important for many of the fundamental tasks of concordance analysis, and should have high priority when creating an encoding.

Establishing initial requirements. When designing visualization tools it is important to involve expert users in the requirements gathering process. However, simply asking domain experts what they need is rarely productive. Instead, two researchers from the GoK project were asked to list 20 questions which they would like to be able to answer about a corpus. They were asked to list them in order of importance where possible and to make themselves available afterwards to discuss the lists. The request for the lists was made a week before a meeting to discuss the results. It is also important to note that as tool development took place alongside corpus gathering and the preliminary steps of analysis by the GoK researchers, the answers were influenced by the discussions that took place during this phase of the project, as well as by the use of early prototypes and other, general purpose, tools such as concordancers typically used in translation studies. The discussion was used to get more details about the challenges facing users when trying to answer the questions they have identified. This technique is an established method for requirements elicitation in human computer interaction (Marai, 2018) and is becoming more widespread for domain characterization in visualization design.

The list created by the first researcher, Henry Jones (HJ), used a very loose ranking system. The categories and questions follow the order in which they occurred to the researcher and as such may be implicitly correlated with question importance. The three categories identified were *keywords*, *collocational patterns* and *temporal spread*.

The first seven questions identified were all associated with the analysis of a keyword, as shown below:

Keywords

1. How many times is the chosen keyword used across all of the corpus texts as a whole?

2. Is the keyword used with more or less uniform frequency in each of the corpus texts individually or are there significant imbalances in the dispersion of the keyword?
3. Which specific corpus texts use a given keyword with proportionally greater frequency? And lesser? What patterns can we see if we rank the corpus texts by number of hits for this keyword?
4. Which linguistic-grammatical form(s) of the term/concept under investigation (e.g. singular vs. plural, forms suffixed with -ship, -like or -ly) is/are more common across the corpus as a whole?
5. To what extent are the relative proportions of these different word-forms the same or different within each of the corpus texts individually?
6. Are there other related keywords we might study in order to expand our investigation? Can the software suggest keywords that are important to these texts but which we might not otherwise have thought of?
7. If so, are the frequencies of these terms similar or different to the first keyword, both across all of the corpus texts as a whole and in each corpus text individually?

Questions 1 and 4 can easily be answered with a concordance query given that the concordancing tools available can already list all instances of a lexical item as it appears across a corpus of texts. Questions 2, 3 and 5, on the other hand, are not so easy to answer using the tools more commonly used for concordancing and collocation analysis, such as WordSmith tools (Scott, 2001), the Sketch Engine (Kilgariff et al., 2014) and TEC (Luz, 2011). During discussions with the researcher, the use of spreadsheets to collect word frequencies from repeated concordance queries was offered as a potential solution. Manipulating the subcorpus selection interface to search each text individually or using the filenames displayed in the left hand column of a concordance were both seen as potentially useful techniques. These questions and the technical difficulties faced by the researcher in answering them were evidence of the need for greater metadata integration into concordance tools so that frequency per filename can be quickly estimated. In the remaining *Keyword* questions, 6 and 7, automated keyword suggestion was discussed as a potentially useful technology which could be added to the GoK suite of tools.

In the questions relating to *Collocational Patterns* (below) we found calculating or estimating frequency via the concordance to be part of the solution in six of the seven questions.

Collocational Patterns

1. What are the adjectives that most commonly modify the chosen keyword (LEFT +1) across all of the corpus texts as a whole?
2. What are the adjectives that most commonly modify the chosen keyword (LEFT +1) in each text individually?
3. Are there any adjectives that modify the chosen keyword significantly more frequently in one text when compared with the others?
4. Are these adjectives only used to describe this keyword or are they connected to other keywords in this text?
5. What verbs are most commonly associated with the keyword (normally, RIGHT +1, RIGHT +2) across all of the corpus texts as a whole?
6. What verbs are most commonly associated with the keyword (RIGHT +1, RIGHT +2) in each of the corpus texts individually?
7. Are there patterns of interest in any of the other word-positions relative to the keyword? For example, if the keyword is a label used to describe a particular kind of political agent, we might be interested to look at what

collective nouns are used to group and characterize these political agents (e.g. a mob of citizens, a tribe of politicians: LEFT +2).

The possibility of sorting a concordance according to the items occurring at different word positions, and filtering the concordance via the subcorpus selection interface were suggested as useful techniques for answering the questions. Again, the use of a spreadsheet for collecting the results of various searches was mentioned. Question 4, the remaining question, seemed to have ties to measures of collocation strength with a keyword. By calculating the collocation strength between the keyword and the adjective the user might get some measure related to how strongly they co-occur in comparison with other combinations. The user pointed out that these questions assume the keyword is a noun, but that they believed the tools required for analysis would not change substantially for words of other grammatical categories, due to the analyst's primary interest in general co-occurrence patterns. Consequently, it seems that any techniques which might make the collection of collocation frequency information more efficient would be beneficial.

It is also worth noting that this analysis was conducted on the English-language subcorpus of GoK. Most of the questions and routines discussed so far apply equally to the other languages of GoK (Arabic, Greek, and Latin), implying support for different languages and scripts as a requirement. The questions above, on the other hand, assume English grammar and syntax more specifically.

The final heading for our first set of questions was *Temporal Spread*:

Temporal Spread

1. In what ways do keyword and collocational patterns correspond to the temporal spread of these texts (i.e. given the fact that some of these texts were published in 1850, others in 2012)?
2. Is a particular keyword more frequent in one time-period or another (e.g. within a specific year, decade, or longer historical period e.g. the Victorian era, post-1945, pre-1989, etc.)?
3. Are there time-periods in which the keyword does not feature at all?
4. Can certain adjectives/nouns/verbs be found to collocate more frequently with the keyword (in a particular word-position) in those corpus texts produced within one time-period versus those produced in an earlier or later time-period?
5. Are the changes in the relative frequency of a keyword over time similar or different to the patterns observed with regard to other keywords?
6. To what extent can these patterns be explained by other factors (especially those pertaining to the construction of the corpus itself e.g. the uneven distribution of tokens across the corpus as a whole)?

These questions, particularly 1–4, address similar issues to those categorized as *Keyword* and *Collocational Patterns* questions. The difference is that these questions need to be answered with regard to the date of a corpus text's publication rather than its filename. Now using just the concordance browser, it seems essential to use subcorpus selections and a spreadsheet or notepad to analyse the patterns across time. Building this metadata frequency information into the tool and using it to interact with and explore the corpus becomes an obvious design goal to solve these issues. Question 5, by contrast, suggests some form of keyword extraction tool would be helpful, where keywords with

similar temporal frequency profiles could be identified and grouped together. Question 6 seems to require expert interpretation of the results of other questions in the context of domain knowledge and some understanding of the limitations/design of the corpus.

The second researcher, Jan Buts (JB), decided to split the 20 questions into four categories *Keyword*, *Text*, *Author* and *Corpus*. The sections are ordered by importance, as are the questions within each. The researcher was keen to point out that the questions categorized within different sections will often intertwine and that the importance ranking is only an approximation.

It is interesting that *Keyword* is again given as a topic and is presented as most important by this researcher:

Keyword ("i.e. in case one wants to study a specific keyword in any number of texts")

1. How frequent is the keyword, and where is it ranked in a frequency list?
2. With which words is the keyword most frequently combined, in a span of four positions to the left and right?
3. What is the approximate strength of the collocational patterns observed?
4. Are there intuitive variations of the keyword (both formally and semantically) that occupy similar positions and display similar collocational patterns?
5. Which position does the keyword take in the clause, the sentence, and the text?

Questions 1 and 2 simply deal with keyword frequency and collocation frequency. Collocation strength is again mentioned explicitly in question 3. This is noteworthy as the concordance browser does not facilitate easy investigation of collocation strength. Question 4 calls for analysis of the formal and semantic variations of the keyword, and suggests that an understanding of what these variations might be comes from intuition. It is possible these questions could be also answered by the automatic keyword suggestion recommendations proposed by the first researcher. For question 5, discussion with the researcher revealed that reading each CL individually, using the 'extract' function of the GoK software to expand the context provided in relation to specific lines, and/or searching the original text externally from the corpus were the methods used at each level.

The first question listed under the heading *Text* concerns word frequency in a text and comparisons with other texts or sets of texts:

Text ("i.e. in case one wants to uncover the properties of a certain text")

1. What are the most frequent content words in the text, and how does this compare to other texts of a similar character?
2. What are the most frequent function words and connective elements in the texts, and with which of the content words above do they recurrently combine?
3. What are the most common proper names in the text?
4. Having identified all the above, do they vary in their dispersion across the document?
5. Having established all the above, where are (dis-)continuities situated in the text? (For instance, does the introduction display a different textual character than the body of the text).

The generation of word frequency lists was discussed as a way of answering this question. However, comparing raw frequencies within lists generated from subcorpora of different sizes can be problematic. It would not be a fair comparison, for example, to

ARTICLE

PALGRAVE COMMUNICATIONS | <https://doi.org/10.1057/s41599-020-0423-6>

contrast the raw frequency of a word in an article of 2500 words with its raw frequency in a book of 100,000 words; some way of automating the calculation of normalized frequencies (e.g. 5 instances per 1000 words) would consequently appear to be an important requirement. Question 2 again makes use of a frequency list to identify frequent words with certain linguistic properties. After the function words have been identified, collocation patterns are analysed. Question 3, on the other hand, relates to proper names and would require manual annotation or some automatic method which recognizes named entities. Finally, questions 4 and 5 are concerned with identifying the position of words within a text with the aim of identifying patterns of dispersion. Visualizing the spatial component of a text and highlighting terms of interest was suggested as a solution. Lexical dispersion plots such as those available in *WordSmith tools*, which display the relative locations of the occurrences of selected words in a text or corpus, are an example of a visualization which helps answer these questions.

Under the *Author* heading four of the five questions are concerned with frequency and collocation strength differences:

Author (“i.e. in case one wants to construct a profile for an author with multiple texts in the corpus”)

1. Which words are the most frequent in each individual text written by the author in question, and how do this compare to the overall frequency of words in all the author’s texts combined?
2. Which words does the author use significantly often in comparison to other authors similar in temporal, spatial, linguistic, or social context?
3. Who does the author frequently cite?
4. Which multi-word expressions occur significantly often?
5. Given all the above, are there temporal changes to be observed in the author’s textual profile?

The questions compare an author to other authors, individual texts of the author, temporal profiles and various other metadata based subcorpus selection options. To answer these questions using the concordance browser would be time consuming, requiring many searches and subcorpus selections combined with note taking or spreadsheets. Question 3 would be especially difficult to answer using existing software: in order to make this process more efficient, the tool would need to be able identify citations and annotate them with metadata tags correctly, a feature that is not well supported in collocation and concordance analysis software. The alternative to automatic methods is manual annotation and linking.

Questions about the *Corpus* focus first on identifying frequent words or patterns, then investigating the collocations of those words and patterns:

Corpus (“i.e. in case one wants to interrogate a corpus varied in textual material”)

1. What are the most frequent words, collocations, and other multi-word expressions in the corpus?
2. Can the frequency of the above be attributed to a limited number of texts, or is it characteristic of the corpus as a whole?
3. If the texts in the corpus display varied patterns regarding the above, how are relevant keywords, collocations, and multi-word expressions distributed across the corpus in terms of publication date, source language, author, etc.
4. What can one say about the specificity of the corpus in question in comparison to another specific corpus? (For instance, do certain keywords occur very often in all texts

studied, while being very low-frequency in another varied corpus set).

5. Are the texts in the corpus explicitly connected through quotation or other types of reference?

Frequency lists and concordance analysis are employed to answer the questions, before moving to examine the distribution of those patterns of interest across the metadata attributes of the corpus. subcorpus searches, note taking and concordance analysis with a focus on frequency are the main techniques that would be useful.

Question 4 could be answered by comparing frequency lists. As we had previously discussed there are limitations when comparing word frequencies across two lists of different size, and our proposed solution of calculating instances per 1000 words does not take into account distribution of word frequencies. So, if for example a small number of words account for the majority of the occurrences the normalized frequencies for most words will be small and close together. This makes comparisons of word usage difficult and misleading. Solving the disconnect between the raw frequency and the distribution could help with frequency list comparison.

From these two sets of questions provided by the two researchers, we have identified some of the goals, tasks and techniques of our colleagues on the GoK project. We have found many instances in which information related to word frequency in a concordance would be useful. In particular, the examination of frequency information is not limited to the keyword of interest, but typically involves comparison and analysis in relation to other texts and subcorpora, as well as frequencies of collocates. This type of analysis was of high importance in the questions elicited from the researchers. During the discussions we established that these frequencies tend to be estimated visually from a concordance or manually counted. Supporting these actions through visualization was identified as a potential area which could benefit the GoK project and corpus analysis in general. Comparison of frequency lists is another area that was identified as a candidate for tool support. Frequency lists came up quite frequently in the discussion and the comparison of frequency lists was raised as a useful means of gaining insight into the particularities of a subcorpus selection. As noted, however, this comparison can be difficult in practice. Issues around the comparability of different sized lists cause problems, due to frequency and rank not being easily comparable for different sized lists. While this problem was not identified as the most important issue by the researchers, it does seem to be a problem which one cannot address without additional tool support. The third issue for which we considered creating visualization based solutions was metadata based frequency analysis. However, the methods currently used to answer questions related to metadata make use of external tools such as spreadsheets or notes and the analysis of several concordance queries together. The precise tasks which we would be attempting to facilitate were at this point not fully clarified. For this reason we gave priority to the other issues identified.

Software prototyping. At the start of the project, one of the authors (SL), attended meetings where the project team members discussed aspects of the scholarly work to be carried out, and the core research questions to be investigated. This provided the developers with basic intuitions as to how the tasks analysed in section “Analysis of published methodology” could be employed to support the kinds of investigation described in section “Observational research”. We then employed low-fidelity prototyping and user interface sketching methods to communicate these intuitions to the research team and create initial designs for

what would eventually become the GoK tools. Note that by “fidelity”, here, we mean how close to a finished product the prototype is. Low-fidelity prototypes or ‘mock-ups’ often take the form of paper prototypes, they are quick to design and easy to alter. An early version of the concordance mosaic visualization was sketched along with different representations of metadata, to be incorporated to the basic software platform (see section “The GoK Software”). Initial ideas for frequency comparison functions were also discussed.

The next step was to implement mock-ups and “bare-bones” working prototypes, illustrating how the tools might work at the level of the user interface. Some of these ideas eventually developed into the tools described next.

The GoK software. Concordancing software (or “concordancers”) are common currency in corpus based studies. While arranging and indexing fragments of text for comparison and study dates from antiquity, the advent of computers enabled the systematic creation of *concordances* through the “keyword-in-context” (KWIC) indexing technique, as well as dramatically increasing the volumes of text that can be analysed. Concordancers have since become widely used tools across a broad variety of fields of research, from studies of lexicography and linguistics, to narratology, discourse analysis, and translation studies. The GoK software is based on a set of language processing software libraries (modnlp (Modnlp, 2020)) and a basic concordancer, which have been used in a number of projects (Luz, 2011). This infrastructure supports tasks such as indexing, data storage, metadata management, access and copyright compliance, as well as the basic user interface. As this has been described elsewhere (Luz, 2011), here we will focus on the visualization tools built for GoK specifically. This section gives a brief overview of the software and how it is used, before we delve into its design process in the following sections.

The modnlp client software provides a traditional concordancer interface, comprising frequency lists, a concordance display which allows sorting the CLs according to words to the right or left of the keyword, and functions to display text “metadata” (e.g. title, author, publication date, source language, subcorpus) and restrict the search and display of data based on these metadata. The basic modnlp concordancer is shown in Fig. 3. Through the concordancer the user can search for specific words, word sequences or “wildcards”. The results are presented in a KWIC style, with the keyword aligned in the central column and surrounded by its immediate “context” (that is, the words that occur to the right and left of the keyword). Also included in the far left-hand column of the interface is the filename of the text in which each CL can be found. The concordance may be sorted by word position, but also by filename. Lines may be removed from the concordance by the user as necessary.

While the concordancer shows all occurrences of the string of characters searched for (refugee in the case of Fig. 4), the concordances rarely fit a single screen, forcing the user to sort and scroll in order to discover possible collocation patterns. The following tools, implemented as “plug-ins” to the modnlp system, were designed to aid this pattern discovery process, addressing the questions and tasks discussed in sections “Analysis of published methodology” and “Establishing initial requirements”.

The concordance mosaic. The *concordance mosaic* tool (referred to as Mosaic for short) provides a concise summary of the KWIC display by presenting it in a tabular, space-filling format which fits a single screen. Mosaic is able to fit hundreds, often thousands of CLs onto a small display by taking advantage of the fact that a small number of types tends to dominate the distribution of

tokens at each position in the concordance’s context with respect to the keyword (a trend known as Zipf’s law (Baek et al., 2011)) and therefore can be represented by a single object (type) on the screen, rather than a repetition of tokens. Thus Mosaic represents positions relative to the keyword as ordered columns of tiles. The design is based on temporal mosaics which were originally developed to display time-based data (Luz and Masoodian, 2004). Each tile represents a word at a position relative to the keyword. The height of each tile is proportional to the word statistic at that position. These tiles can be compared across all positions to evaluate quantitative differences between positional usage. However, the display option labelled *Collocation Strength (Local)* intentionally breaks this cross positional linkage and only allows comparison between tiles at the same position. Colors are used to differentiate between the frequency and collocation strength views of the concordance list. In its simplest form each tile represents the frequency of a word at a position relative to the keyword. In Fig. 4 the Mosaic generated for the keyword *refugee* as it is found in the GoK Internet corpus is presented. The tool is set to display a collocation statistic (cubic mutual information), which emphasizes higher-than-expected word frequencies. Due to the strong visual metaphor of KWIC it should be clear the word *anti* is the most salient (though not the most frequent) word immediately to the left of the keyword (K−1), and that *crisis* is the most salient word immediately to the right of the keyword (K+1); see Baker (this volume). Hovering over any tile will display a tool-tip with the word count and frequency at the relevant position. This relieves the need for manually counting or performing additional searches to retrieve position based word frequencies.

Alternative displays by relative frequency, relative frequency with very frequent words (stop-words such as *the*, *of*, and) removed, and scaled according to global statistics (rather than scaled to fill the space of each column, as in Fig. 4) are also available. The user additionally has the option of applying a range of different collocation statistics, such as mutual information, log-log and z-score (Pecina, 2010).

Concordance tree. The economy of representation in Mosaic comes at the cost of sentence structure. Mosaic is based on an abstraction of the CLs as a graph, where types are connected with other types in linear succession, with each path in the graph corresponding to a sentence in the collocation (Luz and Sheehan, 2014). However, the tabular layout of Mosaic as juxtaposed tiles does not encode these paths visually, and consequently it is impossible for the user to know at first glance whether two words that occur next to each other on Mosaic also co-occur in the same sentence. For example, the words *de-politicize* and *Syrian* occur next to each other on the mosaic of Fig. 4 (K−2 and K−1, respectively) but there are no sentences in this concordance that feature the expression *de-politicize Syrian*. To discover which sentences (paths) exist for a given word the user must click on that word and observe which collocates appear highlighted (as white tiles) on Mosaic.

An alternative rendering of the underlying graph structure of the concordances is the *concordance tree*. This tool is a variant of the Word Tree design, introduced by Wattenberg and Viégas (2008). It shows the left or right context of a concordance as a prefix tree, where each branch (path along the graph, from root to leaves) corresponds to a sentence in the concordance. The font size of each word at a particular position on the branch is scaled according to the frequency of occurrence of that word at that position.

A fragment of a concordance tree corresponding to the right context of the *refugee* concordance is shown in Fig. 5. While the concordance tree preserves sentence structure, it cannot generally

ARTICLE

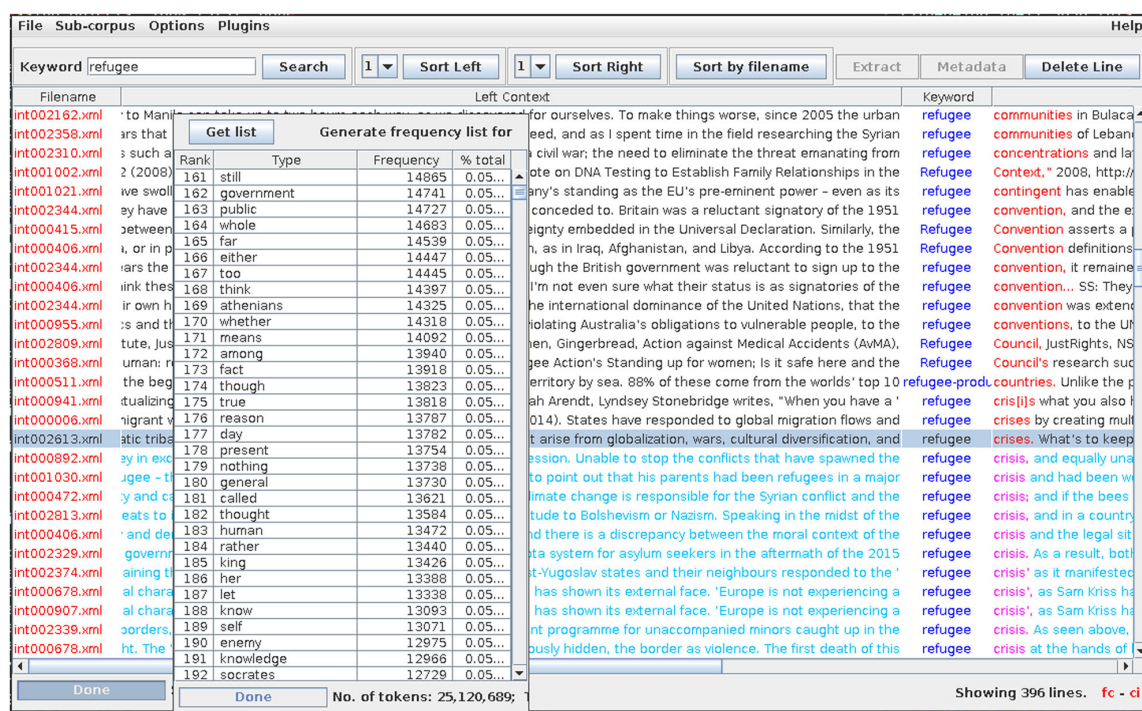
PALGRAVE COMMUNICATIONS | <https://doi.org/10.1057/s41599-020-0423-6>

Fig. 3 Modnlp concordancer and frequency list. Highlighted items have been sorted and selected by the user using Mosaic.

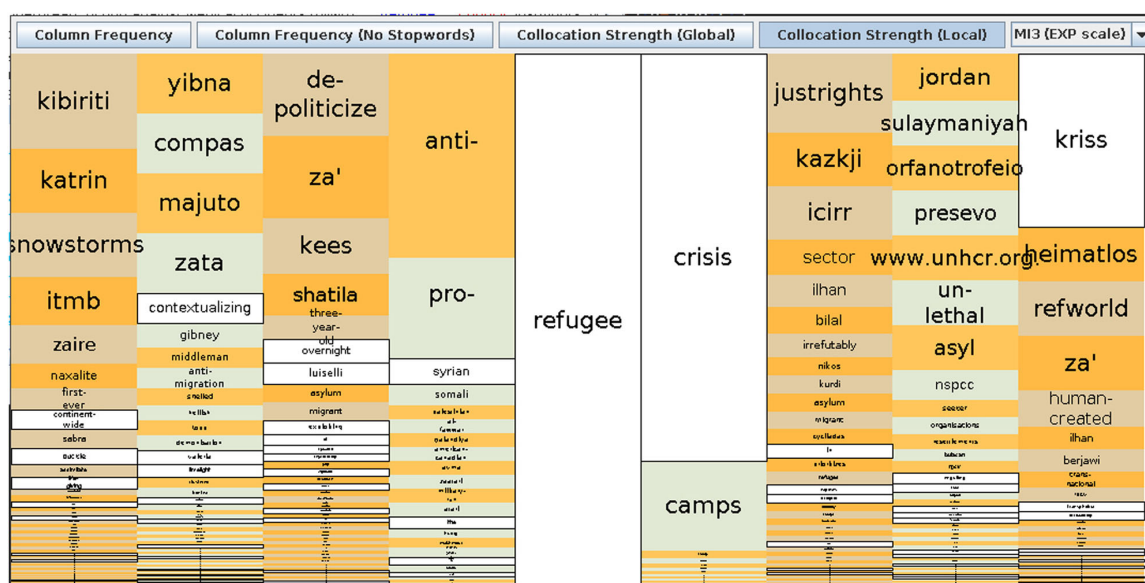


Fig. 4 Mosaic visualization for the concordance of the word "refugee" in the Genealogies of Knowledge corpus. The word "crisis" has been highlighted, causing all words that occur in sentences containing the word crisis at position K+1 to be highlighted as well, and the concordancer to display those occurrences in full context (Fig. 3).

display a full concordance as compactly as Mosaic. In fact, in the worst case it can take up as much space as a full textual concordance. However, as the purpose of visualization is to emphasize frequently occurring patterns, one will typically "prune away" low frequency branches (using the depth-1 width slider; the top right corner of Fig. 5).

Metafacet. The concordancer allows the user to retrieve metadata about each file and section on a line by line basis. This is however a time consuming and challenging process for the corpus analyst if the metadata of a large number of lines need to be investigated. To better support the exploration of metadata in conjunction with concordancing we have implemented the Metafacet tool,

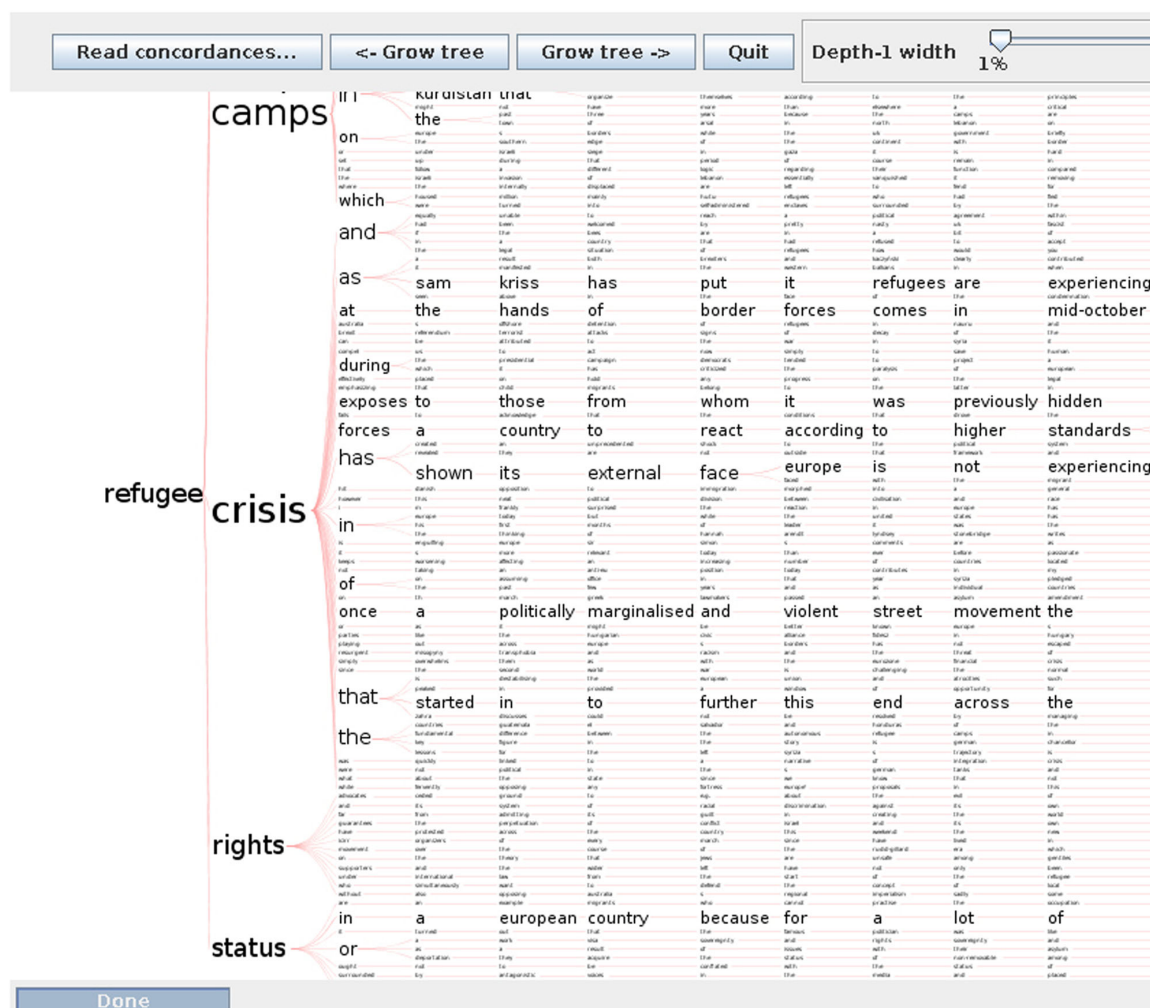


Fig. 5 Fragment of a concordance tree showing the right context of the word *refugee* in the concordance shown in Fig. 4. Concordance tree (fragment) showing the right context of the word *refugee* in the concordance shown in Fig. 4.

which provides a faceted summary of the available metadata. It also allows for the interactive filtering of the concordance list and the Mosaic using all available metadata facets (Sheehan and Luz, 2019).

The interface uses a horizontal bar chart to display CL frequencies per metadata attribute. Colour is used to help with visual comparison of bar length but the gradient otherwise encodes no special meaning. An attribute is a possible value that a metadata facet can take. For example, “The Nation” being the name of an online magazine whose contents the GoK team has included in the GoK Internet corpus, is an attribute of the facet “Internet outlet”.

Figure 6 shows a Metafacet chart for a concordance of *refugee* generated from the Internet corpus, with “OpenDemocracy” selected as the internet outlet for analysis. A drop-down list is used to choose which facet is displayed and the bars can be sorted by frequency or alphabetical order. Moreover, the visualization window can itself be filtered using a sliding scale (positioned on the far right) in order to allow the user to view a smaller portion of the attributes. This conforms to the common visualization design pattern of overview plus detail on demand, whereby users

are initially presented with a general summary of the dataset being analysed but retain the option of conducting finer grain analyses according to their interests (Shneiderman, 1996).

Metafacet, when used on its own, provides an interface with which to explore keyword distribution across different metadata attributes. By combining it with the concordance list and Mosaic, the user can navigate the corpus in a new way, viewing the concordance as attributed sets of collocations that can be interactively filtered, sorted and examined.

Frequency comparison tool. The modnlp frequency list shown in Fig. 3 provides detailed statistics on term frequency overall or by subcorpora. However, it does not allow easy comparison of frequencies in different subcorpora. The *frequency comparison* tool allows frequency lists to be compared visually in a statistically valid manner. The functionality of the tool has been described elsewhere (Sheehan et al., 2018), and it has since been modified to operate as a plugin for modnlp and is briefly presented here.

The modnlp concordancer has a subcorpus selection interface which can be used to save named subcorpora for later reuse. These named subcorpora then become available for comparison

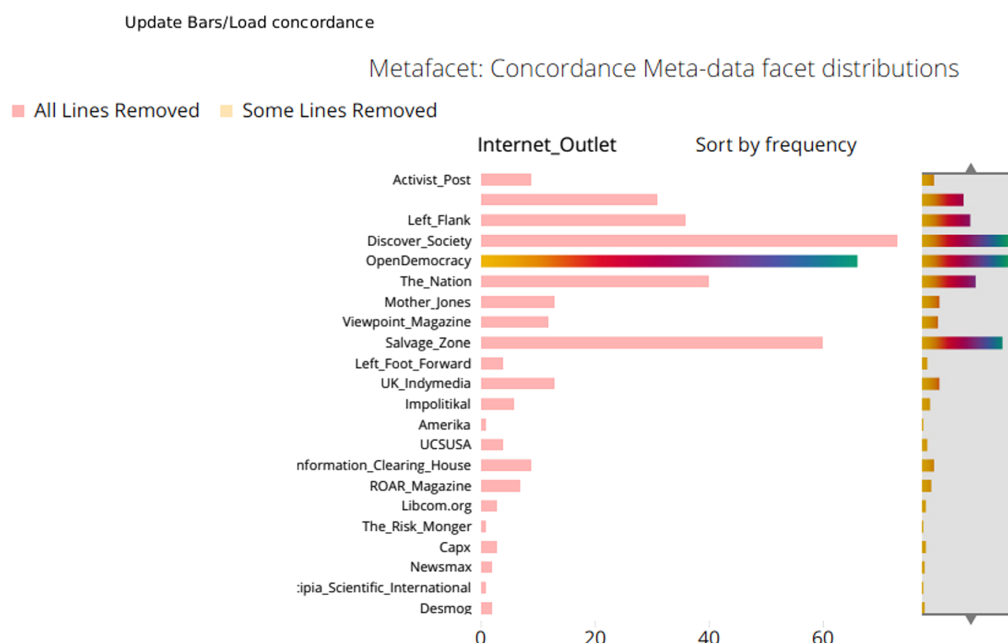


Fig. 6 Metafacet tool showing the attributes (in this case, publication titles) of the facet “Internet outlets”.

through the frequency comparison tool. Figure 7 shows a comparison of two pairs of outlets from the GoK Internet Corpus. Frequency information for the outlets *ROAR Magazine* and *Salvage Zone* is displayed on the left, these outlets explicitly adopt a more radical left agenda than others like *The Nation* and *Open Democracy* which are displayed on the right. In this diagram, both axes are log scaled which should yield a linear frequency diagram if the word frequencies follow a Zipfian distribution. Scaling both ranked lists to the same height and comparing a word’s position in the distributions enables the user to compare subcorpora of vastly different sizes.

In Fig. 7, the majority of the words have been filtered out to reveal the words with the greatest frequency changes between the two corpora. The words are placed at heights that correspond to the rank order in which they occur in the frequency distributions in their respective corpora. The lines in the middle connect words to the positions in which they appear in the other corpus’ ranked distribution. Thus the diagram shows the word “capitalist” is ranked just below 100 (with a frequency of 0.096%) in the radical left outlets, and at nearly 4000th (with a frequency of 0.003%) in the less radical pair. In this case the total number of unique tokens is roughly the same at ~10,000 tokens. The comparison would still be possible if the scales were vastly different, if one of the subcorpora was much larger than the other. The nature of the subcorpora should be evident from the differences in the frequency distributions of these words.

The corpus exploration facilitated by the frequency comparison tool supports hypothesis discovery activities, such as questions 3 and 6 in HJ’s keyword related questions list (in “Keywords”), as well as questions 2 and 5, relating to temporal spread (in “Temporal Spread”), as time intervals can be used to define subcorpora for comparison. Similarly, this tool allows JB to investigate question 2 of his author question list (in “Author”) and questions 2 and 4 of his corpus list (in “Author”). See also the papers by Baker (2020) and Buts (2020) for examples of uses of the frequency comparison tool in scholarly work.

The images of the above described visualizations were checked using the Coblis color blindness tool (Coblis, 2020). For red, green and blue, the displays are readable without issue. Differentiating between the collocation strength and frequency views for red or green, can be slightly challenging for colour blind individuals as the colour profiles become similar. However, there is enough difference in saturation to tell them apart. In addition, the button shading for the selected interface helps the user identify the option that is selected. This works even for monochromatic images. As colour is used only to help differentiate between items and not as a visual encoding of a data attribute the exact color that displays is of minor importance.

Observational research. To get a better understanding of the techniques and methods used by researchers in the GoK project, we requested time to observe research where the concordance browser was in use. Early versions of Mosaic and Compare Frequencies tools had additionally been integrated into the software by this point.

Interview and case study on the concept of democracy. A researcher from the GoK project (Jan Buts) offered to discuss his methodology and give an outline of the typical analysis process. The initial discussion was not based on a specific case study but describes the general method employed by Jan. After the initial discussion Jan agreed to let us observe a partial re-enactment of an analysis which had already been performed: this related to historical changes in the use and meanings of the concept of *democracy*. Jans’s methodology was described as the search for the largest unit of meaning related to a keyword. Meaning in this case should be constructed from the evidence present in the corpus. The corpus is central to the analysis and the technique is in the style of Sinclair (2003), meaning the investigation of collocation, colligation, semantic preference and semantic prosody is performed by an analyst who must be make a conscious effort to

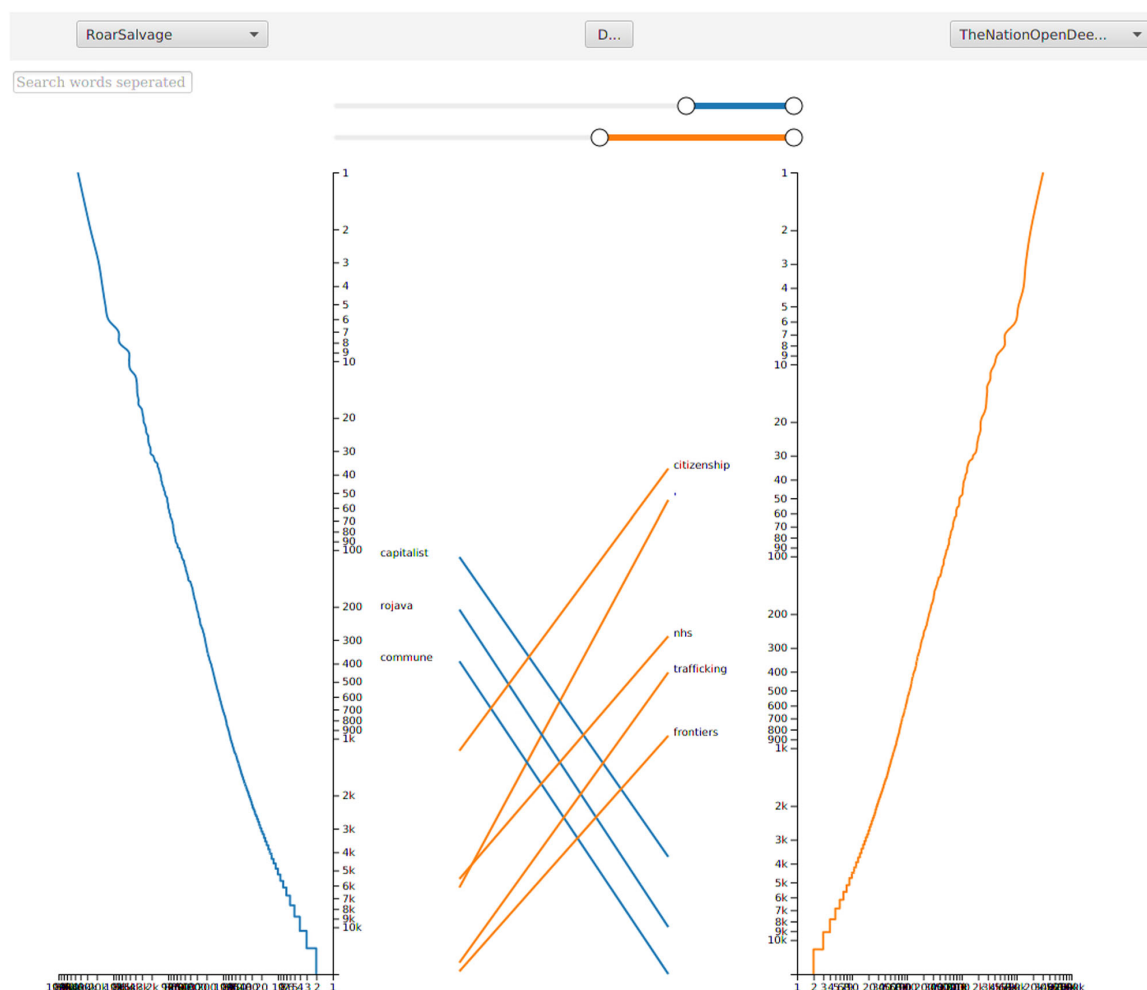


Fig. 7 Compare Frequencies visualization showing words with the largest change in usage between the outlet pair ROAR Magazine and Salvage Zone (left) and the outlet pair The Nation and Open Democracy (right).

ignore personal bias. A list of the ordered steps used to perform this type of analysis was provided:

Construction of meaning:

- Sample
- Describe Patterns
- Sample
- Compare Patterns
- Hypothesize

In the steps provided the term “sample” refers to a subcorpus selection and keyword search in a concordance browser: if the concordance is large, the samples may be thought of as subsets of the full concordance list. “Describe Patterns” refers to the process of analyzing the positional frequencies around the keyword. Jan explained that the analysis begins by looking at the patterns of words occurring next to the keyword and expands to additional positions until the discovered patterns describe the meaning of most CLs. The remaining lines would then be analyzed after the core units of meaning had been established. “Compare Patterns” is the process of examining the differences and similarities

between the described patterns of the samples. The following clarifying question was posed to Jan:

- Can you give practical details of your typical methodological approach?

“Investigate the keyword and its neighbouring collocates (Left and right +1). Investigate deviations from frequent patterns then expand the analysis horizon and repeat until the largest unit of meaning is found. Largest unit of meaning should be read as ‘overarching’, in the sense that the point is not to necessarily go beyond the concordance line, but to construct an abstract unit that can account for as many concordance lines as possible. If interesting patterns which lead to a hypothesis are discovered pursue these. Typical corpus linguistics method applied to unique corpus”.

During the discussion some difficulties were reported in relation to working with the GoK corpus. The unique nature of the corpus makes the generalization and the representativeness of

ARTICLE

PALGRAVE COMMUNICATIONS | <https://doi.org/10.1057/s41599-020-0423-6>

a hypothesis more difficult to explain. Viewing metadata for the CLs is useful but, as it is line specific, it is impractical for analyzing large numbers of CLs. The concordance browser's use for identifying patterns in large numbers of CLs requires sampling multiple times. A broader picture of the concordance which can examine broader and more restricted contexts would be desirable. New tools should complement the concordance, extending its functionality rather than seeking to replace it. The statement that the unique nature of the corpus caused problems required clarification. The following question was posed in relation to this issue:

- How and why does the corpus influence the methodology?

"If speaking about 'typical/traditional' corpus linguistics (which is always a bit of a stretch), one finds actual/practical lexicography, and analysis of register/text-type, etc., drawing on corpora that are constructed to serve as a sample of the full language or sub-language under investigation. Think of the British National Corpus, for example.

Our corpus hosts a variety of texts, but it would be difficult to make the case that it is representative of anything outside the corpus itself. Our Internet corpus, for instance, is not a sample that can tell something meaningful about the Internet 'as a whole'. Therefore, rather than making exhaustive analyses of a certain word across the corpus, or tracing a grammatical pattern across its contents, the corpus urges one to study a specific subset of texts, and to complement the findings with sources outside of its confines to make hypotheses about conceptual developments. Consequently, the method will be less repetitive than one would traditionally see, and more meandering, to a certain extent".

When asked at the end of the interview if it would be possible to observe the method in the future, Jan volunteered to give a demonstration of some analysis. The demonstration which we observed was a partial re-enactment of an analysis which had already been performed. The concept of democracy was investigated in a subcorpus of political texts published in English from 1970 onward. Jan commented that "this is in line with the most fundamental goals of the GoK project". The steps taken which were observed and recorded were:

- Begin by searching the keyword *democracy* without any bias for what will be returned;
- Open Mosaic and see if anything stands out (nothing does);
- Look at Column Frequency (No Stopwords) view. Social democracy appears to be a very strong collocation. Click social and look at the CLs now highlighted in the concordance browser;
- Reading the lines reveals that *Social democracy* only occurs in file mod000008 and refers to one book title and its contents. This is only informative about this specific file and the file is then removed from the subcorpus under investigation for the sake of gaining a more balanced overview. (This appeared an unusual step and was recorded as needing further clarification);
- Re-run the search this time ~500 lines were removed from the concordance. *Common* and *Athenian* were recorded as important collocates;
- Mosaic is consulted again, both the Column Frequency and the Column Frequency (No Stopwords) views. These do not seem to show any unexpectedly frequent results;
- Navigate to Collocation Strength (Global) view and

investigate the words one position to the left of the keyword;

- Do any of these extreme combinations also have interesting frequency profiles (not single occurrences in the concordance)? Investigate by looking for words which stand out in the Mosaic Column Frequency and Collocation Strength views;
- Did not find any particularly interesting frequent and strong collocations at position left +1;
- Use a regular expression to search for "-acy". Interested in keyword frequency and collocations;
- Note *democracy* is 76% of "-acy" occurrences.
- Looking at other frequent keywords (*aristocracy*, *bureaucracy*): they are mostly negatively framed in the CLs;
- Switch to concordance strength view and observe that the highest ranked keyword is *mediaocracy*;
- Search *mediaocracy* 10 lines returned;
- Use Column Frequency (No Stopwords) view of Mosaic and the concordance browser to establish the semantic prosody of the term, that is, whether it is used negatively or positively;
- Hypothesis: Democracy is the dominant "-acy" and is viewed in a positive way. All other "-acy" are presented as negative. They are presented as threats to democracy.

This description reveals heavy use of Mosaic for analysis. The case study presented seemed to be a partial treatment of the problem and may have skipped some steps which were needed to reach the hypothesis. Jan was asked the following clarifying questions:

- You moved swiftly from removing the file mod8 to investigating collocation strength. After removing the file mod8 you did not re-investigate the collocation frequency of *democracy* and instead moved on to collocation strength. Why?

"Just for demonstration purposes. In essence, not only were pieces skipped over, the illustration was also fairly preliminary in the following sense: Removing mod8 because it creates some distortion is of course bad practice [if this were the actual research]. The point in doing so is to quickly weed out material unfit for my purposes, until I reach a suitable point of investigation (in this case: democracy turning from one of the competing systems of rule into the only one available, however constantly beleaguered by threats from within). Once this point of investigation is established, the analysis can start out again and I make sure to construct a suitable subcorpus on clearly defined terms that doesn't require me to be rash at the outset of an analysis. The mosaic view can then be approached again as an entry into the data, and all the collocation patterns examined more closely".

- How do you decide what subcorpus to initially investigate? In this analysis books from 1970 to present date.

Currently the first thing I do (especially when the concordance return is small) is look at overlaps in meta-data property between concordance lines, to get a sense of the whereabouts of the data.

- Would a visualization which shows frequency of a keyword across meta-data facets be useful?

"Yes. One could, for example, look at differences in dispersion in the use of the word 'terror' pre- and post-9/11, look at whether a certain author evades a word (say, anarchy) that is used by all other authors writing on the

same subject (say, democracy), one could look at whether a magazine has a regional, national or international outlook by comparing the proper names used with those in other magazine, etc”.

• In your analysis I struggle to see why you began analyzing the -acy concordance. It does not appear to follow from the previous steps of analysis. Is this an established next step in corpus linguistics? Is it based on experience and domain knowledge or some part of the analysis not presented?

“This has to do with the reduction of bias through the reliance on form. I could, for example, go look at democracy vs. totalitarianism (in my attempt to study contemporary forms of government), but I have no proof that these concepts in fact are alternatives to each other. This would be solely based on intuitions, and as a lot can be argued about language data, I would basically come to prefabricated conclusions if I wanted to (democracy is opposed to totalitarianism in the following senses). Starting out from taking the suffix -cracy and seeing what other terms it attaches to offers a more neutral entry into the data inspired by the actual linguistic form rather than pre-conceived oppositions”.

• You did not appear to investigate the collocations of democracy and other (-acys) to determine the usage or context in which they occur, except for meritocracy, I am assuming that this was done and just shown?

Indeed, in the final analysis every term discussed merits close attention to the immediate co-text.

• You use Mosaic extensively in the method? Is that typical of your work

I use the Mosaic every time I access the corpus. Especially at the beginning of an analysis, to get an idea where to start and to make sure I won't, in a later stage, overlook any significant patterns.

• You appear to use the collocation strength view for analysis, what is your opinion on it?

Useful for analysis as it gives extreme combinations. (where the combination rarity is interesting). As it stands the analysis done using the Collocation strength view is difficult to explain. Justifications for the patterns found using this view are usually easier to re-frame as part of the qualitative analysis which involves reading the concordance lines.

• If other statistical measures were available in Mosaic would that be useful?

Yes, we would benefit from a measure of confidence rather than strength, or from a commonly known measure that can simply be mentioned as such in publications.

Case study on the concept of “the people”. In this case study we observed an analysis of keywords related to the concept of “the people” featuring in a subcorpus of eight different English translations of Thucydides’ *History of the Peloponnesian War*. The researcher (HJ) told us this was early stage exploratory research that, he hoped, would ultimately lead to a publication. The details of the think aloud observation session and interview

can be found in Sheehan and Luz (2019). While think aloud user studies are one of the most common user study techniques they need to be carefully controlled. In performing the study, we carefully explained what was required by Henry. That he should not try to think about what he was going to say and should instead try to narrate what he was doing in real-time. We were careful to not bias the study via prompts, we used repeated phrases from a script to help avoid leading questions and interviewer bias.

To summarize the session, the observed method consisted of an analysis of multiple keywords, related to the concept of “the people” using frequency and collocations. The method made use of a concordance browser to select the subcorpora, retrieve the keyword frequencies and to help list the most frequent collocates of the keywords. A spreadsheet was used to keep track of the keyword frequencies per translation and to list the collocates of interest. The process was time consuming and the researcher would benefit from automated methods of extracting the required numerical information, however he was unaware of any tools which could help with this type of analysis when the corpus must be made available through concordance and not as full texts. When investigating the frequent collocates the mosaic was used to save time counting or estimating frequency from the concordance list.

At the end of the observation session Henry explained how the analysis would progress beyond what was observed. For each file and keyword combination the collocation patterns would be identified using the observed technique. Clearly this will be very repetitive and time consuming. Following the collection of these results, the next stage would be to look at the frequency patterns using the table of results. Making bar charts in a spreadsheet application or external tool will be helpful for examining the trends. Temporal patterns are expected but any identified patterns will be investigated using qualitative analysis, which involves reading the CLs which relate to the identified pattern. Understanding the meaning of the concept of *the people* at different times is the goal of these analysis steps. Any differences identified, temporal or otherwise, must take into account individual translator’s style, the political context and many other factors. The researcher’s knowledge of the domain, the corpus and the individual texts is essential to the analysis. For example, the observation made during discussion with Henry that “There are no translations 1919–1998, during period of huge cultural change in Britain. Possible reasons for this include Suffrage, war or technological revolution” would be difficult to derive from the concordancer as it relates to an absence of texts in the Modern English corpus. Henry explained that information about the authors and texts will influence the analysis. Some examples of the type of information which can be relevant are “the political leanings of the translators which is established relevant knowledge” and knowing that “certain texts are partial translations, abridged versions etc.”.

In regard to methodology, Henry described a process of exploration using the corpus tools and pattern search, drawing on the actions described in section “Analysis of published methodology”. Common lexical knowledge and knowledge of the literature on specific concepts (e.g. the people) helped identify keywords for exploration. Similar patterns of exploration and use of the GoK tools are shared among other GoK scholars to investigate the role of translation in the evolution of political and scientific discourse. Moreover, Henry believes aspects of this methodology can be used by other projects that use corpora. Lack of familiarity with the software, and lack of documentation were mentioned as barriers to the further development and wider adoption of the methodology. However, tools such as Mosaic were considered intuitive and helpful in “any investigation of

ARTICLE

PALGRAVE COMMUNICATIONS | <https://doi.org/10.1057/s41599-020-0423-6>

collocations, as it tells you in a very quick and transparent way which are the most common collocates in each word position for a given keyword". A somewhat different investigation pattern was observed in the next case study.

Case study on the concept of statesmanship. This observation session took place after a piece of analytical work had already been completed, but Henry offered to explain and re-enact a portion of the investigation for the purposes of this study. The analysis we observed contributed to a publication (Jones, 2019). The following is a summary of the observed analysis and explanation.

In the GoK Modern English corpus the term *statesman* was found to exist "almost exclusively (90%) in translations from Classical Greek". This pattern was not observed for other semantically connected keywords such as *governor*, *leader*, *ruler* and *citizen* which are more evenly distributed across translations from all source languages rather than only Classical Greek. The analysis which led to at this conclusion involved a simple keyword frequency comparison across the source language metadata recorded for each text in the corpus. This involved selecting the subcorpus of texts translated from each source language represented in the Modern English corpus individually and recording the number of CLs for the same keyword search.

A spreadsheet with an entry for each of the 261 files in this subcorpus was created and metadata (the author, the title, the translator and the date) was entered for each file. This was done manually and was time consuming. Henry explained that in this spreadsheet "the information could easily be (re)sorted according to each of these meta-data facets and patterns more easily identified". The number of CLs for each file was found by selecting a subcorpus consisting of a single file and searching for *Statesman*. Performing this action for each of the 261 files was very time consuming. After the spreadsheet has been filled in for each file the information could be analyzed to look for frequency patterns across the metadata attributes (such as author or date). Sorting and visualization (bar charts) were the main techniques used to get an overview of the identified patterns.

Henry described a process of exploration that aimed at determining the use of the term *statesman* and its frequency in comparison to semantically related terms. The main difficulty in this process was the time-consuming nature of the collection of frequency data. As analysis of collocations played a minor role in this context, the visualization tools were not used as frequently as in other cases. The most significant outcome of the two case studies was the emergence of the obvious need for a method to support the analysis of concordance lists through the lens of metadata. This topic had previously been raised in the requirements elicitation process. However, without this observational work, the requirements were too general and it was difficult to understand the low level actions we would need to support. From these two sessions it became clear that filtering the concordance list via metadata facets would be worthwhile. In addition having an instant breakdown of the number of CLs per metadata attribute would be useful. This observation session led to further discussion among team members and the eventual development of a metadata analysis tool which eventually became Metafacet. Another problem identified was that in the version of Mosaic available to the researchers at that time only a single collocation statistic was available and it was based on mutual information scores. Due to a lack of proper documentation, Henry did not know exactly what the scaling scheme was for the collocation strength view of Mosaic and so could not accurately interpret or use it for publication. This led to the writing of a detailed user manual and the addition to the Mosaic tool of optional scaling schemes based on well known collocation

metrics. More collocation measures are still being added to the tool.

Discussion

One of the main themes that emerged from the iterative design process presented in this paper is the role of visualization tools as a means of sensitizing the investigator to overall patterns which may not be easily captured by extensive sequential reading. Detecting such patterns is not however, as one might expect, merely a matter of compiling and interpreting text statistics. It blends data representation and statistical elements with aspects of the laborious reading and interpretation process that characterizes more traditional scholarly enquiry. Although certain statistical aspects of analysis have been raised by researchers, especially in connection with the discovery of collocation patterns through Mosaic, they are not seen as the main product, or even necessarily part of the main product of analysis. They rather serve the purpose of guiding the exploration when a corpus of text is represented as graphical summary. The analytical work remains essentially qualitative and interpretative.

The methods detailed by Sinclair are at the core of many research areas. We do not however expect them to capture the full range of tasks in these areas, or in a research project such as GoK. Researchers working with text build upon and sometimes diverge from the core methods as needed. It is not surprising then that we find differences between the methods identified in our hierarchical task analysis of Sinclair's work and our domain characterization efforts with the GoK project. Sinclair's tasks do not feature any comparative methods similar to those we observed. In particular subcorpora are not discussed at all. In our experience working with translation scholars this type of comparison is very common. Since these methods are not explicitly described in foundational texts it is less likely that they would be well supported by tools.

Avoidance of bias is another critical issue. The sparsity of language virtually guarantees that any sample (corpus), however large, will include an element of bias. While this is unavoidable, it is important that the text processing and visualization tools allow the researcher to identify possible sources of bias in text. This is important from two perspectives. First, identification of bias in the corpus, tracing it back to the texts and contexts in which it occurs is often an important element of analysis, and may sometimes be the object of analysis itself. Second, the highlighting of biases through overviews of textual patterns may help the researcher become aware of their own pre-conceptions that adversely affect the interpretation of data.

A use of the visualization tools which has not been explicitly discussed in the previous section but which featured in studies produced by project members (see Baker's and But's contributions to this special issue, for instance) is the use of graphics for communicating conclusions and viewpoints originating from corpus analysis. In communicating one's views to others through visualization, simplicity is imperative (Bertin, 1981). In this sense, the intuitive simplicity of Mosaic, for example, has enabled investigators to present much clearer illustrations of usage patterns than would be possible to do by means of tables and concordance displays. This is an aspect of the use of the GoK visualization tools that we would like to explore further.

Finally, it became apparent that tools need to be documented with the end users in mind, the language should be the language of the domain and not of the designer. Detailed documentation is needed and this documentation must be written collaboratively by developers with the help of users to ensure both its accuracy and its relevance to key research problems.

Conclusion

“Translator invisibility” is an issue with which many language scholars are familiar (Venuti, 1995). Venuti offers a critique of the view of “transparency” as an ideal in translation, regarding it as an illusory notion that tends to render the work of the translator invisible, in detriment of culture. Although this is not the place for an in-depth discussion of this argument, we note that in many ways software developers working with humanities researchers often find themselves in a similar position of inconspicuousness. Software tools and methods, much like statistical tests, are taken to provide an objective (transparent) means through which analytical work that is often of a markedly qualitative nature is done, and the developer or computational researcher regarded simply as the (invisible) “service provider”. The methods and analysis presented in this paper suggest that this is neither an accurate characterization of the developer’s contribution nor a desirable situation in interpretation projects. Effective interdisciplinary engagement is necessary if progress is to be made in the digital humanities.

Data availability

All data generated or analysed during this study are included in this published article.

Received: 30 October 2019; Accepted: 3 March 2020;

Published online: 23 March 2020

References

- Annett J (2003) Hierarchical task analysis. In: Erik H (ed.) *Handbook of cognitive task design*. 2. Lawrence Erlbaum Associates, New Jersey, pp 17–35
- Anthony L (2004) Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In: Anthony L, Fujita S, Harada Y (eds) *Proceedings of IWLLeL*, pp 7–13
- Baek SK, Bernhardtsson S, Minnhagen P (2011) Zipf’s law unzipped. *New J Phys* 13 (4):043004
- Baker M (1993) Corpus linguistics and translation studies: implications and applications. In: Baker M, Francis G, Tognini-Bonelli E (eds) *Text and technology: in honour of John Sinclair*. John Benjamins Publishing Company, pp 233–250
- Baker M (1993b) Corpus linguistics and translation studies: implications and applications, chapter 11. John Benjamins Publishing Company, Netherlands
- Baker M (1995) Corpora in translation studies: an overview and some suggestions for future research. *Target* 7:223–243
- Baker M (2020) Rehumanizing the migrant: the translated past as a resource for refashioning the contemporary discourse of the (radical) left. *Palgrave Commun* 6(1):1–16
- Baker P (2006) Using corpora in discourse analysis. Bloomsbury discourse. Bloomsbury Academic
- Bernardini S, Kenny D (2020) Corpora. In: Baker M, Saldanha G (eds) *The Routledge handbook of translation studies*. Routledge, pp 110–115
- Bertin J (1981) Graphics and graphic information-processing. de Gruyter
- Bertin J (1983) Semiology of graphics. University of Wisconsin Press
- Biber D, Douglas B, Biber P, Conrad S, Reppen R, University C (1998) *Corpus linguistics: investigating language structure and use*, cambridge approaches to linguistics. Cambridge University Press
- Bonelli ET (2010) Theoretical overview of the evolution of corpus linguistics. In: Anne O’K, Michael McC (eds) *The Routledge handbook of corpus linguistics*. Routledge, p 14
- Buts J (2020) Community and authority in ROAR Magazine. *Palgrave Commun* 6 (1):1–12
- Cleveland W, McGill R (1985) Graphical perception and graphical methods for analyzing scientific data. *Science* 229(4716):828–833
- Coblis (2020) Coblis color blindness tool. <https://www.color-blindness.com/coblis-color-blindness-simulator/>. Accessed Feb 2020
- Culy C, Lyding V (2010) Double tree: an advanced kwic visualization for expert users. In: Banissi E, Bertschi S, Burkhard R, Counsell J, Dastbaz M, Eppler M, Forsell C, Grinstein G, Johansson J, Jern M, Khosrowshahi F, Marchese FT, Maple C, Laing R, Cvek U, Trutschl M, Sarfraz M, Stuart L, Ursyn A, Wyeld TG (eds) *Proceedings of the 14th International conference on information visualisation (IV)*. pp 98–103
- Culy C, Lyding V (2011) Corpus clouds—facilitating text analysis by means of visualizations. In: Mariani J (ed.) *Human language technology. Challenges for computer science and linguistics*, vol. 6562 of Lecture notes in computer science. Springer, Berlin, Heidelberg, pp 351–360
- Culy C, Lyding V, Dittmann H (2011) Structured parallel coordinates: a visualization for analyzing structured language data. In: Pastor MC, Trellis AB (eds) *Proceedings of the 3rd international conference on corpus linguistics, CILC-11*. pp 485–493
- Frank AU, Ivanovic C, Mambrini F, Passarotti M, Sporleder C (eds) (2018) *Proceedings of the second workshop on Corpus-based Research in the Humanities CRH-2*, vol. 1 of Gerastree proceedings
- Green M (1998) Toward a perceptual science of multidimensional data visualization: Bertin and beyond. *ERGO/GERO Hum Factors Sci* 8:1–30
- Hareide L, Hofland K (2012) Compiling a norwegian-spanish parallel corpus. In: Andersen G (ed.) *Quantitative methods in corpus-based translation studies: a practical guide to descriptive translation research*, *Studies in corpus linguistics*. John Benjamins Publishing Company, pp 75–114
- Jänicke S, Scheuermann G (2017) On the visualization of hierarchical relations and tree structures with tagspheres. In: Braz J, Magnenat-Thalmann N, Richard P, Linsen L, Telea A, Battiato S, Imai F (eds) *Computer vision, imaging and computer graphics theory and applications*. Springer International Publishing, pp 199–219
- Jones H (2019) Searching for statesmanship: a corpus-based analysis of a translated political discourse. *Polis* 36(2):216–241
- Jänicke S, Blumenstein J, Rücker M, Zeckzer D, Scheuermann G (2018) Tagpies: comparative visualization of textual data. In: Telea A, Kerren A, Braz J (eds) *Proceedings of the 13th international joint conference on computer vision, imaging and computer graphics theory and applications*, vol. 2: IVAPP. INSTICC, SciTePress, pp 40–51
- Jänicke S, Franzini G, Cheema MF, Scheuermann G (2015) On close and distant reading in digital humanities: a survey and future challenges. In: Borgo R, Ganovelli F, Viola I (eds) *Proceedings of the Eurographics conference on Visualization (EuroVis)—STARs*. The Eurographics Association
- Kilgarriff A, Baisa V, Bušta J, Jakubíček M, Kovář V, Michelfeit J, Rychly P, Suchomel V (2014) The sketch engine: ten years on *Lexicography* 1(1):7–36
- Kilgarriff A, Rychly P, Smrz P, Tugwell D (2004) Itri-04-08 the sketch engine. *Inf Technol* 105:116
- Kocincová L, Jakubíček M, Kov V, Baisa V (2015) Interactive visualizations of corpus data in sketch engine. In: Grigonyte G, Clematide S, Utká A, Volk M (eds) *Proceedings of the workshop on innovative corpus query and visualization tools at NODALIDA 2015*. pp 17–22
- Léon J (2007) Meaning by collocation. In: *History of linguistics 2005*. John Benjamins, pp 404–415
- Luhn HP (1960) Key word-in-context index for technical literature (kwic index). *Am Doc* 11(4):288–295
- Luz S (2011) Web-based corpus software. In: Kruger A, Wallmach K, Munday J (eds) *Corpus-based translation studies—research and applications*, chapter 5. Continuum, pp 124–149
- Luz S, Masoodian M (2004) A mobile system for non-linear access to time-based data. In: Costabile MF (ed.) *Proceedings of Advanced Visual Interfaces AVI’04*. ACM Press, pp 454–457
- Luz S, Sheehan S (2014) A graph based abstraction of textual concordances and two renderings for their interactive visualisation. In: *Proceedings of the international working conference on Advanced Visual Interfaces, AVI ’14*. ACM, New York, pp 293–296
- Lyding V, Nicolas L, Stemle E (2014) Interhist—an interactive visual interface for corpus exploration. In: Calzolari N, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland. European Language Resources Association (ELRA)
- Mackinlay J (1986) Automating the design of graphical presentations of relational information. *ACM Trans Graph* 5(2):110–141
- Marai GE (2018) Activity-centered domain characterization for problem-driven scientific visualization. *IEEE Trans Vis Comput Graph* 24(1):913–922
- McEnery T, Wilson A (2001) *Corpus linguistics: an introduction*. Edinburgh University Press Series. Edinburgh University Press
- Miller A (2018) Text mining digital humanities projects: Assessing content analysis capabilities of Voyant Tools. *J Web Librariansh* 12(3):169–197
- Modnlp (2020) ModNLP software repository. <http://modnlp.sf.net>. Accessed Feb 2020
- Moretti F (2005) Graphs, maps, trees: abstract models for a literary history. Verso
- Munzner T (2009) A nested model for visualization design and validation. *IEEE Trans Vis Comput Graph* 15(6):921–928
- Newman W, Lamming M (1995) *Interactive system design*. Addison-Wesley
- Oelke D, Kokkinakis D, Keim DA (2013) Fingerprint matrices: uncovering the dynamics of social networks in prose literature. *Comput Graph Forum* 32(3 part 4):371–380

ARTICLE

PALGRAVE COMMUNICATIONS | <https://doi.org/10.1057/s41599-020-0423-6>

- Olohan M (2002) Corpus linguistics and translation studies: interaction and reaction. *Linguist Antwerp* 2002(01):419–429
- Paley W (2002) Textarc: showing word frequency and distribution in text. In: Wong PC, Andrews K (eds) *Proceedings of IEEE symposium on information visualization. Poster compendium*. IEEE CS Press
- Pecina P (2010) Lexical association measures and collocation extraction. *Language Resour Eval* 44(1–2):137–158
- Rabadán R, Labrador B, Ramón N (2009) Corpus-based contrastive analysis and translation universals: a tool for translation quality assessment english and spanish *Babel* 55(4):303–328
- Schmied J (1993) Qualitative and quantitative research approaches to English relative constructions. In: Souter C, Atwell E (eds) *Proceedings of the International Computer Archive of Modern English, conference*. pp 85–96
- Scott M (2008) *Wordsmith tools version 5. Lexical Analysis Software*, Liverpool, p 122
- Scott M et al. (2001) Comparing corpora and identifying key words, collocations, and frequency distributions through the wordsmith tools suite of computer programs. In: Ghadessy M, Henry A, Roseberry RL (eds) *Small corpus studies and ELT*. John Benjamins Publishing Company, pp 47–67
- Sheehan S, Luz S (2019) Text visualisation for the support of lexicography-based scholarly work. In: *Proceedings of the eLex 2019 conference on electronic lexicography in the 21st century*, Sintra, Portugal. pp 694–725
- Sheehan S, Masoodian M, Luz S (2018) COMFRE: a visualization for comparing word frequencies in linguistic tasks. In: Catarci T, Leotta F, Marrella A, Mecella M (eds) *Proceedings of Advanced Visual Interfaces AVI'18. Association for Computing Machinery (ACM)*, pp 36–40
- Shneiderman B (1996) The eyes have it: a task by data type taxonomy for information visualizations. In: Green TRG (ed.) *VL '96: Proceedings of the 1996 IEEE symposium on visual languages*. IEEE Computer Society, Washington, pp 336–343
- Sinclair J (1991) *Corpus, concordance, collocation*. Oxford University Press
- Sinclair J (2003) *Reading concordances: an introduction*. Pearson/Longman
- Svartvik J (2011) *Directions in corpus linguistics: proceedings of nobel symposium 82*, vol. 65, Stockholm, 4–8 August 1991. Walter de Gruyter
- Tufte ER (1990) *Envisioning information*. Graphics Press, Cheshire
- van Ham F, Wattenberg M, Viegas FB (2009) Mapping text with phrase nets. *IEEE Trans Vis Comput Graph* 15(6):1169–1176
- Venuti L (1995) *The translator's invisibility: a history of translation*. Routledge
- Viégas F, Wattenberg M (2008) Tag clouds and the case for vernacular visualization. *Interactions* 15(4):49–52
- Voyant (2020) *Voyant tools*. <https://voyant-tools.org/>. Last accessed Feb 2020
- Wattenberg M, Viégas FB (2008) The word tree, an interactive visual concordance. *IEEE Trans Vis Comput Graph* 14(6):1221–1228
- Zanettin F (2001) Swimming in words: corpora, translation, and language learning. In: Aston G (ed.) *Learning with corpora*. Athelstan, p 177
- Zanettin F (2013) Corpus methods for descriptive translation studies. *Procedia - Soc Behav Sci* 95:20–32

Acknowledgements

This research was supported by the Arts and Humanities Research Council, UK (Grant number: AH/M010007/1), and by the European Union's Horizon 2020 research and innovation programme under Grant agreement No. 825153, project EMBEDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflect only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.L.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

Appendix C: Text Visualization for the Support of Lexicography-Based Scholarly Work

Proceedings of eLex 2019

Text Visualization for the Support of Lexicography-Based Scholarly Work

Shane Sheehan, Saturnino Luz

Usher Institute of Population Health Sciences & Informatics,
The University of Edinburgh, UK
E-mail: Shane.Sheehan@ed.ac.uk, S.luz@ed.ac.uk

Abstract

We discuss three visualisation techniques for corpus analysis, Concordance Mosaic, Metafacet and ComFre, and explore the design rationale based on a characterization of the corpus linguistic domain. The Concordance Mosaic visualization is designed for the investigation of collocation patterns. It encodes word positions in a concordance list in a manner that emphasizes quantitative analysis of frequency or collocation statistics. Metafacet provides an interface for investigating concordance lists through the lens of meta-data. When combined with the Mosaic it provides a powerful technique for investigating collocations in the context of meta-data. ComFre can be used to compare word frequencies between two corpora of different size, it has potential use as a technique for identifying terms which are representative of the corpora under investigation. The domain characterization shows how the visualizations were designed with corpus linguistic methodologies at the core. It consists of a task analysis based on the methodology outlined in Sinclairs' *Reading Concordances: An Introduction*, and the analysis of methodology case studies from language scholars.

Keywords: visualization; concordance; frequency; meta-data; collocation

1. Introduction

Concordance analysis is a core activity of scholars in a number of humanities disciplines, including corpus linguistics, classical studies, and translation studies, to name a few. Through the advent of technology and the ever increasing availability of textual data this type of structured analysis of text has grown in importance (Sinclair, 1991; Bonelli, 2010).

In concordance analysis, every corpus occurrence of a keyword of interest is displayed along with its context. The context is an ordered list of words which precede and follow the keyword. The analyst then seeks to discover the linguistic properties of the keyword and the contextual patterns which predict them by observing the frequencies of occurrence, in the keyword's context, of words (collocations), word combinations, parts of speech (colligations) or the various other lexical classifications (Sinclair, 2003; Scott, 2010).

The most widely used tool in this kind of analysis is a form of tabular visualization known as keyword-in-context (KWIC). The creation of concordances through the

keyword in context indexing technique was first proposed by Hans Peter Luhn in the 1950's (Luhn & Division, 1959). KWIC displays, enhanced in interactive systems by features such as search, context sorting and statistical analysis, are widely used not only by academics and scholars, but also by professional translators and post-editors (Karamanis et al., 2011; Doherty et al., 2012).

While these KWIC interfaces provide support for exploring the linear structure of the concordance, word frequency and other statistics rarely form any part of the visualization. This statistical information is essential to the work of the text analyst. However, in the presence of large corpora, it is difficult to explore statistical regularities armed solely with the KWIC display. External statistical tools are often used to complement the concordance. We argue that integration of this analysis step into the concordance visualization fits in well with the task structure of corpus linguists, and will be of great benefit to the text analyst.

There have been calls for the creation of more advanced concordance analysis tools (Rockwell, 2003), and advancements such as Sketch Engine have provide new analytic paths (Kilgarriff et al., 2014). However, the adoption of visual analysis tools for concordance analysis is very limited. That does not mean that visual representations of the concordance do not exist, it is simply that they have not been adopted by analysts or integrated into analysis tools.

It has been suggested that the publication of more domain characterization papers for visualization would be beneficial for tool adoption (Munzner, 2009). It is at this level of design that relevant problems are identified, and creating visual solutions to problems that are not relevant to domain experts is wasted effort. Publication of domain characterization should also encourage wider conversation and help identify and characterize overlooked areas of investigation.

In this paper we outline the functionality of three corpus analysis tools, Concordance Mosaic, Metafacet and ComFre. Concordance Mosaic displays positional collocation statistics for any corpus word or regular expression. Interactive restructuring of a concordance browser is enabled through the interface. This restructuring combined with colour highlighting of the concordance lines creates a powerful technique for investigating significant collocation patterns.

The MetaFacet visualization enables exploration of corpora through the lens of meta-data. Keyword frequency can be investigated across any combination of meta-data attributes associated with corpus source files. The concordance browser and Mosaic can be interactively filtered by these attribute combinations, allowing investigation and comparison of lexical information across combinations such as date, author and topic.

ComFre is a tool for corpus frequency comparison, which provides a method of comparing corpora of different size in a visual and statistically valid manner.

These visualizations were designed in close collaboration with language scholars with an emphasis on translation studies. The design rationale is rooted in a domain characterization which encompasses a literature-based task analysis and ethnographic studies of methodology. Relevant portions of this domain characterization are presented following the visualization descriptions.

2. Modnlp plugins

The visualization tools are developed as plugins for the open source concordance browser included in the Modnlp toolkit. Significant contributions were also made to the core Modnlp project to better integrate the plugins and enable interactions with the concordance list. Modnlp provides a modular architecture and tools for natural language processing, it comes with an indexer, feature rich concordance browser and server implementation (Luz, 2011, 2000). Previous versions of the Modnlp software have been used by the European Parliamentary Comparable and Parallel Corpora project¹ (ECPC) and by the Translational English Corpus² (TEC). The toolkit is currently being developed as part of the Genealogies of Knowledge project³ (GoK) and the plugins are fully integrated into the GoK corpus browser.

The goal when developing these plugins is to improve the efficiency and capability of corpus linguistic methodologies and tools. Here we present the visualization plugins from a purely functional perspective to provide an overview of the capabilities and context for later discussion of the relevance to lexicography and corpus linguistics.

The English GoK corpus is used to exemplify the usage of the visualizations. This corpus is quite varied, it includes translations and re-translations of texts from antiquity as well as modern internet blogs and magazine articles. The corpus is designed to enable researchers to trace the trajectory of key concepts as they enter different cultural and temporal spaces, predominantly but not exclusively through the mediation of various forms of translation. The corpus is specialized and the examples used may not exhibit general lexical properties due to the issues of representativeness in relation to frequency (Summers, 1996).

In the discussion of the visualization functionality we do not try to analyse or interpret the linguistic properties of the words or corpus. Any analysis choice or comments on linguistic properties are to help clarify the examples and should not be viewed as an attempt to perform corpus analysis.

¹ <http://www.ecpc.uji.es/>

² <http://genealogiesofknowledge.net/translational-english-corpus-tec/>

³ <http://genealogiesofknowledge.net/>

2.1 Concordance Mosaic

The first visualization designed was the Concordance Mosaic. This visualization has the concept of keyword in context at its core. The visualization is designed to display word statistics per position extracted from a concordance list. The underlying graph based abstraction of the concordance list and an early prototype were presented in an earlier work (Luz & Sheehan, 2014).

Using the visual metaphor of the KWIC, Mosaic represents positions relative to the keyword as ordered columns of tiles. The mosaic is created using a space-filling approach introduced by Luz and Masoodian (2007), where each tile represents a word at a position relative to the keyword, and the height of each tile is proportional to the word statistic at that position. In its simplest form each tile represents the frequency of a word at a position relative to the keyword. In Figure 1 the Mosaic of the keyword “hazard” is presented along with the concordance list for the 335 occurrences in the corpus. The Mosaic is set to display column frequencies. Due to the strong visual metaphor of KWIC it should be clear the word “to” is the most frequent word immediately to the left of the keyword (K-1) and also at positions K-2 and K-3. Hovering over any tile will display a tool-tip with the word count and frequency at the position, this relieves the need for manually counting or performing additional searches to retrieve position based word frequencies.

Words with high corpus frequency tend to dominate the positional frequency distributions for most keywords. The second view Mosaic affords is a stop-word filtered view of column frequency. The columns are filtered using a threshold based on corpus frequency. In Figure 2, the stop-words are removed and column heights are no longer uniform. The reduction in a column’s height represents the density of stop-word frequency at that position. At K-1 we notice stop-words were the most frequent for any position. At K-1 the next most frequent word after stop-words is “moral”. Tile heights and thus frequency are comparable across positions, from the Mosaic we can see that “moral” at position K-1 and “run” at position K-2 have similar positional frequencies.

The mosaic and concordance browser have been presented together but we have not yet commented on the interaction. The data is linked to both interfaces, and interactions with the mosaic can be reflected on the concordance list. In Figure 2 the tile for the word “run” at K-2 has been left clicked with the mouse. This interaction colours white any position word tiles on the Mosaic that are found in concordance lines, including “run” at K-2.

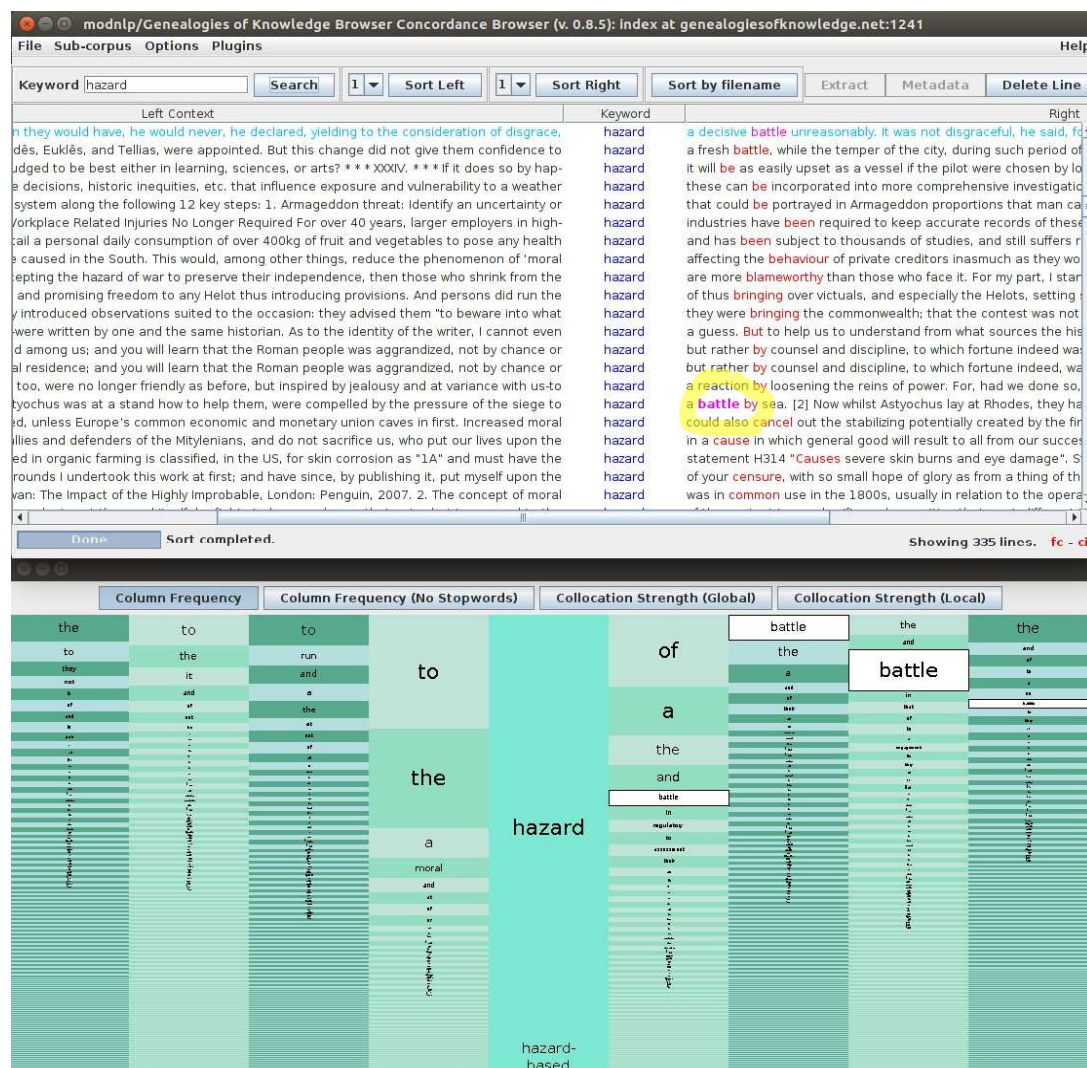


Figure 1: Concordance Mosaic for keyword “hazard”. Right click selection of “battle” at position K+3.



Figure 2: Concordance Mosaic for keyword “hazard”, stop words have been removed. Left click selection of “run” at position K-2.

Looking at the Mosaic we see that at least one concordance line with “run” at position K-2 also contains “battle” at K+2. The concordance list has been sorted at the selected position and scrolled automatically to the selected word. For emphasis the sorted position words are coloured red and the selected word coloured pink. The horizontal concordance lines for the selected word are coloured blue for easy identification. In addition, any occurrences of the selected word at other positions are also highlighted in pink, and as you investigate the entire list it is possible to get a sense of global patterns which may not be restricted to the selected position. In Figure 3 the selection of the word “to” at K-1 and a sample of its many occurrences at other positions are visible.

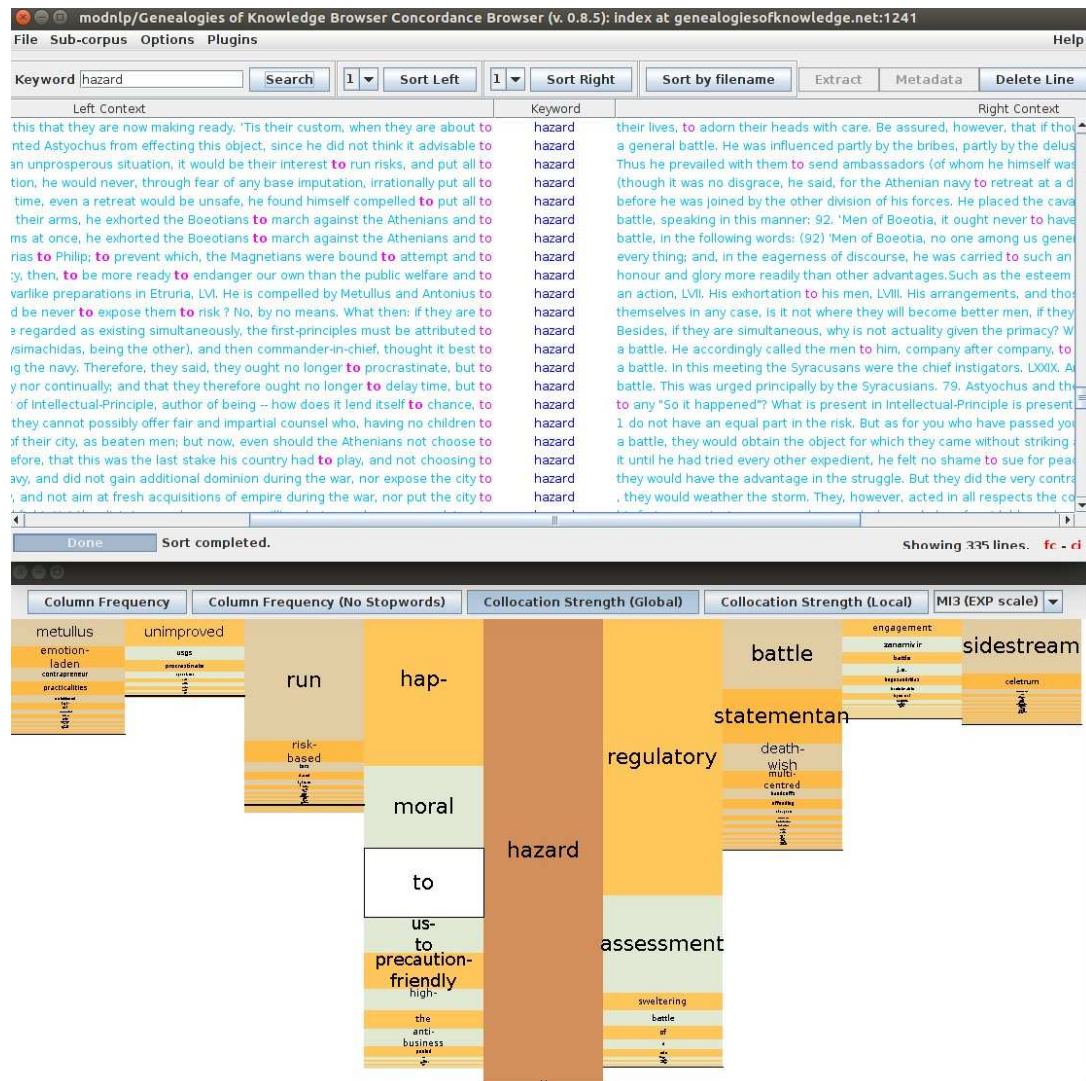


Figure 3: Concordance Mosaic for keyword “hazard”. Global view of MI3 is selected. Right click selection of “to” at position K-1.

The second click is activated by right clicking on a Mosaic tile. This interaction has the same effect on the concordance list as the left click interaction but differs in its change to the Mosaic. Right clicking on a Mosaic tile highlights other occurrences of the word at all positions in the mosaic. This is useful for getting a better sense of the frequency distribution of a word across all positions in a concordance list. In Figure 1, “battle” at K+3 is selected. Tiles representing “battle” at positions K+1 K+2 and K+4 are coloured white for easy identification. In the concordance list we can see one of these additional occurrences of “battle” at K+2 highlighted in pink.

Positional word frequency is a fundamental property of the concordance list, but other quantitative measures are used extensively to reason about collocations. Statistics such as Mutual Information (MI), Cubic Mutual Information (MI3) and Z-Score are often used to investigate collocation statistics in a window surrounding a keyword (Manning

& Schütze, 1999). This windowed approach most often groups word positions together and presents the results as a list. However we wish to preserve the positional aspect of these statistics and present them as a Mosaic. Figure 3 shows the *global collocation strength* view of Mosaic. Global in this setting means the tiles can be compared across positions and have not been scaled to fill the space. This contrasts with Figure 4 where the *local* view of collocation strength makes each column full height, and this allows easier investigation of each position but removes the ability to compare tiles across positions.

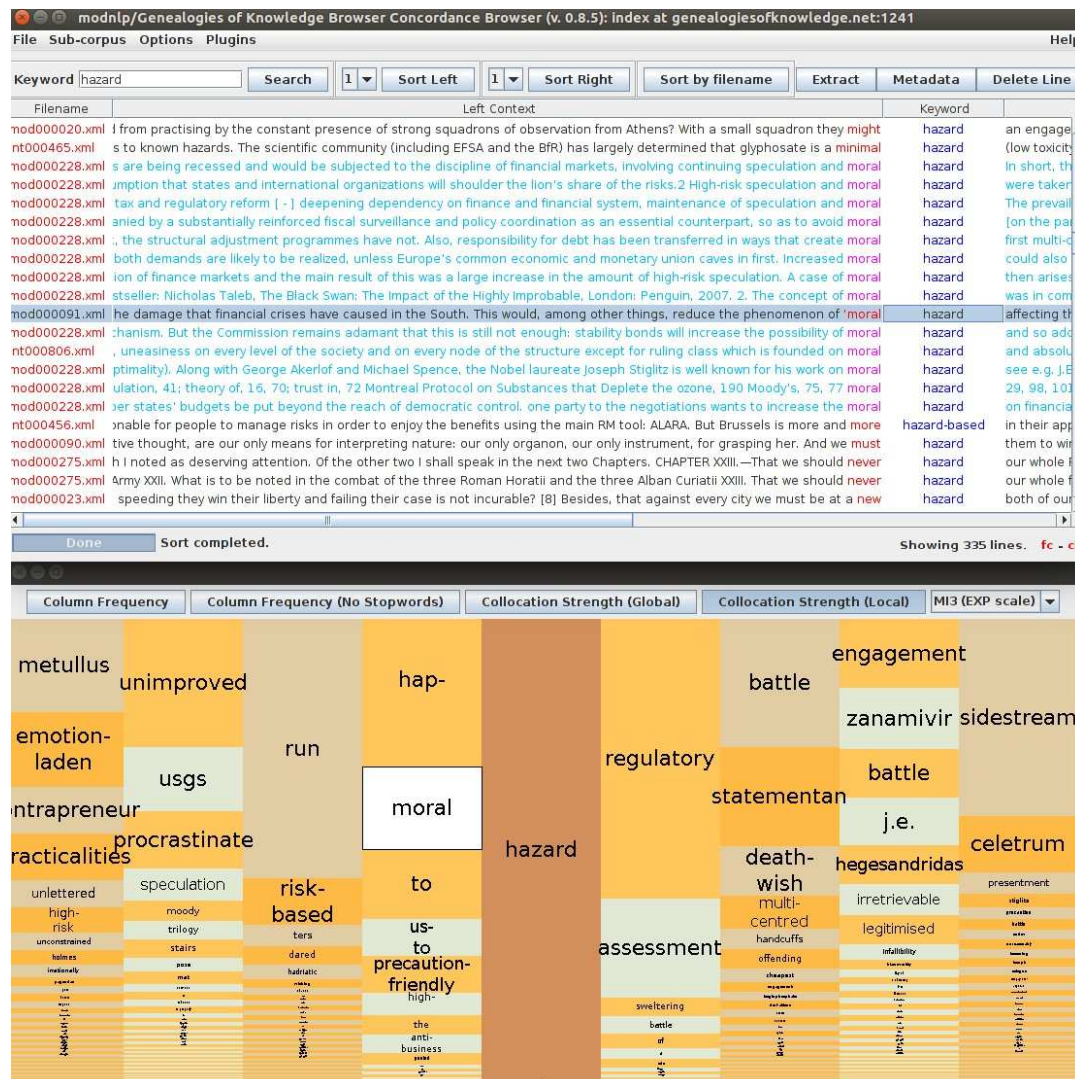


Figure 4: Concordance Mosaic for keyword “hazard” Local view of MI3 is selected. Right click selection of “moral” at position K-1. Concordance list scrolled horizontally to reveal filenames.

In the *Global* view shown in Figure 3 the column heights give an indication of the word positions relative to the keyword where the statistical association is highest. Each individual tile’s height is proportional to the value of the statistic calculated for that

word at that position. In this example MI3 is selected as the statistic under investigation. The strongest association based on MI3 is the word “regulatory” at K+1. It may be worth noting that the stop-word “to” is shown to have a strong association at K-1.

If we investigate the concordance lines of the tile “moral” at K-1 (since it has both high frequency and MI3 score) we find that all but two of its 14 occurrences originate from the same file, see Figure 4.

2.2 Metafacet

The Modnlp concordance browser presents the file-names along with concordance lines. An interaction is available in the browser to view meta-data about each file and section on a line by line basis. However, this is a time consuming and challenging process for the corpus analyst if the meta-data of a large number of lines need to be investigated. The Metafacet plugin is a proposed solution to this issue and provides interactive filtering of the concordance list and the Mosaic using all available meta-data facets.

The Metafacet interface is quite simple, and uses a horizontal bar chart to display concordance line frequency per meta-data attribute. An attribute is a possible value that a meta-data facet can take. As an example “Plato” is an Attribute of the Facet “author”. A drop-down list is used to choose which facet is displayed and the bars are sortable by frequency or lexicographical order and the window can be filtered using a sliding scale to view a smaller portion of the attributes. This conforms to the common visualization design practice of first presenting an overview, and then more detail on demand (Shneiderman, 1996).

In Figures 5 and 6 the Metafacet interface for the concordance of “hazard” is shown for the facet “author” sorted by frequency. Figure 6 shows a window of this data focusing on the nine most frequent attributes of this facet in the concordance list. The hover interaction is shown for “Thucydides”, who is the most frequent author of the keyword “hazard” in the GoK corpus, with a total of 94 concordance lines out of a list of 335.

Metafacet when used alone provides an interface to quickly explore keyword distribution across meta-data attributes. By interactively combining it with the concordance list and Mosaic we can navigate the corpus in a new way, viewing the concordance as attributed sets of collocations that can be interactively explored. In Figure 7 the stop-word Mosaic shown in Figure 2 is filtered to remove any concordance lines with the attribute “book” from the “format” facet. Books account for the majority of the concordance lines, and removing them from the concordance significantly changes the collocation structure of the Mosaic. During interactive filtering the current selection can be kept by pressing the “Update Bars” button, and this will refresh the Metafacet window with filtered concordance.

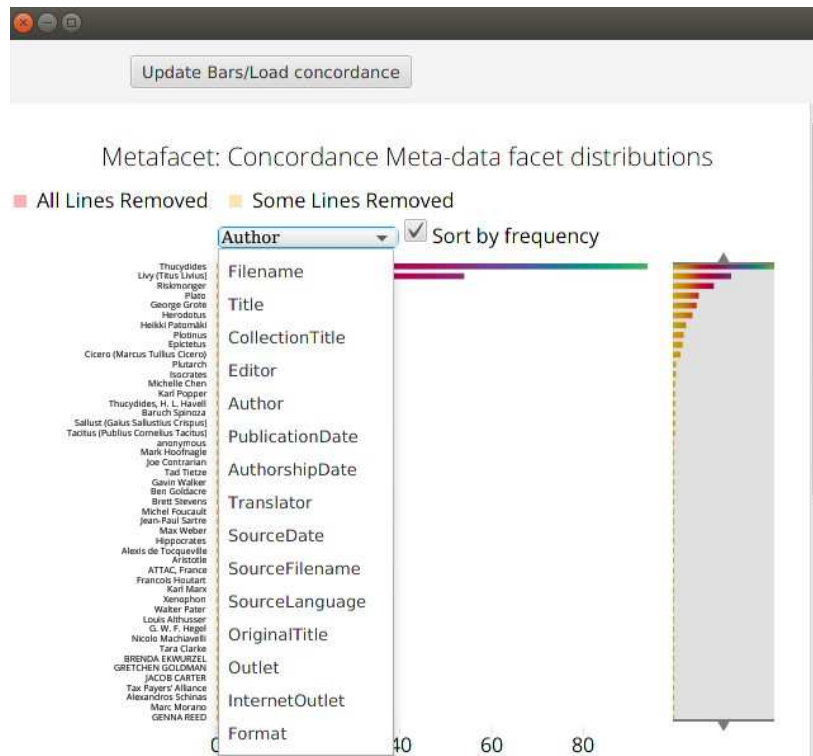


Figure 5: Metafacet interface showing all available meta-data facets. Fully zoomed out but obscured view of all authors in the concordance of “hazard”.

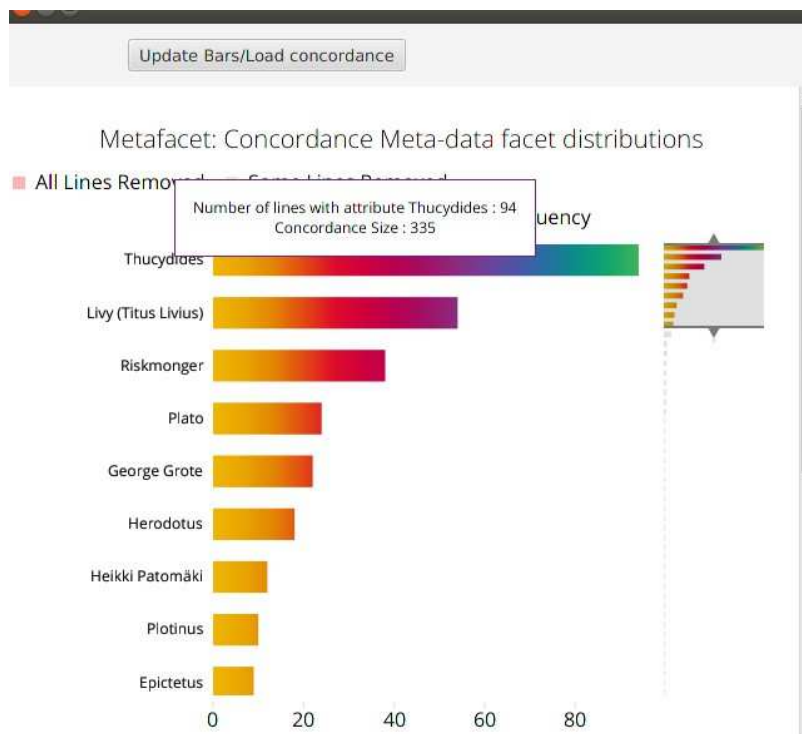


Figure 6: Metafacet zoomed to most frequent authors. Hover interaction displaying attribute name, associated concordance lines and total concordance lines for “hazard”.



Figure 7: Left click interaction filtering out any lines from the concordance associated with the attribute Format= “book”. Both Concordance Mosaic and List respect the click interaction.

Left clicking a bar removes an attribute from the concordance list, right clicking removes everything but the clicked attribute. Once an attribute or multiple attributes have been selected it is possible to switch to another facet to explore further. In Figure 8 the facet “author” is displayed after books have been removed. We can see from the red bars that the most frequent author was only found in books. The second and fourth most frequent authors are coloured yellow, this indicates that some of the lines associated with these authors have been removed but others have not. To view how much these yellow bars have been reduced the “Update Bars” button must be pressed to generate a new Metafacet for the filtered concordance. It is possible on this author facet window to add attributes back into the list by clicking on the red or yellow bars, and this would generate a filtered list where all books except those of the selected authors have been removed.



Figure 8: Viewing frequent authors after filtering out attribute Format = “book”. Partially removed attributes coloured yellow, fully removed attributes coloured red.

The combination of facets and attributes which can generate a single filtered list is limited only by the attribute crossover of the concordance lines. Finally the only author not colouring a block red or yellow in the nine most frequent is “Riskmonger”, who does not have any concordance lines associated with the attribute “book”. We stop the analysis here, but further exploration could be done to investigate the concordance lists and Mosaics for facets such as authorship/source dates and outlets. We would find that “Riskmonger” is a modern internet author who is responsible for the collocation patterns of “hazard” + “regulatory” and “assessment” at position K+1.

2.3 ComFre

The ComFre visualization is a corpus comparison tool where frequency lists can be compared visually in a statistically valid manner. The functionality of the tool has been detailed elsewhere (Sheehan et al., 2018), it has since been modified to operate as a plugin for Modnlp and is briefly presented here.

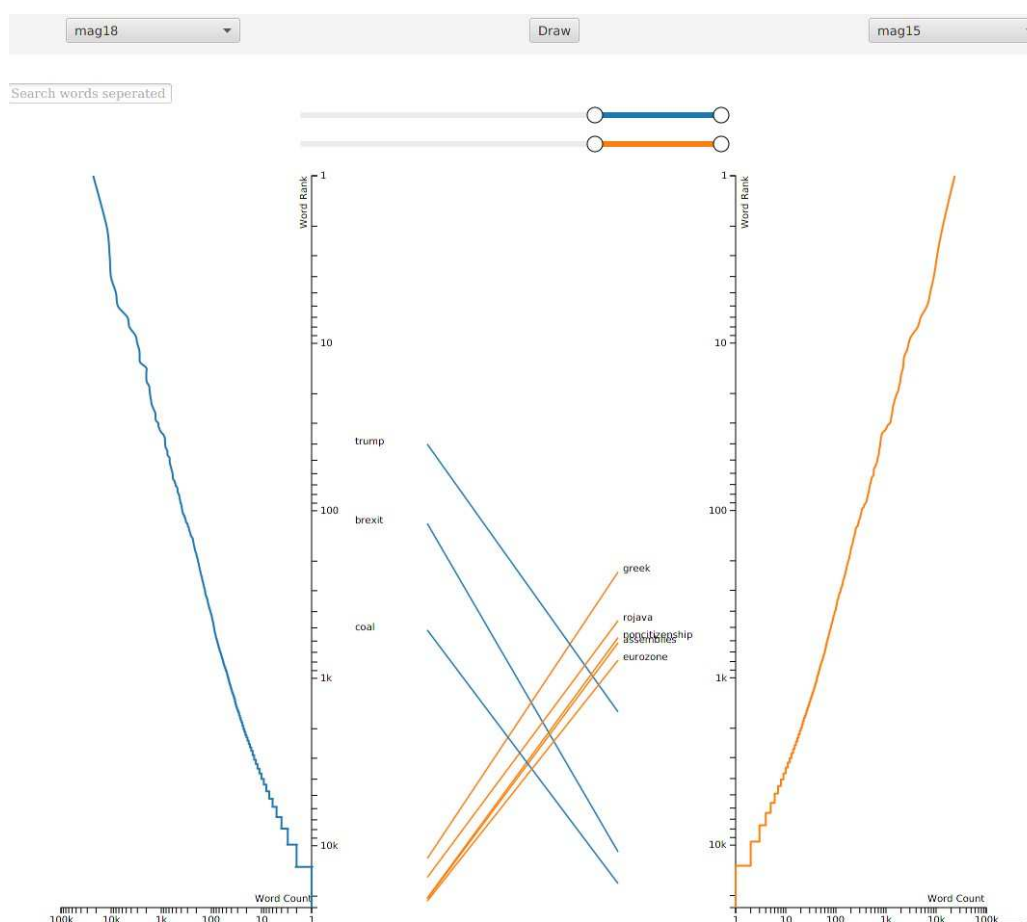


Figure 9: ComFre visualization comparing the words with the largest change in distribution rank between magazine articles from the GoK corpus authored in 2018 and 2015.

The Modnlp software has a sub-corpus selection interface which can be used to save the named sub-corpora for later reuse. ComFre makes these named sub-corpora available for comparison in dropdown lists. In Figure 9 “mag18” and “mag15” are selected for comparison, these sub-corpora are magazine articles from the GoK corpus which were authored in 2018 and 2015, respectively.

In ComFre both axis are log scaled, which should yield a linear frequency diagram if the word frequencies follow a Zipfian distribution. Scaling both ranked lists to the same height and comparing a word’s position in the distributions lets us compare sub-corpora

of vastly different size.

In Figure 9 the majority of the words have been filtered out to reveal the words with the greatest frequency changes between the two corpora. We can see that “Trump”, “Brexit” and “coal” were used much more often in the 2018 corpus, while words such as “Greek” and “eurozone” had much higher usage in 2015.

3. Domain characterization summary

This section explores the domain of corpus linguistics to identify problems and methods which will benefit from visualization. Visualizations which try to address the needs of corpus linguists are much more likely to be effective if those needs are well understood. The inclusion of domain experts in this visualization design stage is very beneficial, however just talking to users is typically not sufficient to achieve a full and accurate domain characterization. Expert users are extremely important when defining the high level goals and tasks of the domain and with ranking the importance of tasks. The characterization can be made more detailed by using methods such as examination of domain literature, contextual studies (Sedlmair et al., 2012) and needs assessments (Marai, 2018).

By performing a domain characterization, as outlined in the nested model (Munzner, 2009), the methodologies used to achieve the identified goals can be systematically investigated. The aim is to extract the low level tasks which are performed in the process of working towards the higher level goals. This analysis can be arranged as a hierarchy of goals, tasks and low level actions. The hierarchy can then be used to gain insight into the challenges faced by corpus linguists and how they have been previously addressed.

At its core domain characterization for visualization design is about identifying real problems which are relevant to the domain under investigation. This process is fluid and iterative, a level of domain understanding must be reached before work can begin on a visualization, but the design process should be reviewed as opportunities to refine the problems and domain characterization emerge.

The analysis presented here is not a full detailing of our characterization efforts. Rather, it is a presentation of some of the clearer insights and how they relate to the design choices which can be observed in the created visualization tools.

3.1 Literature-based domain analysis

Consultation and collaboration with the language scholars of the GoK project who interrogate corpora as an essential part of their analytical work lead to the natural discussion of visual tools to support analysis.

These collaborations revealed how integral the KWIC-based concordance display is to

the work of the text analyst. These visual representations provide an essential view of the context in which the keyword occurs. However, examining the relative frequencies of the words which surround the keyword is also a commonly performed task using these tools, for which it would appear these tools are not well suited. In practice, the analyst usually complements the textual information provided by the KWIC display with lists of words sorted by frequency of occurrence in the sub-corpus under examination, as well as other statistics. Different processes and sub-tasks mediate the analysis as a whole.

To study this type of concordance analysis in a practical context we turned to a reference work entitled *Reading Concordances: An Introduction* (Sinclair, 2003). This book is intended as a tutorial on how to look for certain linguistic properties of a keyword (such as word sense, phrasal usage, part of speech and many others) using a KWIC concordance list. The reader is invited to perform eighteen tasks which introduce the key practical actions and usage of linguistic knowledge required to make decisions about the properties of a word or collocation. For each of these tasks we performed a hierarchical task analysis (Annett, 2003) by combining or splitting the steps into a series of actions and sub-actions.

Each of the eighteen tasks was analysed and tagged to assist with the classifying and counting of the actions and sub-actions. Before explaining the exact meaning of the tags, an example of the tagging procedure for task 4 is given. This tagging procedure can allow a visualisation researcher with limited knowledge in the problem domain to extract meaningful actions.

Task 4 is concerned with identifying literal and metaphorical usage phrases. The preamble to the task provides some linguistic insight explaining that “some idiomatic phrases in English are recognizable because they contain a word which is not found anywhere else, like *at loggerheads*”. They may also be recognizable because the literal meaning is absurd. But others are more subtle and don’t have the aforementioned identifying marks. As an example the phrase *he got cold feet* seems to be a literal way of saying that his feet are cold. How do we as readers know when it means he is cowardly? The task studies the example of the phrase “free hand”. A concordance of 30 lines is provided and a set of twelve directions in how to analyse the concordance are given to the reader. An answer key is also provided which expands on the analysis and the insights that can be gained.

The first direction tells the reader to look at the position directly to the left of the phrases which have been sorted alphabetically “and list them in order of frequency. Can you associate any of the SINGLETONS with any of those that recur?” (Sinclair, 2003: 21) We tag this action with the *frequency* tag, *word position* tag, *group* tag and *expert decision* tag. The key gives a breakdown of the words at the position and notes that “her, your” are in the same word class as “his” and that “completely, fairly, totally” are in the same word class as “relatively”.

Step two asks the reader to

“Look again at the five lines where N—1 is an adverb of degree. What is the word at N—2? Then consider the two lines where N—1 is one. What is the word at N—2? Can you associate these seven lines with the two big groups of a and his . . . ?”

The positional notation N—2 means the set of words two positions to the left of the keyword. The same tags are applied to this action as word position, exact frequency counts and linguist knowledge are used. The answer key states

“Where N - 1 is an adverb of degree, N—2 is a; so these five lines join the group of the indefinite article. Where N—1 is the word one, in no. 25 N - 2 is her and so this line joins those with possessive adjectives. The other one, no. 24, has only at N - 2 , which is unlike all the other lines in this sample, so we will fit it in later on.”

Step three starts by explicating that in the previous step 28 of the 30 lines were extracted and divided into two groups based on “choice of determiner in front of the noun hand” the reader is then told “here the difference is not just the type of determiner; consider the meaning of free hand in the two types of line and comment on the distinction in meaning.” This task is tagged with *Similar Meaning*, *expert decision* and *read context*. For this examples the meanings of the keyword must be analysed by reading the contexts and using linguist knowledge to compare the meanings The answer key explains that when a possessive adjective is the determiner the word “free” means “available” and the word “hand” is a part of the human body. When the determiner is a the phrase “a free hand” it means “an unrestricted opportunity”.

Skipping forward to step seven the reader is narrowing in on the linguistic patterns which are used to determine literal or metaphorical usage of the phrase “free hand”. The reader is asked to group concordance lines according to whether the verb is active or passive and to examine if this accounts for the use of the word “given” exclusively before “a free hand”. Tags *group*, *read context* and *expert decision* all apply. Step 8 then combines all of the previous analysis to describe an algorithm for determining metaphorical or figurative usage of the phrase “free hand”. Many of the lines which have been discarded as not matching any patterns are not included in the construction of the algorithm.

Condition 1 of the algorithm is that there is a form of the word “give” or a word with similar meaning to the left of the phrase. If not is there an occurrence of the verb “have” or “get”, or one with a similar meaning and use?

Condition 2 is that the indefinite article precedes the core phrase, either directly or with only an adverb of degree in between.

If both conditions hold the phrase “free hand” means “to be set a task without restrictions on resources or methods to accomplish it”.

Steps nine to twelve examine all that had not previously examined in the concordance. The word frequencies and patterns to the right of the keyword are analysed and used to help account for the lines which could not be explained by the left context analysis.

This example should help clarify how the tags were assigned to the individual steps of the tasks. There was a significant amount of variation across the tasks, but the core actions could be described with a relatively small set of tags.

The actions and sub-actions generalize the descriptive analysis steps into operations which are common to many of the tasks. Taking an overview of our classifications of these actions we created the hierarchy shown in Figure 10.

At the first level of the hierarchy, the primary actions (second level) are split into quantitative and qualitative groups. Qualitative actions are classified on the criteria that a decision, in which it would be possible for experts to disagree, needs to be made to complete the action. These experts could be human users or algorithmic classification processes. Quantitative actions may form a part of a qualitative action, for example, frequent patterns must be identified before they can be classified as phrasal or non-phrasal usage (Sinclair, 1991).

The quantitative actions are those in which the steps involved in the action can be clearly stated, and, given the classifications have already been made, the results will be the same when performed by a reliable analyst. For example, for a concordance word frequencies at a specific word position can be accurately and repeatably determined. The quantitative actions often make use of the results of a qualitative action, such as estimating the frequency of words to the left of a meaning group where the group has to first be identified by expert decision.

The second level of the hierarchy contains the primary actions. These are the actions which most often describe the spirit of the instructions given in the eighteen tasks. Deeper into the hierarchy the sub-actions required to perform these primary actions are presented.

At the third level of the hierarchy the *area of analysis* is displayed, this is the level at which we perform the primary action. Looking first at the quantitative actions, we found that in three of the primary actions (filter, frequency and estimate frequency) a word’s position relative to the keyword is the area at which the actions are applied. A fourth quantitative action, frequent patterns, has an area of analysis, estimate frequency, which is one of the other primary actions. This means the action is performed on a collection of results from estimate frequency actions i.e. the analysis is performed on frequency estimations across word positions. It is worth noting that in four of the five quantitative tasks identified the word position or multiple word positions is the area at which the action is performed. The final action identified, *significant collocates*, uses

the results of statistical analysis of the keyword and its context from the corpus under investigation. This analysis is usually undertaken as a separate piece of analysis, which has its results reported as a list of frequent collocations with a keyword.

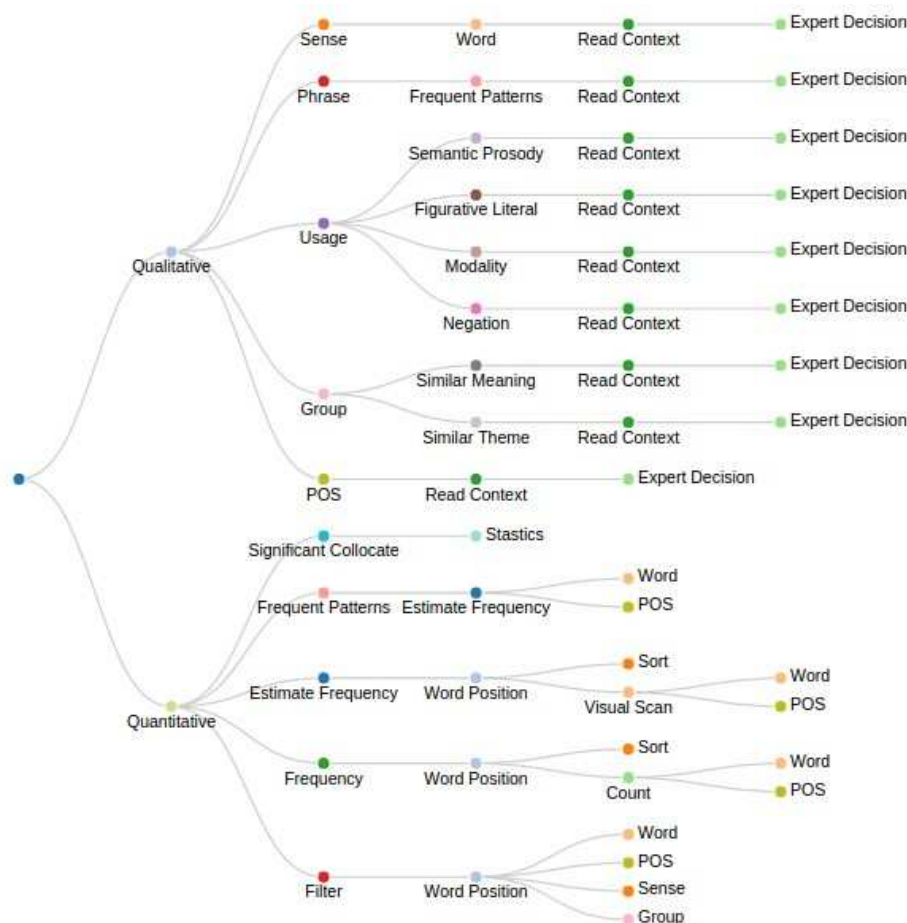


Figure 10: Hierarchical visualization of concordance-based corpus analysis actions.

Turning to the qualitative actions and, again, looking at the area of analysis at level three, we see that the analysis always occurs at the sentence level, which is implied by the read context action. This is in contrast with quantitative actions where positions are the most common area of analysis, and for qualitative actions it appears the horizontal structure of the KWIC list is emphasized while the qualitative actions make better use of the vertical alignment. Each of the actions requires an expert (or algorithm) who evaluates the context of individual occurrences of the keyword and makes a classification decision based on the semantic and syntactic content of the concordance line. This *Expert Decision* can often be the result of a combination of reading the individual contexts (the linear structure of the text) and performing some of the quantitative actions (positional statistics of the text). In essence, the *Expert Decision* action encapsulates the process of using the information extracted by other primary actions to answer questions about the keyword using linguistic knowledge.

Tag	No. of tasks in which an action appears	Total action appearances
expert decision	18	60
estimate frequency	16	34
read context	16	31
frequent patterns	15	21
frequency	14	18
word position	13	24
POS: Part of speech	11	23
filter	11	18
sense	10	19
group	7	9
significant collocate	5	7
usage	5	6
phrase	5	6

Table 1: Action counts from task analysis. Total numbers of actions found in the 18 tasks and numbers of the 18 tasks which feature the action.

While most of the tags represent actions, a few additional tags were chosen to help clarify and add information about the tasks and sub tasks. The tags *word*, *semantic prosody*, *Similar Meaning* and others are not themselves actions, but are useful in clarifying the objective or operation of the sub-actions. The part of speech (*POS*) tag is both a primary action tag and a clarifying tag. The POS primary action is to determine the part of speech of a word occurrence. The POS clarifying tag represents the use of part of speech information in another action. The purely clarifying tags are omitted from the analysis of tag frequency.

We recorded the distribution of the tags according to the number of tasks in which it appeared and the total number of actions which received the tag, as shown in Table 1. At a high level, this table tells us that both qualitative actions enabled by reading concordance lines and quantitative actions which require positional statistics are necessary for the style of concordance analysis outlined by Sinclair (2003).

3.1.1 Influence on visualization design

The structure that the task analysis and tag weightings add to the descriptive methodological steps was very useful for the early visualization design. The initial prototype of the Concordance Mosaic followed directly from this analysis. By focusing on frequent yet difficult aspects of the methodology we were able to create an interface which was likely to be of interest to corpus linguists. This gave us the opportunity to

engage with domain experts in the iterative development of tools and methodology starting with a useful prototype.

3.2 Methodological descriptions for GoK case studies

During the development of the visualization tools many interactions with GoK researchers occurred in situations such as progress meetings, design reviews and informal meetings. One set of interactions which made significant contributions to identifying relevant domain problems is presented here.

This takes the form of an initial presentation and follow up observation session with one GoK researcher. In the initial meeting a simplified methodology for a case study was described and a visualization which could be useful was suggested. The follow up observation session took place a number of months later after the Mosaic interface had been improved and made available to the researchers.

3.2.1 Methodology presentation

In the methodology discussion meeting a brief presentation outlining an example methodology and its challenges was given by a member of the GoK project to help with the initial definition of visualization goals for the project. The methodology was explained in the form of a case study. The case study made use of the portion of GoK English corpus which was available at the time. The task was defined as comparing the patterning around the keyword “*citizen**”. The * represents a regular expression search for continuations of the word citizen such as citizens and citizenship. The patterns identified were compared across two large sub-corpora.

- **Sub-corpus 1** A sub-corpus of modern English translations from Classical Greek (1850 onwards);
- **Sub-corpus 2** A sub-corpus of translated and non-translated texts written by contemporary authors, published between 1992 and the present day.

The method itself consisted of two techniques. The goal of the first technique is the identification of explicit definitions of “citizenship” contained within each sub-corpus. To find these definitions the researcher wants to compile a list of frequently used verbs and prepositions at position “keyword+1”. To achieve this the GoK corpus browser is used. Sub-corpus 1 was selected using the sub-corpus selection tool, the regular expression “citizen*” was searched and the concordance was sorted at position “keyword+1”. The researcher then spends time scrolling through the concordance and compiling a list of relevant frequent words at the position of interest, Figure 11 shows the concordance window sorted and scrolled to the preposition *as*. With this list in hand more accurate searches can be run such as:

- citizenship+“(is/as/was/defined/conceived/are/equals /considered/appears/means)”
- citizenship+“(has/should/must/will/may)”
- citizen+“(is/as)”
- citizens+“(are/as)”



Figure 11: Visualization proposed by GoK researcher

By reading the concordance lines generated by these new searches definitions can be extracted. Some examples of the definitions found are:

- Citizenship is a status bestowed on those who are full members of a community.
- As well as enjoying rights, citizens are required to undertake responsibilities such as paying taxes, and jury or military service.
- Citizenship should be based purely on residency
- US citizenship has represented a safe haven from oppressive regimes around the world

The second technique is the observation of patterns in the kinds of adjectives used to modify “citizenship”, as well as constructions such as “citizens+of+*”. The researcher explained that this technique is more difficult and time consuming using a concordance browser. To quote the researcher.

“Specifically, it is difficult to get a quick overview of such patterns using the concordancer given that the number of lines returned for my searches is quite large:

e.g. 4420 hits for “citizen*” in my sub-corpus of translations from Classical Greek.”

The researcher had some experience with linguistic visualization having used early versions of Mosaic and in the past had used word clouds, such as Wordle (Viegas et al., 2009), to present research results. There are some challenges to overcome to use word clouds for the methodology. The first which the researcher noticed is that stop-words dominate the frequency distributions of the word positions, so some technique has to be applied to get meaningful results. The suggested technique was to use a stop-word list to filter the visualization. The concordance would need to be processed to extract the words at particular positions for visualization, since the concordance is structured as a list of aligned text extracts. The result of the researchers reasoning was an interface for displaying positional word clouds with the option to exclude stop-words. The presentation included a mock-up of what a visualization to solve this problem would look like, as shown in Figure 12. The mock-up displays a word cloud for either a full concordance or a chosen word position, and has the option to remove stop-words. Looking at the mock-up in Figure 12 the words modifying citizen are presented in a manner that emphasizes frequency and provides an overview on a single screen of a position relative to the keyword.

At the end of the presentation the idea and its feasibility were discussed and some questions were asked to clarify the methodology. The notes taken were later discussed with the researcher and the following questions and answers were prepared.

- What is the domain in which the case study is situated?

“Translation and Reception studies. How have we received classic Greek texts? How has translation shaped this reception? The role of translation is often overlooked.”

- Is this methodology (excluding the proposed visualization) typical of the field?

“Translation Studies as a discipline tends to encourage close qualitative analysis of a small selection of examples chosen from specific texts to illustrate a particular argument.

Corpus analysis enables the translation scholar to identify and investigate with significantly greater ease differences between and patterns within translations, taking into account the full length of each work as a complete text.

Corpus analysis has been extensively used in translation studies before (e.g. within the TEC project and many others) but the field has tended to focus mainly on more micro-level linguistic concerns, rather than the socio-political implications of translators’ word-choices etc.”



Figure 12: Visualization proposed by GoK researcher.

- How did the idea for this example arise?

“GoK seeks to understand the constellation of concepts related to the body politic across time and space. Citizenship is a lexical item in that constellation. Comparing meaning, frequency and usage of related terms is an exploratory process used to discover obvious patterns.”

3.2.2 Methodology presentation: Design influence

The presentation helped confirm that the tasks and actions identified in the task analysis were relevant to at least one linguistics researcher. The early design of Mosaic did not take into account the need for removal of stop-words to make the Mosaic more usable. The researcher identified this flaw but did not notice the equivalence between a mosaic column and a word cloud. By removing the stop-words from the Mosaic you present the same information as a positional word cloud with a greater visual emphasis on word position and frequency. This was a very beneficial meeting, and led to the addition of this “No Stop-word” view of Concordance Mosaic.

3.2.3 Methodology observation: Case study of “the people”

After a significant amount of time follow-up observation sessions were organized to gain further insight into the methodologies of the researcher who gave the presentation. This took place after the development and release of the mature Concordance Mosaic, but prior to the development of Metafacet.

Prior to the observation session a spreadsheet was created with the headings filenames, date, translator, people, citizens, commons, Athenians, public. The meta-data information related to filename, date and translator were added to the table. The remaining headings are keywords which will be investigated as part of this study. The spreadsheet used in the study can be seen in Figure 13. Partitioning the frequencies by date, file or translator is equivalent for this sub-corpus, as each file has a unique author and date.

	A	B	C	D	E	F	G	H
1	Filename	Date	Translator	people	Citizen	commc	Atheni	public
2	mod000023.xml	1629	Hobbes	167				
3	mod000098.xml	1848	Dale	158				
4	mod000148.xml	1873	Wilkins	27				
5	mod000020.xml	1874	Crawley	145				
6	mod000019.xml	1881	Jowett	185				
7	mod000214.xml	1910	Havell	29				
8	mod000016.xml	1919	Smith	182				
9	mod000048.xml	1998	Lattimore	211				
10			Total	1112	551	151	8310	405

Figure 13: The spreadsheet which was used in the study of “the people” in translations of “Thucydides” from the GoK corpus.

The first steps of the study focused on the keyword frequencies in the entire sub-corpus.

- The sub-corpus of “Thucydides” was selected.
- The keyword “people” was searched and the total frequency in the corpus was recorded
- Regulator expressions for the other “citizens?”, “commons?”, “Athenians” and “public” were searched and the total frequency in the sub-corpus was recorded.

The researcher commented, after the keyword frequencies had been recorded, that the keyword “Athenians” is much more frequent than other keywords. This is unexpected and will need to be investigated.

The next step was to gather the keyword frequencies for individual files.

- Make a sub-corpus selection for each individual file. Record in the spreadsheet the number of lines returned for the keyword “people”.

The analysis now turns from keyword frequency to the identification of collocation patterns. Mosaic was used extensively to identify collocation patterns and frequency of occurrence. The steps observed were:

- Make a sub-corpus selection for the first file.
- Perform a search for the first keyword “people” in the concordance browser.
- Open the Mosaic visualization and remove stop-words.
- Examine word frequencies.
- Open a document for taking notes and record in it the most frequent collocations directly to the left of the keyword. The words “common and “Athenian” were recorded.
- Return to the sorted concordance list and check if any continuations (such as “Athenians”) are present.
- Record the counts for the frequent collocated words. (common 8, Athenian 6).
- Open the frequency mosaic with stop-words included.
- Record in notes “lots of hits for the+people (i.e. unmodified)”
- Similar analysis for second file.
- Frequent collocates directly to left of “people” (common 34, Athenian 5).
- Record “A few more different adjectives modifying this noun:entire, experienced, free, dynamic, adventurous.”
- Similarly for the third file the noted collocates were (Athenian 13, whole 13, common 5).

The recording was ended and the researcher explained how the analysis would progress. The collocation pattern method is repeated and would continue in the same manner for each file and keyword. The next stage of the analysis would be to analyse the frequency patterns using the table. Possibly making bar charts in a spreadsheet application. Temporal patterns are expected. Identified patterns will be investigated using qualitative analysis, which involves reading the concordance lines related to the identified patterns. Understanding the meaning of the concept of “the people” at

different times is the goal.

This analysis is performed in the context of the knowledge the researcher has about the corpus and texts. She states that it is interesting that there are

“No translations 1919-1998, during period of huge cultural change in Britain. Possible reasons for this include Suffrage, war or technological revolution. The researcher explained that information about the authors and texts will influence the analysis. Some examples of information which is relevant are “the political leanings of the translators which is established relevant knowledge” and “certain texts are partial translations, abridged versions etc.”

Any differences identified, temporal or otherwise, must take into account translator style, politics and more.

Some questions were asked the researcher to elicit more information about the methodology

- How did you come up with this methodology?

“Playing around with the corpus tools, generating concordances for interesting keywords, trying to find patterns in the data.”

- How did you choose the keywords?

“Obvious keywords associated with the concept of “the people”. The idea for the study emerged through reading the literature on citizenship.”

- Would this methodology be useful for other researchers in the field?

“Other scholars using the GoK software to investigate the role of translation in the evolution of political and scientific discourse use similar methods. Other projects developing other corpora may also adopt some aspects of the methodology.”

- What are barriers to the adoption of your methodology?

“Not sure. Perhaps better documentation of the corpus software, detailing what it can and can’t do, with lots of example analysis. The publication of case-studies by members of the team will also help demonstrate the potential of the tools.”

- Mosaic was used in this analysis, is this typical when you investigate collocation patterns?

“Yes. Mosaic will be very useful for this case-study and any investigation of collocations, because it tells you in very quick and transparent way which

are the most common collocates in each word position for a given keyword.”

- You did not make use of collocation strength in your analysis, do you intend to?

“No. The collocation strength Mosaic is not immediately clear, and so (to be brutally honest) would tend to slow down analysis rather than speed it up.”

- Have you used this methodology for other studies?

“The collocation pattern aspect of this study is unique in my work. I have in previous studies studied keyword frequency in larger sub-corpora where there are multiple files for each author and date. I can show you an example for the concept of “Statesman”.”

3.2.4 Methodology observation: Case Study of “Statesmanship”

An unpublished paper on a case study of the concept of “Statesmanship” was supplied by the researcher and the major conclusions and analysis were described.

In the GoK corpus the term “statesman” was found to exist “almost exclusively (90%) in translations from Classical Greek”. This pattern was not observed for other similar keywords such as “governor”, “leader”, “ruler” and “citizen”, which are more evenly distributed across all language pairs. The analysis which arrived at this conclusion was a simple keyword frequency comparison across the translation facets of the corpus. This involved selecting each sub-corpus individually and recording the number of concordance lines for the keywords in each sub-corpus.

The frequency of the keyword “statesman” in the sub-corpus of Classical Greek translations was analysed. A spreadsheet with an entry for each of the 261 files in the sub-corpus was created and meta-data (the author, the title, the translator and the date) was entered for each file. This was done manually and was time consuming. The researcher explained that in this form “the information could easily be (re)sorted according to each of these meta-data facets and patterns more easily identified”. The number of concordance lines for each file was found by selecting a sub-corpus of a single file and searching for “statesman”. Performing this action for each of the 261 files was also time consuming. A sample of the completed spreadsheet can be seen in Figure 14.

By examining the spreadsheet and generating bar charts, such as Figure 15, the faceted distributions of “Statesman” can be understood. “statesman” seemed to be “bursty”, to use the author’s term, and to exhibit a temporal pattern.

The frequency of “statesman” in these corpora suggest most recent translations (1950-2012) of ancient Greek texts use “statesman” much less frequently. This is surprising because the corpus contains several recent re-translations

(published within the last seventy years) of classical texts such as Aristotle's Politics or Plato's Dialogues which in earlier English-language interpretations included the keyword "statesman" very prominently.

Filename	Author	Title	Translator	Date	Hits for statesm*
mod000023	Thucydides	History of the Peloponnesian War	Thomas Hobbes	1843	1
mod000149	Herodotus	Histories	Henry Cary	1847	0
mod000098	Thucydides	The history of the Peloponnesian war by Thucyd	Henry Dale	1848	0
mod000179	Plato	Apology	Henry Cary	1848	0
mod000180	Plato	Crito	Henry Cary	1848	0
mod000181	Plato	Gorgias	Henry Cary	1848	0
mod000182	Plato	Phaedo	Henry Cary	1848	0
mod000026	Hippocrates	Oath	Francis Adams	1849	0
mod000027	Hippocrates	Airs, Waters, Places	Francis Adams	1849	0
mod000035	Hippocrates	Law	Francis Adams	1849	0
mod000186	Plato	Republic	Henry Davis	1849	0
mod000178	Plato	Statesman	Georges Burges	1850	79
mod000212	George Grote	History of Greece Vol. 7		1851	2
mod000213	George Grote	History of Greece Vol. 8		1851	17
mod000211	George Grote	History of Greece Vol. 6		1851	19
mod000152	Plato	Republic	John Llewelyn Davies	1852	6
mod000177	Plato	Laws	Georges Burges	1852	17
mod000150	Thucydides	THE HISTORY OF THE PLAGUE OF ATHENS; Translat	Charles Collier	1857	1
mod000147	Herodotus	Histories	George Rawlinson	1858	0
mod000188	Plato	Gorgias	E. M. Cope	1864	32
mod000163	Plato	Apology	Benjamin Jowett	1871	0
mod000164	Plato	Crito	Benjamin Jowett	1871	0
mod000165	Plato	Phaedo	Benjamin Jowett	1871	0
mod000172	Plato	Theaetetus	Benjamin Jowett	1871	1
mod000169	Plato	Meno	Benjamin Jowett	1871	12
mod000170	Plato	Sophist	Benjamin Jowett	1871	19
mod000153	Plato	Republic	Benjamin Jowett	1871	33
mod000168	Plato	Laws	Benjamin Jowett	1871	39
mod000167	Plato	Gorgias	Benjamin Jowett	1871	41
mod000171	Plato	Statesman	Benjamin Jowett	1871	100
mod000148	Thucydides	Speeches from Thucydides	Henry Musgrave Wilkins	1873	8
mod000020	Thucydides	The History of the Peloponnesian War	Richard Crawley	1874	2
mod000252	G. W. F. Hegel	Hegel's Logic (Part One of Hegel's Encyclopaedia	William Wallace	1874	2

Figure 14: A sample from the spreadsheet used in the study of "statesman" in translations of Classical Greek from the GoK corpus. The full spreadsheet contains 261 lines of analysis.

Some clarifying questions were asked and answered:

- You mentioned the process of completing the spreadsheet was time consuming, how long did it take?

"Probably around 5-6 hours because of the amount of manual processing required. It would take a lot longer if I were to investigate more than one keyword."

- Where did the idea for this study and methodology come from?

"This was exploratory. I was not trying to establish anything in particular, only to understand whether the term "statesman" was used, how frequently (in comparison with other semantically related terms), and if any obvious patterns could be found from these initial quantitative analyses.

The terms “statesman” and “citizenship”, which I have investigated previously, are very closely related concepts, especially in classical Greek thought.”

- Were the visualization tools used in this case study?

“My focus on the use of a single keyword (“statesman”) and alternative word choices did not require and collocation pattern analysis. This is more typical of translation studies research. The corpus tools lend themselves particularly well to the analysis of collocations (this is one of their clear advantages), and this is why I want to push my research in this direction with my next case study.”

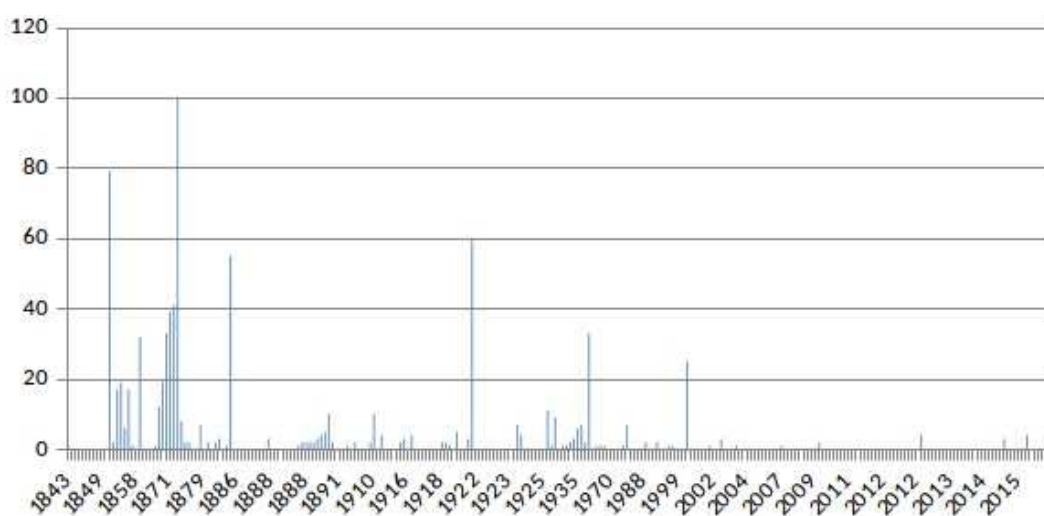


Figure 15: Bar chart examining temporal spread in translations of ancient Greek.

- Are there any areas of your methodology where you current or new visualization tools could be beneficial?

“Constructing the spreadsheets is time consuming. A tool which can help identify patterns in the dispersion of a concept according to different meta-data facets would be extremely helpful, at least for the kinds of research I intend to carry out as part of this project.”

3.2.5 Case study observation: Influence on visualization design

The most significant outcome of the two case studies was the emergence of the obvious need for a method to support the analysis of concordance lists through the lens of metadata. This observation session led to further discussion and needs assessment for a meta-data analysis tool which eventually became Metafacet.

Another problem identified was that in the version of Mosaic available to the researchers at that time only a single collocation statistic was available, and it was based on Mutual Information. The researcher did not know exactly what the scaling scheme for the collocation strength of Mosaic View was, and so could not accurately interpret or use it for publication. This led to the creation of optional scaling schemes based on well-known collocation metrics. More collocation measures are still being added to the tool.

4. Discussion and conclusions

We have presented three visualization techniques for corpus analysis. We hope that they can be adopted where appropriate by lexicographers and the wider corpus linguistic community. In addition, discussion of the tools and techniques by the community is welcomed.

We would be glad to hear any ideas, comments or criticisms of our ideas, understanding and designs. We believe the problems the tools address are general enough to have wide applicability in corpus linguistics, but we do not doubt that specific domains, such as lexicography, will have nuanced requirements that may need specialized interactions or entire redesigns to make them useful enough to be widely adopted.

The domain characterization detailed here can be another point of discussion, perhaps leading to more specialized future work on specific domain problems. We believe it is extremely important to provide a rationale for design decisions and to engage with domain experts when designing or modifying a tool or technique. Future work in this area will take the form of modifications which are identified during further domain exploration, and new visualization techniques where entire new problem areas are uncovered.

5. Acknowledgements

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflect only the author's view and the Commission is not responsible for any use that may be made of the information it contains. Pierre Albert has been funded through the INCA project. We thank the INCA project members in Ireland for granting us access to the trainee data.

6. References

- Annett, J. (2003). Hierarchical task analysis. *Handbook of cognitive task design*, 2, pp. 17–35.
- Bonelli, E. T. (2010). Theoretical overview of the evolution of corpus linguistics. *The*

- Routledge handbook of corpus linguistics*, p. 14.
- Doherty, G., Karamanis, N. & Luz, S. (2012). Collaboration in Translation: The Impact of Increased Reach on Cross-organisational Work. *Computer Supported Cooperative Work (CSCW)*, 21(6), pp. 525–554.
- Karamanis, N., Luz, S. & Doherty, G. (2011). Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation*, 25(1), pp. 35–52. URL <http://dx.doi.org/10.1007/s10590-011-9093-x>.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36. URL <http://dx.doi.org/10.1007/s40607-014-0009-9>.
- Luhn, H. & Division, I.B.M.C.A.S.D. (1959). *Keyword-in-context Index for Technical Literature (KWIC Index)*. ASDD Report. International Business Machines Corporation, Advanced Systems Division. URL <http://books.google.ie/books?id=Dk7pAAAAAAAJ>.
- Luz, S. (2000). A Software Toolkit for Sharing and Accessing Corpora Over the Internet. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhauer (eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation: LREC-2000*. pp. 1749–1754.
- Luz, S. (2011). Web-based corpus software. In A. Kruger, K. Wallmach & J. Munday (eds.) *Corpus-based Translation Studies – Research and Applications*, chapter 5. Continuum, pp. 124–149.
- Luz, S. & Masoodian, M. (2007). Visualisation of Parallel Data Streams with Temporal Mosaics. In E. Banissi et al. (eds.) *Procs. of the 11th International Conference on Information Visualisation*. Zurich: IEEE Computer Society, pp. 197–202.
- Luz, S. & Sheehan, S. (2014). A Graph Based Abstraction of Textual Concordances and Two Renderings for their Interactive Visualisation. In *Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '14*. New York, NY, USA: ACM, pp. 293–296.
- Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.
- Marai, G. E. (2018). Activity-Centered Domain Characterization for Problem-Driven Scientific Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), pp. 913–922.
- Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), pp. 921–928.
- Rockwell, G. (2003). What is Text Analysis, Really? *Literary and Linguistic Computing*, 18(2), pp. 209–219.
- Scott, M. (2010). What can corpus software do. *The Routledge handbook of corpus linguistics*, pp. 136–151.
- Sedlmair, M., Meyer, M. & Munzner, T. (2012). Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp. 2431–2440.

- Sheehan, S., Masoodian, M. & Luz, S. (2018). COMFRE: A Visualization for Comparing Word Frequencies in Linguistic Tasks. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces, AVI '18*. New York, NY, USA: ACM, pp. 42:1–42:5. URL <http://doi.acm.org/10.1145/3206505.3206547>.
- Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings the IEEE Symposium on Visual Languages*. pp. 336–343.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Describing English language. Oxford University Press.
- Sinclair, J. (2003). *Reading Concordances: An Introduction*. Longman Publishing Group. URL <http://books.google.ie/books?id=Ms9nQgAACAAJ>.
- Summers, D. (1996). Corpus lexicography—the importance of representativeness in relation to frequency. *Longman Language Review*, 3, pp. 6–9.
- Viegas, F. B., Wattenberg, M. & Feinberg, J. (2009). Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), pp. 1137–1144.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Appendix D: TeMoCo: A Visualization Tool for Temporal Analysis of Multi-Party Dialogues in Clinical Settings

2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)

TeMoCo: A Visualization tool for Temporal Analysis of Multi-Party Dialogues in Clinical Settings

Shane Sheehan*, Pierre Albert*, Masood Masoodian[†] and Saturnino Luz*

**Usher Institute, Medical School, The University of Edinburgh, UK*

Email: {Shane.Sheehan,Pierre.Albert,S.luz}@ed.ac.uk

[†]School of Arts, Design and Architecture, Aalto University, Finland

Email: masood.masoodian@aalto.fi

Abstract—We present a tool for visualization of transcripts of multi-party dialogues, with application to the analysis of communication in medical teamwork. The visualization is based on a “temporal mosaic” metaphor, which provides a temporal overview of dialogues and supports the tasks of transcript browsing and information access, by segmenting the dialogue and laying out the keywords of the different segments on interactive visual “tiles”. The tool has been tested on a corpus of transcribed dialogues among the members of a (simulated) critical care team. An analytical evaluation is presented which demonstrates the potential uses of the tool in an educational setting and highlights areas for improvements.

Keywords—Medical team communication; temporal visualization; temporal mosaics; speech visualization.

I. INTRODUCTION

Effective verbal communication is crucial to the success of clinical encounters, including clinician-patient consultations, multidisciplinary medical team meetings, accident and emergency contexts, among others. Analysis of communication in such contexts is important for care quality assessment, individual appraisal, assessment of interventions, as well as training and education. However, this kind of analysis tends to be very time-consuming, requiring substantial input from healthcare research experts. While frameworks such as the widely used Roter Interaction Analysis System (RIAS) [1] have helped standardize and guide such work, analyzing medical communication at scale remains a challenge.

While recent advances in speech and language processing technologies promise to facilitate the job of healthcare communication analysts [2], visual tools are still needed to harness the power of these technologies, without requiring analysts to understand their underlying complexity.

Here, we introduce a new visualization, called *TeMoCo*, which aims to support temporal analysis of conversations. We have developed an interactive prototype tool based on this visualization which is designed for use in clinical settings. We present this prototype, illustrate its use with a case scenario of analysis using a selected corpus of multi-party dialogue of conversations in an A&E unit, and perform a cognitive walk-through to evaluate the prototype.

II. COMMUNICATION IN CLINICAL SETTINGS

Clinical conversations play a major part in medical communication, and extensive literature exists on the topic. Medical communication is a complex process, with both biomedical objectives (e.g. establishing a diagnosis, curing the patient) and humanistic objectives (e.g. mutuality of the relationship, effective communication). The communication is impacted by widely different aspects related to socio-demographic, cultural, and even personality aspects, and will vary with diseases-related characteristics, such as the stage of the illness and patient expectations [3]. In medical team settings, communication takes place while team members are cooperating toward a common goal. For example, in A&E the common goal is saving the patient while conducting a number of tasks under complex constraints, including time pressure, information overload, ambiguous situations, and the risk of severe consequences in case of error. The impact of poor communication in such settings is evident.

Teaching and training for good communication skills are therefore necessary. In medical education, training happens at different stages of the professional life of doctors and nurses. It usually includes simulated interventions, where technical and non-technical skills are assessed. To evaluate medical communication training sessions, and provide feedback, the health community has looked at systematic analysis and problem-solving approaches developed in other life critical domains (e.g. crew resource management from aviation) and have implemented similar solutions in clinical settings [4].

Dedicated frameworks have been developed for the assessment of communication – often referred as non-technical skills – each of which assess different sets of skills. For instance, the Observational Teamwork Assessment of Surgery (OTAS) [5] assesses clinical and technical skills, and also interpersonal skills and behaviours. The Anesthesiologists Non-Technical Skills (ANTS) [6] assesses four different sets of skills: task management, team working, situation awareness, and decision making. Similarly, the Communication And Teamwork Skills (CATS) [7] assesses four sets of skills: situation awareness, coordination, communication,

and cooperation. The above mentioned RIAS framework [1] has also been used in assessing communication skills in these settings.

Despite these efforts, overall, there is a lack of consensus regarding the evaluation of clinical team communication [8], and so far, there are no globally accepted theoretical models for assessment of team performance. However, general studies of team performance and studies specific to healthcare have identified some necessary skills. These studies rely on the observation and analysis of certain conversational behaviours during training encounters.

In this context, the use of tools to extract and visualize conversations can support the temporal analysis of team communication. A visualization tool could provide the analyst with a simple and natural way to navigate conversation sessions and to search for different interactional aspects related to the monitored skills – either punctual (e.g. verbalization of plans and changes, requests for help, use of key phrases) or spanning the whole interaction (e.g. acknowledgement of the concerns of others, closed-loop communication, updates).

A. Related work on visual tools for clinical communication

The most common method used for medical communication training is the video recording of a session followed by an after-action-review. The review is performed either by a professional, or provided to the students – e.g. to write self-reflective structured assessment of their performance. Visualization tools exist to help with the analysis of sessions.

The Lab-in-a-box system [9] uses sensors (3D camera, eye tracking, computer activity) to track the clinician's workflow during a medical consultation. The collected data are presented as events along a timeline representing the consultation. Simple events (key strokes and mouse clicks) are presented directly, and visual attention toward the computer is displayed as blocks. The selection of a block opens a picture showing the corresponding gaze direction track.

EQclinic [10], a fully-fledged system for training of health professionals, records and analyses online sessions with simulated patients. Live feedback is provided by the assessor in the form of comments with positive or negative valence. Post-interaction tools includes manual assessment (forms) and automated analysis of non-verbal communication (turn patterns, prosody, visual cues).

While these systems support complex analytic tasks, the display of single features separately without context is difficult to interpret by non-experts. To facilitate interpretation, specific visualizations of the content of the interactions need to be developed. Addressing this issue, Angus et al. [11] provided a visual representation of content to track conceptual recurrence in the conversation structure of medical consultations. Our approach also aims to address the issue of providing a temporal structure to dialogue content, but we employ a different visual representation that scales

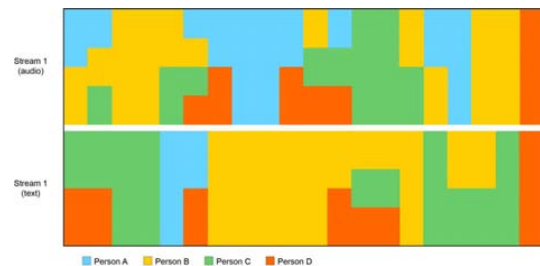


Figure 1. A sketch of temporal mosaic visualization, adapted from [12].

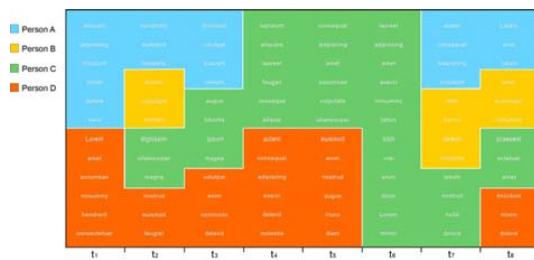
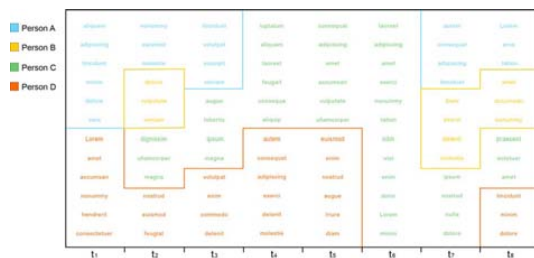
to dialogues with more than two participants (multi-party dialogues), as explained below.

III. *TeMoCo* VISUALIZATION

We have designed the *TeMoCo* (Temporal Conversation Mosaics) visualization to better support visual analysis of conversations. It uses the temporal mosaics visualization [12] as its basis. The original temporal mosaics visualization (see Figure 1) represents the individual time-based data streams separately as synchronized rows of visualizations.

In the case of Figure 1, the top row shows the audio conversations between 4 people, while the bottom row shows their contributions to a text document. Within each row, the temporal mosaics visualization allocates the vertical space equally between the number of participants active in each time-slice (i.e. the horizontal space) – with the resulting sum of individually coloured rectangular shapes showing the contributions of each participant across time. A temporal mosaic visualization is, therefore, used to represent temporal contribution patterns rather than the content of individual contributions. However, when used as an interactive visualization [13], each rectangular segment of a temporal mosaics visualization can be linked to the corresponding part of the data stream it represents – thus supporting access to media content, both temporally as well as contextually.

In the case of analysis of audio recorded conversations, we are only dealing with a single data stream. As such, *TeMoCo* can utilize the visualization space to represent a single data stream using the convention of dividing the vertical space equally between the active conversation participants for each time-slice – similar to the top row of Figure 1. While in an interactive version each mosaic segment can be linked to its corresponding audio recording, *TeMoCo* uses the visual space of each segment to also superimpose a textual summary of the transcript of the corresponding audio speech, making it more useful even in a static mode. This textual summary can take a number of forms, depending on the application area for which the visualization is used. Here, we have chosen to provide a list of keywords from each speech segment, ranked according to their occurrences. Other options could include a word-cloud of keywords for

Figure 2. A sketch of *TeMoCo* visualization.Figure 3. An alternative version of *TeMoCo* with coloured keywords.

each segment (e.g. in a manner similar to Wordle [14]), a representative sentence (e.g. first sentence), etc. Figure 2 provides a sketch of the *TeMoCo* visualization.

Even though *TeMoCo* is visually similar to a temporal mosaic visualization, the additional visual encodings must be carefully considered. The first issue to consider is the visual contrast between the text and the background mosaics. Although an increased contrast would make the text more readable, it would also cause visual distraction – thus reducing visual detection of the background mosaic patterns. As mentioned, detection of these mosaic patterns is an important aspect of the original temporal mosaics visualization, allowing the user to easily view contributions of each of the conversation participants across time, to detect, for instance, any imbalance in levels of contribution, dominance of one participant, and so on. Therefore, although in Figure 2 we use white text on colour mosaics with only hue variations between their colours, ultimately such variations need to be adjusted to suit the static or interactive uses of the visualization. For example, Figure 3 shows an alternative version of *TeMoCo* which might be better for printing in static form.

Another issue to consider in *TeMoCo* visualization is the choice of the number of words for each mosaic segment, as well as the size and style of typefaces used to show the selected words. Further to considering the issue of contrast discussed above, these variations are dependant on the visual and temporal length of each time-slice. Increasing the length of time-slices visually allows for better accommodating

Figure 4. The *TeMoCo* prototype, with the visualization on the left, and transcripts pane on the right.

longer keywords and/or making their typeface size bigger – thus increasing readability. However, increasing the visual length of time-slices may require increasing their temporal length as well. This in turn has its own consequences. For instance, longer temporal time-slices would have longer transcripts to be represented (e.g. requiring more keywords to be selected). Furthermore, if the time-slices are too long, then they may end up including every conversation participant in each slice, and as such, reduce visual effectiveness of mosaic patterns. Once again, these issues are application dependant and must be considered for each use case.

A. Prototype

We have developed an interactive prototype tool which uses *TeMoCo* to visualize multi-party conversations, aiming at supporting the visualization of communication among medical team members. Figure 4 shows the interface of the *TeMoCo* prototype. As can be seen, the left-hand panel is the interactive visualization showing the temporal mosaic patterns of the conversation – along with the top keywords selected from each speaker turn – and the right-hand panel shows the transcript of the entire conversation session. In this conversation session there are five participants (Patient 1, Nurse 1, Doctor 1, Doctor 2, and Medical Registrar 1), who have been talking for 13 minutes and 30 seconds.

While the static view of *TeMoCo* is sufficient for seeing the patterns of conversation, and a summary of its main keyword points, the user can get a detail-on-demand view by clicking on a speaker turn mosaic on the visualization to access the relevant parts of the conversation on the transcript. Figure 5 shows a selected mosaic (in gray colour) on the left for participant D1, between 04:30 and 06:00. By selecting a mosaic, the prototype tool locates the start of the transcript text related to the selected time-slice (04:30–06:00), grays out the background of all the text for that time-slice, and then highlights the segments of the transcript text for the chosen speaker during that time-slice using the colour assigned to that speaker (the orange colour for D1 in Figure 5, the blue colour for P1 in Figure 6).



Figure 5. The *TeMoCo* prototype, with a speaker turn selected on the visualization (grayed out mosaic on the left), and the relevant parts of the transcripts highlighted (orange background text on the right).



Figure 6. The *TeMoCo* prototype, with another speaker turn selected on the visualization (grayed out mosaic on the left), and the relevant parts of the transcripts highlighted (blue background text on the right).

B. Implementation

Figure 7 shows the architecture of the *TeMoCo* prototype which has been implemented as a single-page web application using the *D3.js* framework [15]. The current system creates the visualization using a transcript file made available to it on the server. The transcript text is time-stamped and tagged with the labels of the conversation participants.

The system starts by pre-processing the transcript text to create two data streams. The first stream generates a data source containing relevant keywords, in which the keywords

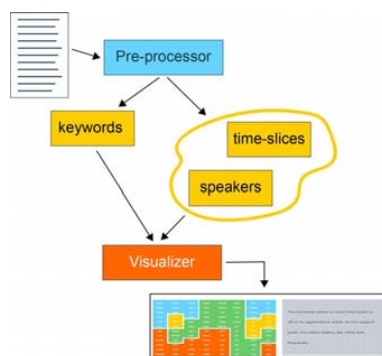


Figure 7. Architecture of the *TeMoCo* prototype.

are selected for each time-slice and participant combination. Keyword salience is dependant on context and use case – measures such as word frequency or frequency in a domain specific reference corpus are an obvious starting point. In our tests we found raw frequency to be uninformative, and subsequently decided on the manual selection of seemingly salient words, this simulates the word selections that could be achieved automatically using a medical reference corpus. Depending on the corpus and use case, any statistical measure of word salience or keyness could be injected to produce the keywords for a speaker in a time-slice.

The second stream of data is generated by extracting the time-slice and speaker information. This information is then used for tagging the input transcript with HTML attributes. This enables dynamic manipulation of the raw transcript as a part of the system interface.

Once the two data streams have been processed, the system constructs the temporal mosaics of the *TeMoCo* visualization from the time-slices, speakers and keywords information. The visualization and transcript panels are then positioned in the same web page. Both views are linked via the data, allowing interactions between the two. Selection of a time-slice mosaic scrolls the transcript to the corresponding time-slice, as describe above.

IV. EVALUATION

We have conducted an initial evaluation of *TeMoCo* using an existing corpus of transcribed medical conversations. We employed the cognitive walk-through analytical evaluation methodology to assess the use of *TeMoCo* on these data, in a medical education scenario.

A. Test conversations data set

To test our visualization, we selected a corpus of multi-party dialogues recorded in a hospital in Ireland, as part of the INCA (Interaction Analytics for Automatic Assessment of Communication Quality in Primary Care) project. The corpus was created for the development of tools for automatic analysis of verbal and non-verbal communication, to assess communication quality in different contexts of medical interaction. The corpus consists of simulation-based team training sessions for health professionals intervening in medical emergencies (e.g. accident and emergency services).

Each session follows a scenario with a specific medical problem selected by teaching staff. At the start of the training, the simulated patient – a dummy on a bed played by an actor outside the room – is showing rapid signs of health deterioration. The team must jointly establish a diagnosis and provide relevant care. Vital signs of the simulated patient are displayed on a patient monitoring equipment by the bed. Each recording features a nurse and two doctors. The nurse is present from the beginning and calls a doctor after detecting the abnormality. As the problem get more serious, a second doctor is called, If specific difficulties or questions

Total	Duration		Participants	Utterances
	Session	Turn		
207min	14min47s (11'20 - 23'01)	01.87s (1.1 - 3.9)	5.6 (5 - 8)	450 (296 - 680)

Table I
A SUMMARY OF THE CHARACTERISTICS OF THE INCA MEDICAL CONVERSATIONS CORPUS.

arise, the medical team could call specialist registrars (e.g. anaesthetic, orthopedics, etc.) on a telephone. A third doctor is sometimes present, as well as a second medical registrar.

A total of 14 training sessions have been recorded, segmented and transcribed. An overview of the main statistics of the data set used in our testing is provided in Table I.

B. Cognitive walk-through

Cognitive walk-through requires setting the objectives and task for the user of the system, and walk through each to assess the usability and capacity of the system to fit its role.

1) *Persona*: A persona is a prototypical user whose knowledge and behaviour are representative of the target users of the system. The following personae were identified:

Trainer: Dr Grey

The trainer is a professor and a doctor, expert in the field of medical education. She is 40 years old, has trained students for 5 years and has already established routines for each pedagogic goal. She has average competency in computing, and will use the software as a tool to improve the impact of her feedback. She is not interested in the underlying technology, and the tool interface must be easy to understand and use for her purposes. During the training she will observe the session and take notes on paper regarding the different skills under evaluation. After the session, she will want to navigate the session and illustrate global and punctual aspects of the communication and certain events that happened during the session, either good or bad.

Learner: B. Bagins

The learner is a medical student, she has already studied for 4 years and finds it difficult to see the value of non-medical skills. Once she has completed the training session successfully, she needs to be given feedback on her own behaviour, as well as her role in the team. She needs to visualize directly the points made by the trainer.

2) *Individual actions*: We defined a set of criteria for evaluating the training interactions between the two personas participating in debriefing sessions. Each of the training interactions are then evaluated using these four criteria:

- 1) Is the effect of the user's action corresponding to the user's goal?
- 2) Is the action visible?
- 3) Is the action identifiable as being the correct one?
- 4) Is the feedback understandable?

Each task and comments on each of these criteria are presented below:

Access to the session and overview of the conversation and its main points. The trainer starts the *TeMoCo* prototype, and both personas look at the visualization, with no further action needed (see Figure 4). The trainer wants to see the general structure of the conversations and identify any global patterns (e.g. distribution of speech related to communication and cooperation).

- 1) the effect is immediate with the visualization and corresponding transcript visible.
- 2) yes.
- 3) if a single session is accessed, yes. If multiple sessions are available, a label with the ID/date/participants of the sessions would be needed to identify the correct session.
- 4) feedback is natural: the legend displays each participant labelled with a single colour, corresponding to the one used in the visualization. Temporal visualization of conversations is immediately visible. For each mosaic segment, the set of keywords provides an overview of the main items in the conversation. A visualization of the links between recurring terms – or semantically similar terms – would help materialize this recurrence.

Navigation through the session to select points of interest (e.g. new problems, new participant, etc.) The trainer uses the visualization and the shown keywords to select a specific segment where something of interest occurred. The use of task specific keywords (e.g. medical terms, requests, concerns) illustrates cooperative behaviour. The selected area is grayed out, and the participant's utterances of the corresponding time-slice are displayed and highlighted using the participant's colour in the transcript window (see Figure 5). The learner can see the focus of the current point being discussed.

- 1) yes.
- 2) yes.
- 3) the trainer will need to search the keywords across participant's utterances that were produced in the selected time-slice. Highlighting salient keywords in the transcript would help to contextualize them.
- 4) yes.

Illustration of a specific exchange over a few conversation turns by going through the details of the conversation. The trainer is looking at the visualization to pick up keywords, and scrolls the transcript window to use the corresponding detail of the conversation and switches between participants to visualize their turns. The trainer will use the mosaic segments of visualization and the transcript window to search for events marked in her notes (see figures 5 and 6).

- 1) yes.
- 2) yes, the side bar and utterances of the selected participant scrolls.
- 3) yes, but part of the utterances of a time-slice do

not fit in the browser. However, utterances belonging to the selected time-slice are delimited by a gray background.

- 4) yes, the side bars are standard and are commonly used in most interfaces.

Identification of conversation patterns from the global interaction to local interactions (e.g. cooperation, coordination, turn-taking behaviour, etc.) The trainer and learner look at the visualization to see the global structure of the conversations (e.g. contributions of each participants, occurrences of keywords, etc.). Specific behaviour leading to patterns of interest (turn taking behaviour) is accessed through the transcript (see Figure 6)

- 1) Utterances of other speakers are grayed out, making it difficult to see the sequence of speakers. The use of faded colour would allow an easier interpretation while keeping the significance of the gray/user coloured duality.
- 2) yes.
- 3) yes, but the user will need to read and point her selected local points of interest.
- 4) Local points of interest that are not shown on the visualization require to scroll through the transcripts and may be difficult to find within long temporal visualization time-slices.

V. CONCLUSION

We presented a temporal text visualization tool to support the analysis and exploration of transcripts of medical team communication. Testing was carried out on data from simulated situations in a critical care (accident and emergency) setting, and the usefulness and usability of *TeMoCo* to support medical education was assessed. The mosaic-based design was found to be effective in providing a contextualized, temporal view of the conversation. Future work will focus on incorporating further textual structure, such as topics and conversational threads, to the visualization, and on testing the tool in other analysis tasks, such as assessment of patient-doctor communication.

VI. ACKNOWLEDGEMENT

This paper is supported by European Unions Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains. Pierre Albert has been funded through the INCA project. We thank the INCA project members in Ireland for granting us access to the trainee data.

REFERENCES

- [1] D. Roter and S. Larson, "The roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions," *Patient education and counseling*, vol. 46, no. 4, pp. 243–251, 2002.
- [2] P. Ryan, S. Luz, P. Albert, C. Vogel, C. Normand, and G. Elwyn, "Using artificial intelligence to assess clinicians' communication skills," *BMJ*, vol. 364, 2019.
- [3] L. M. L. Ong, J. C. J. M. de Haes, A. M. Hoos, and F. B. Lammes, "Doctor-patient communication: A review of the literature," *Social Science & Medicine*, vol. 40, no. 7, pp. 903–918, Apr. 1995.
- [4] D. M. Gaba, S. K. Howard, K. J. Fish, B. E. Smith, and Y. A. Sowb, "Simulation-Based Training in Anesthesia Crisis Resource Management (ACRM): A Decade of Experience," *Simulation & Gaming*, vol. 32, no. 2, pp. 175–193, Jun. 2001.
- [5] A. N. Healey, S. Undre, and C. A. Vincent, "Developing observational measures of performance in surgical teams," *BMJ Quality & Safety*, vol. 13, no. suppl 1, pp. i33–i40, 2004.
- [6] R. Flin and N. Maran, "Identifying and training non-technical skills for teams in acute medicine," *BMJ Quality & Safety*, vol. 13, no. suppl 1, pp. i80–i84, 2004.
- [7] A. Frankel, R. Gardner, L. Maynard, and A. Kelly, "Using the Communication and Teamwork Skills (CATS) Assessment to Measure Health Care Team Performance," *The Joint Commission Journal on Quality and Patient Safety*, vol. 33, pp. 549–558, Sep. 2007.
- [8] D. P. Baker, S. Gustafson, J. Beaubien, E. Salas, and P. Barach, "Medical teamwork and patient safety: the evidence-based relation," *AHRQ publication*, vol. 5, no. 53, pp. 1–64, 2005.
- [9] N. Weibel, S. Rick, C. Emmenegger, S. Ashfaq, A. Calvitti, and Z. Agha, "LAB-IN-A-BOX: semi-automatic tracking of activity in the medical office," *Personal and Ubiquitous Computing*, vol. 19, no. 2, pp. 317–334, 2015.
- [10] C. Liu, R. L. Lim, K. L. McCabe, S. Taylor, and R. A. Calvo, "A web-based telehealth training platform incorporating automated nonverbal behavior feedback for teaching communication skills to medical students: a randomized crossover study," *Journal of medical Internet research*, vol. 18, no. 9, 2016.
- [11] D. Angus, B. Watson, A. Smith, C. Gallois, and J. Wiles, "Visualising conversation structure across time: Insights into effective doctor-patient consultations," *PloS one*, vol. 7, no. 6, p. e38014, 2012.
- [12] S. Luz and M. Masoodian, "Visualisation of parallel data streams with temporal mosaics," in *Proceedings of the 11th International Conference Information Visualization*, ser. IV '07, July 2007, pp. 197–202.
- [13] —, "A model for meeting content storage and retrieval," in *Proceedings of the 11th International Multimedia Modelling Conference*, ser. MMM '05, Jan 2005, pp. 392–398.
- [14] F. B. Viegas, M. Wattenberg, and J. Feinberg, "Participatory visualization with wordle," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1137–1144, 2009.
- [15] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011.

Appendix E: TeMoCo-Doc: A visualization for supporting temporal and contextual analysis of dialogues and associated documents

TeMoCo-Doc: A visualization for supporting temporal and contextual analysis of dialogues and associated documents

ABSTRACT

A common task in a number of application areas is to create textual documents (e.g. summary reports) based on recorded audio data (e.g. interviews). Visualizations designed to support such tasks require linking temporal audio data with contextual data contained in the resulting documents. In this paper, we present a tool for the visualization of temporal and contextual links between recorded dialogues and their summary documents.

CCS CONCEPTS

• **Human-centered computing** → *Visualization design and evaluation methods; Information visualization*; • **Applied computing** → *Computer-assisted instruction*.

KEYWORDS

Assessment of medical communication; temporal visualization; temporal mosaics; speech visualization; contextual visualization

ACM Reference Format:

. 2020. TeMoCo-Doc: A visualization for supporting temporal and contextual analysis of dialogues and associated documents. In *Proceedings of AVI 2020: International Conference on Advanced Visual Interfaces (AVI 2020)*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Written reports of events or actions which have related recorded audio or video data are widespread across a variety of domains. For instance in the media domain, web-based news articles are often presented with video clips of the reported events. Similarly, medical contexts where recording and reporting is commonplace, audio and video recordings are often made during clinician-patient consultations, multidisciplinary medical team meetings and training, and so on. In these medical contexts, effective verbal communication is crucial to the success of clinical encounters, and as such, several frameworks have been developed to help standardise the analysis of medical communication (such as the widely used Roter Interaction Analysis System (RIAS) [6]).

Such analyses are, however, often very time-consuming and not well supported by existing software or visualization tools. For instance, the link between an audio recording and a textual document is rarely made explicit, thus making it difficult to quickly switch between the textual document and the recording. To identify which spans of text in a report are linked to particular times in a recording

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

AVI 2020, June 8–12, 2020, Island of Ischia, Italy
© 2020 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . \$15.00
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

requires either a full examination of both sources, or some other linking mechanism.

In this paper we present an interface which maps a textual document via contextual links to a temporal visualization of the contents of a recording and individual speaker's contributions. In addition, both the document and temporal visualization are implicitly linked to a transcript of the recording (which could easily be replaced by a highlighted timeline with accompanying video or audio). This interface is an extension of the *TeMoCo* visualization [removed for review] which visualizes the speakers involved in each time-slot (without showing the amount of their speech contributions) in a recording, and enables exploration of the transcript via interaction with a temporal mosaic visualization. While this prior work has focused on identifying the temporal content, the work presented here focuses on identifying the links between the contents of textual documents and the temporal speech segments of recorded audio. In this paper we describe this visualization tool (called *TeMoCo-Doc*).

2 THE TEMOCO-DOC VISUALIZATION

Our proposed visualisation *TeMoCo-Doc* combines three juxtaposed views [2]: the temporal mosaic (Figure 1 left), the transcript (Figure 1 right), and the view containing the report (Figure 2) which links all three views via manually encoded contextual edges. The interface is rendered across two webpages, the first contains the report, the second contains both the temporal and transcript views. These two webpages have interactions enabled via the JavaScript broadcast channels API.

The temporal aspect of the visualization, seen in the prototype Figure 1, uses temporal mosaic visualizations [4] to render the speaker contribution per time-slot and displays the salient word for each speaker in the corresponding slot. This choice of encoding is informed by Mackinlay's ranking [5] of visual variables [1]. The order of the time-slots is mapped to horizontal position, the quantity of speaker contribution is mapped to length, and the categorical speaker information is mapped to color hue. The transcript is placed next to the temporal mosaic, and the vertically-ordered speech segments are placed in a scroll box.

When used as an interactive visualization [3] each rectangular segment of a temporal mosaics visualization can be linked to the corresponding part of the data-stream it represents – thus supporting access to media content, both temporally as well as contextually. In Figure 1 one speaker has been selected in four time-slots. This causes the transcript to scroll to the beginning of the first selected time-slot, and the corresponding speech segments are highlighted. This follows the well-known visual information seeking mantra [8]. This combination of temporal mosaic and transcript is based on the *TeMoCo* visualization [7] which displays speakers for each time-slot evenly (i.e. not based on the amount of their speech contribution) and does not allow for contextual linking across multiple time-slots.

AVI 2020, June 8–12, 2020, Island of Ischia, Italy

Mini temporal mosaics (with the salient words removed) are rendered next to each section of the report view. Spans of text in each section which were identified as being contextually linked to the transcript are colored according to the speaker they map to. By hovering over any of these contextual spans the user is able to see the related time-slots on the mini mosaic and the highlight is maintained if the span is clicked (as can be seen in the third mosaic in Figure 2), which gives a preview of the slots that will be selected by clicking on the span. Each of the document sections is presented in a fixed height scroll-box, so that even large documents can be explored in a single screen.

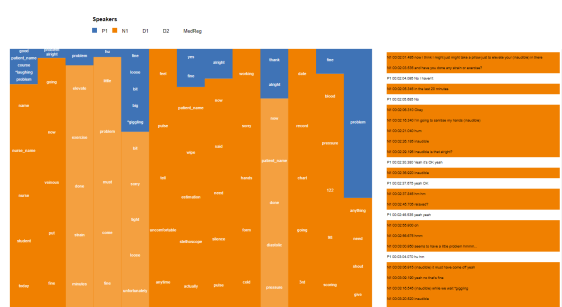


Figure 1: Temporal mosaic (left) and transcript (right) in the TeMoCo-Doc visualization. A contextual edge has been selected and the related speaker time segments have been highlighted. The related sections of the transcript have also been highlighted.

The document example shown across 1 and 2 represents a reflexive report based on a medical communication training exercise. In Figure 2 the user can see the span “I should not have become so flustered as I then forgot to change the obesity cuff to a standard cuff perhaps leading to the high diastolic value I recorded” is clicked, outlining the related time-slots on the mini mosaic. Looking at the large mosaic 1, the user can see the same time-slots are again selected and the transcript is highlighted accordingly. This allows the user to explore the transcript, and quickly see the time-slots and content in the conversation which relate to “I plan to check all equipment before beginning procedures”.

The design of the TeMoCo-Doc visualization was informed by collaboration with colleges involved in medical education. Preliminary evaluation of the prototype led to some adjustments after experts’ feedback. For instance, inconsistent use of highlighting/greying-out techniques across views were changed to unify the representation of active sections of interest in the mosaic visualization and the transcripts, and the contrast and selection of colour was checked for accessibility of the tool to colour-blind users. Some of the medical domain experts had difficulty with the representation of time using different axis: top to bottom for the transcripts, left to right for the mosaic. While the use of 3 panels for the visualization was received positively, the number of words rendered in the temporal view were seen as potentially overloading, and their meanings were not immediately obvious, needing further explanation (salient words not necessarily reflecting medical issues or problems).

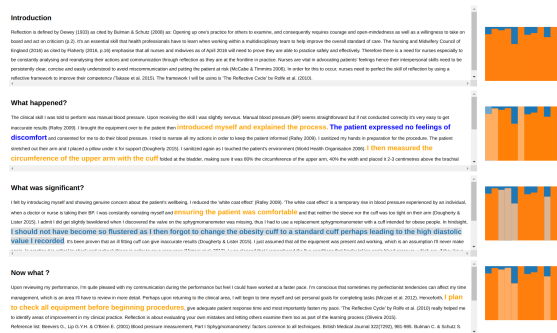


Figure 2: Textual document with contextual links in the TeMoCo-Doc visualization. A contextual link is selected and the related speaker time segments have been highlighted on the corresponding mini temporal mosaic. The related sections of the transcript have also been highlighted.

Overall, the experts found the visualization interesting as a learning tool, but not sufficient as a primary assessment tool for scenario based medical training. A major limitation for the use in assessment is due to the diversity and complexity of aspects used to evaluate communication and technical skills which cannot be reflected by text alone. As such, the experts stated that the prototype could not be used to replace watching the full session of a student’s interaction. However, the visualization tool was considered for its primary purpose and the visualization is expected to be useful to compare students.

The potential of this tool for distant learning in medical training was suggested by the domain experts. The visualization could either allow teachers to point specific behaviours (to share feedback with students), or it could be used by the students to share comments on practice performances. Further proposed development of the tool for e-learning involved adding the capacity to visualise non-verbal communication. The experts suggested adding a video stream, synchronised with the contextual links and transcript. Another proposed application was the visualization of the context of free-text searches in Electronic Health Records.

3 CONCLUSIONS

In this paper we have presented a visualization tool which provides temporal links between a recorded speech audio (dialogues) and its transcript, and the contextual links with their related text document. The tool was tested on a corpora of medical and clinical skills training exercises, used to assist in the assessment and evaluation of students’ communication skills.

We are currently in the process of further developing the interactive prototype. Future developments will include expanding information contained in each of the three panels of the visualization (e.g. dialogue aspects, natural language processing and text analysis). We will also investigate other application areas and tasks for which our visualization tool could be use. Following these developments we will carry out formal evaluations of the visualization tool in a number of application specific areas.

ACKNOWLEDGMENTS

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflect only the author's view and the Commission is not responsible for any use that may be made of the information it contains. Pierre Albert has been funded through the INCA project. We thank the INCA project members in Ireland for granting us access to the trainee data.

REFERENCES

- [1] Jacques Bertin. 1983. *Semiology of Graphics*. University of Wisconsin Press.
- [2] W. Javed and N. Elmqvist. 2012. Exploring the design space of composite visualization. In *Pacific Visualization Symposium (PacificVis)*, 2012 IEEE. 1–8. <https://doi.org/10.1109/PacificVis.2012.6183556>
- [3] S. Luz and M. Masoodian. 2005. A Model for Meeting Content Storage and Retrieval. In *Proceedings of the 11th International Multimedia Modelling Conference (MMM '05)*. 392–398. <https://doi.org/10.1109/MMMC.2005.12>
- [4] S. Luz and M. Masoodian. 2007. Visualisation of Parallel Data Streams with Temporal Mosaics. In *Proceedings of the 11th International Conference Information Visualization (IV '07)*. 197–202. <https://doi.org/10.1109/IV.2007.127>
- [5] Jock Mackinlay. 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics* 5, 2 (1986), 110–141. <https://doi.org/10.1145/22949.22950>
- [6] Debra Roter and Susan Larson. 2002. The Roter interaction analysis system (RIAS): utility and flexibility for analysis of medical interactions. *Patient education and counseling* 46, 4 (2002), 243–251.
- [7] Shane Sheehan, Pierre Albert, Masood Masoodian, and Saturnino Luz. 2019. TeMoCo: A Visualization tool for Temporal Analysis of Multi-Party Dialogues in Clinical Settings. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 690–695.
- [8] Ben Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings the IEEE Symposium on Visual Languages*. 336–343. <https://doi.org/10.1109/VL.1996.545307>

Appendix F: The NetViz terminology visualization tool and the use cases in karstology domain modeling

*Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pages 55–61
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC*

The NetViz terminology visualization tool and the use cases in karstology domain modeling

Senja Pollak¹, Vid Podpečan¹, Dragana Miljkovic¹, Uroš Stepišnik² and Špela Vintar²

¹Jožef Stefan Institute, Ljubljana, Slovenia
{senja.pollak,vid.podpecan,dragana.miljkovic}@ijs.si

²Faculty of Arts, University of Ljubljana, Slovenia
{uros.stepisnik,spela.vintar}@ff.uni-lj.si

Abstract

We present the NetViz terminology visualization tool and apply it to the domain modeling of karstology, a subfield of geography studying karst phenomena. The developed tool allows for high-performance online network visualization where the user can upload the terminological data in a simple CSV format, define the nodes (terms, categories), edges (relations) and their properties (by assigning different node colors), and then edit and interactively explore domain knowledge in the form of a network. We showcase the usefulness of the tool on examples from the karstology domain, where in the first use case we visualize the domain knowledge as represented in a manually annotated corpus of domain definitions, while in the second use case we show the power of visualization for domain understanding by visualizing automatically extracted knowledge in the form of triplets extracted from the karstology domain corpus. The application is entirely web-based without any need for downloading or special configuration. The source code of the web application is also available under the permissive MIT licence, allowing future extensions for developing new terminological applications.

Keywords: Terminology visualization, Karstology, Domain modeling, Networks

1. Introduction

Visual representations of specialized domains are becoming mainstream for several reasons, but firstly as a natural response to the fact that “concepts do not exist as isolated units of knowledge but always in relation to each other” (ISO 704, 2009). In recent terminological projects, visualization has been considered an important asset (Faber et al., 2016; Carvalho et al., 2017; Roche et al., 2019). We believe that the visualization of terminological knowledge is especially well-suited to the needs of frame-based terminology, aiming at facilitating user knowledge acquisition through different types of multimodal and contextualized information, in order to respond to cognitive, communicative, and linguistic needs (Gil-Berrozpe et al., 2017). Moreover, it has been shown that domain experts are often able to interpret information faster when viewing graphs as opposed to tables (Brewer et al., 2012). More generally, as has become evident in the rising field of digital humanities, digital content, tools, and methods are transforming the entire field of humanities, changing the paradigms of understanding, asking new research questions and creating new knowledge (Hughes et al., 2015; Hughes, 2012).

As this workshop demonstrates, terminological work has undergone a significant change with the emergence of computational approaches to extracting various types of terminological knowledge (e.g., term extraction, definition extraction, semantic relation extraction), which enhances the potential of visualization not only to represent manually annotated data, but also for automatically and semi-automatically extracted knowledge, which we also show in our use cases.

We focus on the field of karstology, the study of specific relief which develops on soluble rocks such as limestone and is characterized by caves, typical depressions, karst springs, ponors and similar. It is an interdisciplinary subdomain of

geography bordering on geomorphology, geology, hydrology and chemistry. In karstology, the main objects of interest are its typical landforms usually described through their form, size, location and function, and the environmental and chemical processes affecting their development such as dissolution and weathering.

The proposed semantic network visualization tool NetViz¹ used in the presented karstology domain modeling experiments, complement our previous research in the TermFrame project including work of Vintar et al. (2019) where frame-based annotation of karst definitions is presented, Pollak et al. (2019) presenting results of term and definition extraction from karst literature, Miljkovic et al. (2019) with term co-occurrence network extraction and Grčić-Simeunović and De Santiago (2016) where semantic properties of karst phraseology are explored.

2. Related Work

There are several projects which consider *terminology visualization* as an important asset of specialized knowledge representation. One such project is the EndoTerm, a knowledge-based terminological resource focusing on endometriosis (Carvalho et al. 2016, Roche et al. 2019). EndoTerm includes a visual concept representation developed via CMap Tools and organizes knowledge into semantic categories linked with different types and levels of relations, while ensuring compatibility with existing medical terminology systems such as SNOMED. The most closely related project to ours using a visual representation of specialized knowledge is the EcoLexicon (Faber et al., 2016), where terms are displayed in a semantic network linking the central query term to related terms and its translation equivalents in up to 5 other languages. The edges of the network represent three types of relations, namely the generic-

¹<https://biomine.ijs.si/netviz/>

specific (is_a) relation, the part-whole relation and a set of non-hierarchical relations (made_of, located_at, affects etc.). While the EcoLexicon remains impressive with the abundance and complexity of data it offers, our own approach differs mainly in that we use natural language processing techniques to infer data, and that we envisage different types of visual representation depending on the task or end-user.

In terms of *domain modeling of terminological knowledge*, we can first mention the field of terminology extraction. In automatic terminology first the distinction was between linguistic and statistical approaches, but most state-of-the-art systems are hybrid. Many terminology extraction algorithms are based on the concepts of termhood and unithood (Kageura and Umino, 1996), where termhood-based approaches include work by Ahmad et al. (2000) and Vintar (2010), while Daille et al. (1994) and Wermter and Hahn (2005) use unithood-based measures, such as mutual information and t-test, respectively. More recently, deep learning and word embeddings (Mikolov et al., 2013) have become very popular in natural language processing, and several attempts have already been made to utilize these techniques also for terminology extraction (Amjadi et al., 2016; Zhang et al., 2017; Wang et al., 2016) and terminology expansion (Pollak et al., 2019). Next, for defining relations between terms, there are several relation extraction methods, which can roughly be divided into categories: co-occurrence-based, pattern-based, rule-based and machine-learning approaches (Bui, 2012; Sousa et al., 2019). Co-occurrence is the simplest approach which is based on the assumption that if two entities are frequently mentioned together in the same sentence, paragraph or document, it is probable that they are related (Song et al., 2011). The pattern- and the rule-based differ in that the former use template rules, whereas the latter might additionally implement more complex constraints, such as checking negation, determining the direction of the relation or expressing rules as a set of procedures or heuristic algorithms (Kim et al., 2007; Fundel-Clemens et al., 2007). Machine-learning approaches usually set the relations extraction tasks as classification problems (Erkan et al., 2007). Recently, the proposed approaches often use the power of neural networks as in Lin et al. (2016), Sousa et al. (2019), Luo et al. (2020). The focus of this paper is the visualization tool and its use in karstology domain modeling. For data extraction, we employ several techniques mentioned above. Pattern-based methods (Pollak et al., 2012) are used for definition extraction in the first use case (Section 4.3.) providing definition candidates for further manual annotation of domain knowledge, while in the second use case (Section 4.4.) we use statistical term extraction techniques (Vintar, 2010; Pollak et al., 2012) coupled with co-occurrence analysis and relation extraction using Reverb (Fader et al., 2011).

3. NetViz

Network visualization is of key importance in domains where an optimized graphical representation of linked data is crucial in revealing and understanding the structure and interpreting the data with the aim to obtain novel insights and form hypotheses. There is a plethora of software which

deals with network analysis and visualization. For example, Gephi (Bastian et al., 2009), Pajek (Batagelj and Mrvar, 2002) and Graphviz (Ellson et al., 2001) are among the most popular classic software tools for these tasks and have been used in very diverse domains. However, every domain and every task poses specific requirements and using tools which are too general is often a poor choice which has adverse effects on usability. Therefore, our aim was to provide a minimal environment which enables zero effort network visualization for specific tasks such as terminology. We developed NetViz (<https://biomine.ijs.si/netviz/>), a web application which enables interactive visualization of networks. NetViz builds upon our previous work on visualization and exploration of heterogeneous biological networks (Podpečan et al., 2019), where several large public databases are merged into a network which can then be explored, analyzed and visualized. We applied the same principles and created a domain independent network visualization tool which was then applied to karstology domain modeling and exploration.

3.1. Features

- **Open source.** Netviz is available under the liberal MIT license on the open source portal GitHub².
- **Single page, client-only web application.** NetViz is implemented as a client-only web application. As a result, NetViz requires no hosting and server configuration and can be also run locally simply by downloading and opening its html page in a web browser.
- **High performance network visualization.** NetViz implements a user interface around the `vis-network` module of the `vis.js` visualization library. `vis-network` is a fast, highly configurable library for network visualization in the browser and NetViz builds upon its visualization engine.
- **Visualization and editing features.** A set of fundamental network editing and visualization features are implemented. The network can be modified after visualization by adding or removing nodes and edges. Several settings controlling the physics simulation which does the layouting can be adjusted before, during or after the visualization. Context menus which are available on all elements (node, edges and the canvas itself) provide a few basic options which can be extended according to the requirements of the specific domain.
- **CSV data format.** In order to make the use of NetViz as simple as possible its data input format is a comma separated file (CSV) with header. Two files are used: the first one which is mandatory defines edge properties while the optional second file defines node properties. The header for edge definition file supports the following columns: `node1`, `node2`, `arrow`, `label`, `text`, `color`, and `width` where `node1`, `node2`, and `arrow` are mandatory and the rest is optional. The header for node definition file supports the following columns:

²<https://github.com/vpodpecan/netviz>

node, text, color, and shape. We expect that the list of supported columns (features) will grow and adapt to specific domains where NetViz will be used. We will also add the option to export the current network so that the user modifications of the network will not be lost upon closing the application.

The intended users are domain experts in the process of construction of a domain ontology, terminologists, as well as students and teachers. It also has potential for being used by larger public with some modifications and a fixed domain knowledge base.

4. Karstology Domain Modeling

4.1. The TermFrame Project

The context for this research is the TermFrame project which employs the frame-based approach to build a visual knowledge base for karstology in three languages, English, Slovene and Croatian. The main research focus of the project is to explore new methods of knowledge extraction from specialized text and propose novel approaches to knowledge representation and visualization (see previous work in the project described in Vintar et al. (2019), Pollak et al. (2019), Miljkovic et al. (2019)).

The frame-based approach in terminology (Faber, 2012; Faber, 2015) models specialized knowledge through conceptual frames which simulate the cognitive patterns in our minds. According to this view, a frame is a mental structure consisting of concept categories and relations between them. Unlike hand-crafted ontologies, frame-based terminology uses specialized corpora to induce frames or event templates, thus consolidating the conceptual and the textual level of a specialized domain.

Such an approach to knowledge and terminology modeling has a lot to gain from graph-like representations, because its building blocks are concept categories, concepts and terms as nodes, and various types of hierarchical and non-hierarchical relations as edges. By selecting different layers of representation it is thus possible to visualize the dynamic and multidimensional nature of specialized knowledge.

In the TermFrame project we combine manual and computational methods to extract domain knowledge. However, in an ideal scenario, as many steps as possible would be automated requiring only minimal manual validation. The main steps of our proposed domain modeling workflow can be summarized as follows:

- Convert documents to plain text format.
- Identify domain terms.
- Identify domain definitions.
- Identify semantic categories.
- Identify semantic relations.
- Select information for network visualization.
- Visualize the network.
- Interactively explore and modify the terminological resource.

Details on automated knowledge extraction for several of these steps are provided in Pollak et al. (2019). In the following subsections, we present the corpus, as well as two experiments on karstology domain modeling, where a subset of steps above are performed manually or automatically, before the final steps of visualization and interactive exploration using NetViz, which is the focus of this paper and common to both experiments.

4.2. Corpus

The English part of the TermFrame corpus, which was used in these experiments, contains 56 documents of different length, all pertaining to karstology. It includes books, research articles, theses and textbooks (for more details see Vintar et al. (2019)). We used Google Documents feature for conversion of documents from pdf to text format. Frequently such conversion introduced errors into the document such as additional line breaks or orphaned figure captions in the middle of paragraphs. Such errors were corrected in the post-processing phase either manually or using simple scripts.

4.3. Visualizing Manually Annotated data

In this experiment we use manual annotations of domain definitions. Specialized definitions were first either identified in dictionaries and glossaries or using definition extractor from domain texts (Pollak et al., 2012)³, and next annotated with a hierarchy of semantic categories and a set of relations which allow to describe karst events. For an example of annotated definition see Figure 1. The annotation process—performed by linguists and domain experts—is described in detail in Vintar et al. (2019) and briefly summarized below.

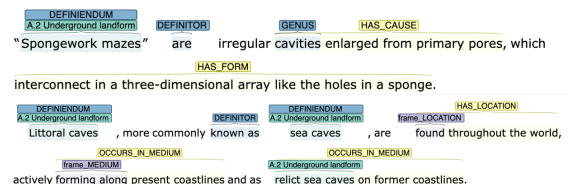


Figure 1: Manual annotation of automatically extracted definitions.

The semantic categories were inspired by the concept hierarchy in the EcoLexicon⁴ and adapted to karstology by domain experts. The first three top-level categories, LANDFORMS, PROCESSES and GEOMES, are the most relevant for domain modeling as they contain terms specific to karst, while the rather broad group of ELEMENTS, ENTITIES and PROPERTIES contains broader terms from geography, chemistry, botany and similar. INSTRUMENTS and METHODS are used to categorize karstology-specific

³The evaluation of automated definition extraction is described in detail in Pollak et al. (2019). About 30% of extracted definition candidates were judged as karst or neighbouring domain definitions, while about 16% of definition candidates were evaluated as karst definitions used for the fine-grained manual annotation.

⁴<https://ecolexicon.ugr.es/en/index.htm>

research and/or measurement procedures, but were found to occur rarely in our set of definitions.

The second important level of annotation identifies the semantic relations which describe specific aspects of karst concepts. According to the geomorphologic analytical approach (Pavlopoulos et al., 2009), landforms are typically described through their spatial distribution (HAS_LOCATION; HAS_POSITION), morphography (HAS_FORM; CONTAINS), morphometry (HAS_SIZE), morphostructure (COMPOSITION_MEDIUM), morphogenesis (HAS_CAUSE), morphodynamics (HAS_FUNCTION), and morphochronology (OCCURS_IN_TIME). The ideal definition of a landform would include all of the above aspects, but in reality most definitions extracted from the corpus or domain-specific glossaries specify only two or three. In total, 725 definitions were annotated, 3149 terms were assigned categories.

In this experiment we focus on the visualization of the taxonomy built from manually annotated categories of DEFINIENDUM and their hypernyms, connected by IS_A relation to their subcategories and categories (LANDFORM, PROCESS, GEOME, ELEMENT/ENTITY/PROPERTY, and INSTRUMENTS/METHODS). The top level—taxonomy of categories—can be observed in Figure 2. In Figure 3, we can see lower levels, which correspond to terms from definitions, more specifically terms (definiendums) assigned to specific subcategories of Hydrological forms and Underground landforms. It allows the user to quickly grasp the main conceptual properties of hydrological forms, namely that water in karst continuously submerges underground (*sinking creek, losing streamflow, swallow hole etc.*) and reemerges to the surface (*karst spring, resurgence, vauclusian spring etc.*), depending on the porosity of the underlying bedrock. Amongst underground landforms we can quickly discern various types of caves (*crystal cave, lava cave, active cave, bedding-plane cave, roofless cave*) and typical underground formations found in them (*straw stalactites, flute, capillary stalagmite, column, cave pearl*). The network also shows that certain terms belong to both categories (*blue hole, inflow cave*) as certain forms are both underground and submerged in water or have a hydrological function in karst. In addition, we have noticed that graph-based visualization facilitates the identification and correction of inconsistencies in the manual expert annotation. The final goal is to integrate the visual, graph-based representation into a multimodal knowledge base where frames (Cause, Size, Location, Function etc.) as defined in Vintar et al. (2019) will be presented to the user together with corpus examples, images and geolocations.

4.4. Visualizing Automatically Extracted Knowledge

In this experiment we used sentences where automatically extracted terms co-occurred, and then identified relations between them. The resulting knowledge is shown in Figure 4. The relation extraction was done using ReVerb (Fader et al., 2011), which is a program that au-

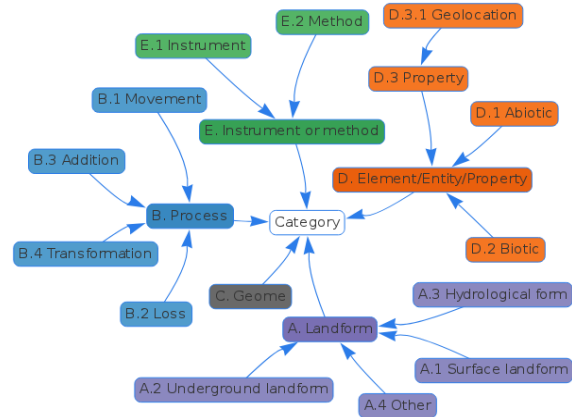


Figure 2: The taxonomy of categories visualized in NetViz.

tomatically identifies and extracts relationships from English sentences, output the triplets in form `<argument1, relation phrase, argument2>`, usually corresponding to subject-verb-object. It is designed for cases where the target relations cannot be specified in advance, which corresponds to the requirements of this experiment with knowledge discovery in mind. The preprocessing includes tokenization, lemmatization and POS tagging. We used the lemmatized forms. We are interested in triplets that include as arguments only terms from the karst domain. The terms were extracted using (Pollak et al., 2012) and were further validated by domain experts.⁵ We also used terms in karstology term list QUIKK⁶. The validated list of domain-specific terms contained 3,149 terms, and triplet arguments extracted with ReVerb were matched against this list. In this way, a huge general triplet network containing less relevant information for domain exploration is reduced and thus made easier for manual inspection. After filtering we retained 302 triplets where arguments exactly match the terms from the list. The most frequent relations include: *be, fill_with, exceed, form_in, associate_with, be_source_of,...*

5. Conclusion and future work

We presented the NetViz terminology visualization tool and two examples of its use for knowledge modeling in the domain of karstology. First, we have demonstrated the visual representation of domain knowledge as extracted from manually annotated definitions. The multi-layer annotations include conceptual categories (Landform, Process, Geome, Element/Entity/Property, Instrument/Method) and their subcategories with which the terms are labelled, and the resulting network can be used by experts, teachers, students or terminologists to explore related groups of concepts, identify knowledge patterns or spot annotation mistakes. Next, we visualized the relations as proposed by the automated term and triplet extraction. This approach is complementary to the manual annotation and may point to previously unknown connections or knowledge structures.

⁵A detailed evaluation of term extraction process is presented in Pollak et al. (2019), ranging from 19.2% for strictly karst terms and 51.6% including broader domain terms and names entities.

⁶<http://islovar.ff.uni-lj.si/karst>

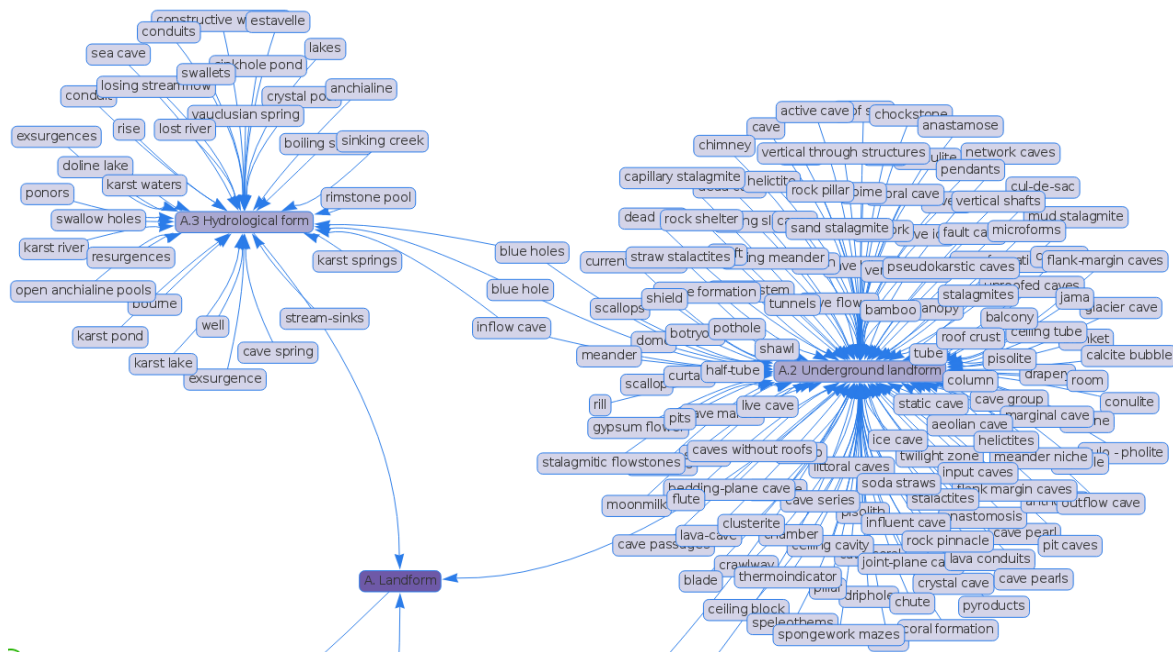


Figure 3: A visualization of a part of categories network which includes hydrological and underground landforms.

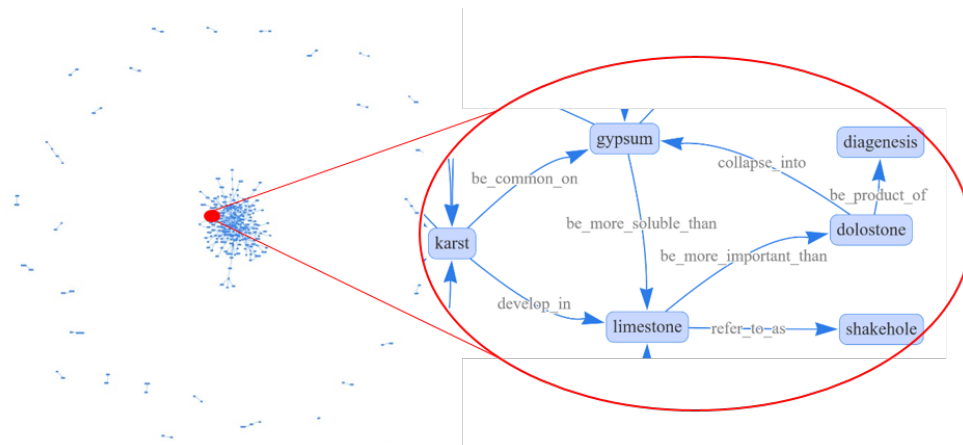


Figure 4: Graph with triplet relations extracted with ReVerb where subject and object match the manually validated list of karst terms.

The simplicity of NetViz allows users to prepare their own input data in the CSV format and create customized visualizations to support their research. For example, in the TermFrame project NetViz is currently used to explore cases where identical or similar concepts have been defined through different hypernyms (e.g. *karst* is a kind of *landscape* / *terrain* / *topography* / *product of processes* / *phenomenon* / *area*).

As future work and the end-result, of the TermFrame project we plan to develop an integrated web-based environment for karst exploration which will combine graphs with textual information, images and geolocations. Since a large number of natural monuments worldwide are in fact karst phenomena, we see the potential of such knowledge

representations not just for science but also for education, environment and tourism.

6. Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103) and the project TermFrame - Terminology and Knowledge Frames across Languages (No. J6-9372). This paper is also supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

7. Bibliographical References

- Ahmad, K., Gillam, L., Tostevin, L., and Group, A. (2000). University of surrey participation in trec 8: Weiridness indexing for logical document extrapolation and retrieval (wilder). 03.
- Amjadian, E., Inkpen, D., Paribakht, T., and Faez, F. (2016). Local-global vectors to improve unigram terminology extraction. In *Proceedings of the 5th International Workshop on Computational Terminology (Comterm2016)*, pages 2–11.
- Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.
- Batagelj, V. and Mrvar, A. (2002). Pajek— analysis and visualization of large networks. In Petra Mutzel, et al., editors, *Graph Drawing*, pages 477–478, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Brewer, N., Gilkey, M., Lillie, S., Hesse, B., and Sheridan, S. (2012). Tables or bar graphs? presenting test results in electronic medical records. *Medical decision making : an international journal of the Society for Medical Decision Making*, 32(4):545–553.
- Bui, Q.-C. (2012). Relation extraction methods for biomedical literature. *Structure*, 01.
- Carvalho, S., Costa, R., and Roche, C. (2017). Ontotermology meets lexicography: the multimodal online dictionary of endometriosis (mode). In *GLOBALEX 2016: Lexicographic Resources for Human Language Technology Workshop at the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 8–15, Portorož, Slovenia.
- Daille, B., Gaussier, E., and Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1, COLING '94*, page 515–521, USA. Association for Computational Linguistics.
- Ellson, J., Gansner, E., Koutsofios, L., North, S., Woodhull, G., Description, S., and Technologies, L. (2001). Graphviz — open source graph drawing tools. In *Lecture Notes in Computer Science*, pages 483–484. Springer-Verlag.
- Erkan, G., Ozgur, A., and Radev, D. (2007). Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '07)*, pages 228–237, 01.
- Faber, P., León-Araúz, P., and Reimerink, A. (2016). EcoLexicon: new features and challenges. In *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, pages 73–80.
- Faber, P. (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin, Boston: De Gruyter Mouton.
- Faber, P. (2015). Frames as a framework for terminology. In Hendrik Kockaert et al., editors, *Handbook of Terminology*, page 14–33. John Benjamins.
- Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, Scotland, UK, July 27–31.
- Fundel-Clemens, K., Küffner, R., and Zimmer, R. (2007). Relex - relation extraction using dependency parse trees. *Bioinformatics (Oxford, England)*, 23:365–71, 03.
- Gil-Berrozpe, J., León-Araúz, P., and Faber, P. (2017). Specifying hyponymy subtypes and knowledge patterns: A corpus-based study. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pages 63–92.
- Grčić-Simeunović, L. and De Santiago, P. (2016). Semantic approach to phraseological patterns in karstology. In T. Margalitadze et al., editors, *Proceedings of the XVII Euralex International Congress*, pages 685–693. Ivane Javakhishvili Tbilisi State University.
- Hughes, L. M., Constantopoulos, P., and Dallas, C. (2015). Digital methods in the humanities: Understanding and describing their use across the disciplines. In J. Unsworth S. Schreibman, R. Siemens, editor, *A New Companion to Digital Humanities*, pages 150–170. John Wiley & Sons.
- Hughes, L. M. (2012). ICT methods and tools in arts and humanities research. In Lorna M. Hughes, editor, *Digital Collections: Use, Value and Impact*, pages 123–134. London, UK: Facet Publishing.
- ISO 704. (2009). ISO 704:2009: Terminology work-principles and methods. Standard, ISO, Switzerland.
- Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289.
- Kim, J.-H., Mitchell, A., Attwood, T. K., and Hilario, M. (2007). Learning to extract relations for protein annotation. *Bioinformatics*, 23(13):i256–i263, 07.
- Lin, Y., Shen, S., Liu, Z., Luan, H., and Sun, M. (2016). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany, August. Association for Computational Linguistics.
- Luo, L., Yang, Z., Cao, M., Wang, L., Zhang, Y., and Lin, H. (2020). A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of Biomedical Informatics*, 103:103384, 02.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings to The International Conference on Learning Representations 2013*.
- Miljkovic, D., Kralj, J., Stepišnik, U., and Pollak, S. (2019). Communities of related terms in Karst terminology co-occurrence network. In *Proceedings of eLex 2019*.
- Pavlopoulos, K., Evelpidou, N., and Vassilopoulos, A.

- (2009). *Mapping Geomorphological Environments*. Berlin Heidelberg:Springer.
- Podpečan, V., Ramšak, v., Gruden, K., Toivonen, H., and Lavrač, N. (2019). Interactive exploration of heterogeneous biological networks with biomine explorer. *Bioinformatics*, 06.
- Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N., and Špela Vintar. (2012). Nlp workflow for on-line definition extraction from English and Slovene text corpora. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 53–60. ÖGAI, September. Main track: oral presentations.
- Pollak, S., Repar, A., Martinc, M., and Podpečan, V. (2019). Karst exploration : extracting terms and definitions from Karst domain corpus. In *Proceedings of eLex 2019*.
- Roche, C., Costa, R., Carvalho, S., and Almeida, B. (2019). Knowledge-based terminological e-dictionaries The EndoTerm and al-Andalus Pottery projects. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 25(2):259–290.
- Song, Q., Watanabe, Y., and Yokota, H. (2011). Relationship extraction methods based on co-occurrence in web pages and files. In *Proceedings of the 13th International Conference on Information Integration and Web-Based Applications and Services, iiWAS '11*, page 82–89, New York, NY, USA. Association for Computing Machinery.
- Sousa, D., Lamurias, A., and Couto, F. M. (2019). Using neural networks for relation extraction from biomedical literature.
- Vintar, Š., Saksida, A., Stepišnik, U., and Vrtovec, K. (2019). Modelling specialized knowledge with conceptual frames: The TermFrame approach to a structured visual domain representation. In *Proceedings of eLex 2019*, pages 305–318.
- Vintar, Š. (2010). Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2):141–158, 12.
- Wang, R., Liu, W., and McDonald, C. (2016). Featureless domain-specific term extraction with minimal labelled data. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 103–112.
- Wermter, J. and Hahn, U. (2005). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 843–850, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.
- Zhang, Z., Gao, J., and Ciravegna, F. (2017). Semre-rank: Incorporating semantic relatedness to improve automatic term extraction using personalized pagerank. *arXiv preprint arXiv:1711.03373*.

Appendix G: Communities of Related Terms in a Karst Terminology Co-occurrence Network

Proceedings of eLex 2019

Communities of Related Terms in a Karst Terminology Co-occurrence Network

Dragana Miljkovic¹, Jan Kralj¹, Uroš Stepišnik², Senja Pollak^{1,3}

¹ Jožef Stefan Institute, Ljubljana, Slovenia

² University of Ljubljana, Ljubljana, Slovenia

³ University of Edinburgh, UK

E-mail: dragana.miljkovic@ijs.si, jan.kralj@ijs.si, uros.stepisnik@gmail.com, senja.pollak@ijs.si

Abstract

Karst science is an attractive field of interdisciplinary research with rich terminology. This study was performed as part of a project aiming at developing novel approaches to terminology extraction and visualization, in line with the understanding of knowledge, as represented in texts, as conceptually dynamic and linguistically varied. The aim of this paper is to investigate how powerful graph-based methods can be used for visualizing and analysing domain terminology. In order to detect communities in karst terminology, we analyse the frequently co-occurring karst terms in a scientific corpus of karstologic literature. The most frequent co-occurrence pairs, which included ten or more co-occurrences within the whole corpus, are delivered as input to the Louvain community detection algorithm and visualized as a domain graph. The resulting data was evaluated by domain experts who found that the detected term groups are meaningful and correspond to different types of karst phenomena. The results are further discussed in relation to more standard topic modelling approaches, using Latent Dirichlet Allocation and Non-negative Matrix Factorization algorithms.

Keywords: karstology; co-occurrence network; community detection algorithm; network visualization; topic modelling

1. Introduction

Karst science, or karstology, is a well-researched discipline with rich terminology, consisting of many expressions referring to regionally specific phenomena. Contemporary research of the topography that is referred to as a 'karst geomorphologic system' or simply 'karst' includes numerous scientific disciplines that study the karst environments worldwide; however, the earliest research on karst primarily regards Classical Karst, which is located in western Slovenia. Consequently, karstologists use many local Slovenian scientific terms and toponyms for typical geomorphological karst structures not only when writing in Slovene, but also in English and other languages. In this paper, we focus on karst texts in English.

This study was undertaken as part of the TermFrame project¹, which is based on contemporary findings in the field of terminology and cognitive linguistics, and aims to

¹ TermFrame project web site: <http://termframe.ff.uni-lj.si/>

develop novel methods that can be utilized in the field of terminology research. The focus of these novel methods is on corpus-based approaches to extraction and visualization of terminological knowledge, including text and graph mining and advanced data representation techniques.

Recent attempts in terminological science understand knowledge, as represented in texts, as conceptually dynamic and linguistically varied (Cabr , 1999; Temmerman, 2000; Kageura, 2002). Research advances in cognition have contributed to the Frame-Based Terminology (Faber, 2012; Faber, et al., 2006), which focuses on representing dynamic knowledge and investigating cultural elements in cognitive structures (Rod r guez Redondo, 2004; Grygiel, 2017), while projects such as EcoLexicon² attempt to visually represent concept networks. While a limited number of studies have used graph-based approaches in the fields of terminology and lexicography (Meyer & Eppinger, 2018; Krek et al., 2017) and for language comparison ( skrlj & Pollak, 2019), we believe that these methods are still to be fully explored, as they present the potential for novel research of specialized knowledge, as well as for new possibilities of knowledge representation that can be inspiring to contemporary lexicography. We believe that the graph-based method for exploring term co-occurrences can contribute to the needs of frame-based terminology, aiming at facilitating user knowledge acquisition through different types of multimodal and contextualized information (Gil-Berrozpe et al., 2017). This type of graph-based tool also has potential for future data representation in the field of e-lexicography (Granger, 2012), where multimodal data and hybridization between different types of language resources (e.g., dictionaries, encyclopaedias, term banks, lexical databases, translation tools) are commonly observed.

The focus of the present work in the scope of the above-mentioned project is to apply graph-based methods to the terminology of karst research. This has motivated us to explore co-occurrences of the specific karstology terms and visualize the results. Another motivation for the visualization of results is that domain experts are often able to interpret information faster when viewing graphs as opposed to tables (Brewer et al., 2012). More generally, as evident by the rising field of digital humanities, digital content, tools, and methods are transforming the entire field of humanities, changing the paradigms of understanding, asking new research questions and creating new knowledge (Hughes et al., 2015; Hughes, 2012). The work complements the results in karst terminology research presented in Vintar et al. (2019), where frame-based annotation of karst definitions is introduced, and in Pollak et al. (2019), where the authors present the results of term, definition, and triplet extraction from karst literature.

This paper is structured as follows: after presenting the background technologies and related work in Section 2, Section 3 introduces our method, which is based on

² <http://ecolexicon.ugr.es/en/index.htm>

community detection of terms extracted from a karstology corpus and their visualization in the form of a network; along with Section 4, the two sections represent the main contribution of the paper. In Section 5, we discuss the results in relation to a more standard topic modelling methods approach, and we conclude this paper in Section 6.

2. Background technologies and related work

This section presents a brief overview of the state-of-the-art of the fields related to our study methods, including co-occurrence and visualization, community detection algorithms and topic modelling.

2.1 Co-occurrence approach and visualization

Scientific literature in different fields can be explored through a search for the co-occurrences of domain-specific terms and their frequencies. A co-occurrence of two terms means that the terms coexist in the text within a certain window. The idea behind detecting co-occurrences of terms is that closely related terms will appear together more frequently. Moreover, co-occurrences can reveal hidden patterns and interesting features in the texts that are being analysed. For example, the co-occurrence analysis might detect spam messages (Krestel & Chen, 2008) or find meaningful knowledge from biological literature in a systematic and automated way (Al-Aamri et al., 2017). Co-occurrence is also used widely in text classification (Figueiredo et al., 2011) and categorization (Luo & Zincir-Heywood, 2004).

There is a difference between first-order and second-order co-occurrence approaches. For the first-order co-occurrence, one would simply count how many occurrences of one token there are within a specified distance of the particular occurrence of another token and build a vector presentation of the results. A second-order co-occurrence vector would represent some aggregation over the token representations, and in the simplest case this is a sum (Maldonado & Emms, 2012).

Representation of co-occurrence pairs in the form of a network is a common way to aid the domain experts with exploration of research results. Such representations can be used for various purposes, such as word sense disambiguation, which represents a challenge in natural language processing field (Duque et al., 2018). Li et al. (2018) report the discovery of new information in the biomedical domain based on the analysis of the structural characteristics of the co-occurrence network. Additionally, co-occurrence networks are increasingly used when analysing users' behaviour on social media (Correia et al., 2016).

In the field of lexicography, co-occurrence networks have been used with the aim of building a new Slovene thesaurus from data available in a comprehensive English–Slovene dictionary (Krek et al., 2017).

2.2 Community detection algorithms

When co-occurrence networks become too large and complex, their visual inspection becomes difficult. One way to explore complex networks more easily is to use community detection algorithms.

Community detection algorithms can be split into several classes based on the underlying idea that guides the algorithms. It must be noted that a strict split between the different methods is impossible, as these methods are not developed in isolation. For example, many methods that are not strictly classified as modularity-based algorithms still use the concept of modularity in one of their steps.

Divisive algorithms are algorithms that find the community structure of a network by iteratively removing edges from the network. The most widely used algorithm among divisive algorithms is the Girvan Newman algorithm (Girvan & Newman, 2002), which removes the network edges with the largest centrality measure. The reasoning behind this is that edges which are more central to a graph are the edges most likely to cross communities. An alternative algorithm is the Radicchi algorithm, which calculates the edge-clustering coefficient of edges in order to determine which edges must be removed. Here, the reasoning is that edges between communities belong to fewer cycles than edges within communities.

Modularity-based algorithms form the majority of community detection algorithms. While, as mentioned above, the concept of modularity (Newman & Girvan, 2004) is used in almost all algorithms to an extent (especially when attempting to determine the best clustering from a hierarchical clustering of nodes), the algorithms in this class use modularity more centrally than other algorithms. The most prominent modularity-based methods are the Louvain algorithm (Blondel et al., 2008) and the Newman greedy algorithm (Newman & Girvan, 2004). Other methods include variations of the greedy algorithm (Wakita & Tsurumi, 2007), simulated annealing (Guimerà & Amaral, 2005), spectral optimization of modularity via a modularity matrix (Newman, 2006a; Newman, 2006b) or via the graph adjacency matrix (White & Smyth, 2005), and deterministic optimization approaches (Duch & Arenas, 2005).

Spectral algorithms find communities in networks by analysing the eigenvectors of matrices derived from the network. The community structure is extracted either from the eigenvectors of the Laplacian matrix of the network (Donetti & Muñoz, 2004) or from the stochastic matrix of the network (Capocci et al., 2005). In both cases, the idea behind the algorithms is that eigenvectors extracted from the network will have similar values on indices that belong to network vertices in the same community. First, a computation of several eigenvectors belonging to the largest eigenvalues is performed. The resulting eigenvectors form a set of coordinates of points, each belonging to one network vertex, with clustering of these points corresponding to community detection of network vertices.

Another important community detection algorithm is the InfoMap algorithm (Rosvall et al., 2009). This is based on the idea of minimal description length of the walks performed by a random walker traversing the network. The communities in InfoMap are determined by constructing so-called codebooks, which are used to describe walks on the network – corresponding to communities in the network, codebooks yield on average shorter average descriptions of walks. Finally, in the most recent rapid development of network embedding algorithms, some researchers have begun using embedding-based methods for network community detection (Li et al., 2018).

2.3 Topic modelling

In this section, we cover topic modelling, i.e. methods used for discovering various topics that appear in a collection of documents. Topic modelling methods are well-established in the field of text modelling, and can be considered as alternative approaches to co-occurrence community detection. Methods for topic modelling can rely on linear algebra, such as Vector Space Model (VSM) (Becker & Kuropka, 2003) or Matrix Factorization (NMF) (Paatero & Tapper, 1994), while others are based upon statistical distributions, for example Latent Dirichlet Allocation (LDA) (Blei et al., 2003). When using both NMF and LDA for topic modelling, two matrices are constructed from the document-term matrix: the document-topic and topic-term matrices. The topics are derived from the contents of the documents, and the topic-document matrix describes data clusters of related documents. LDA usually performs well when it comes to identifying coherent topics, whereas NMF provides incoherent ones (Stevens et al., 2012). While VSM is based on a similar principle as NMF, it has significant limitations when processing long documents as they have poor similarity values. Because the corpus analysed for the purposes of this paper includes both short and long documents (doctoral dissertations, dictionaries, etc.), this specific method was excluded from consideration.

The aim of this paper is to analyse the communities in karst terminology by analysing the co-occurrence network of frequently co-occurring karst terms in the scientific corpus of karst literature. We defined a co-occurrence of terms as their coexistence in the same sentence, while in order to qualify as frequently co-occurring, a term pair had to occur at least ten times over the span of the entire corpus. We decided to start inspecting karst corpus gathered for the purpose of the TermFrame project with basic first-order co-occurrence vectors and present the results of co-occurrence terms in the form of community network, as it is easily comprehended by domain experts. For our research, we used three leading algorithms in the community detection field: Label propagation, Louvain, and InfoMap. The InfoMap and Label propagation algorithms did not yield meaningful results: both identified one large community and several singletons. For this reason, the Methodology, Results, and Discussion sections all focus exclusively on the results obtained using the Louvain algorithm. We also discuss the results from the community detection experiment in relation to two topic modelling approaches, LDA

and NMF, while the exploration of second-order co-occurrence approaches will be explored in future work.

3. Methodology

First, we tokenized and lemmatized our collection of scientific literature and the corresponding term list. Next, first order co-occurrences of pre-specified terms were identified within the corpus. After this, the Louvain community detection algorithm was used to find the communities of co-occurrence pairs. The schematic of the methodology used in this study is shown in Figure 1, with each step further explained below.

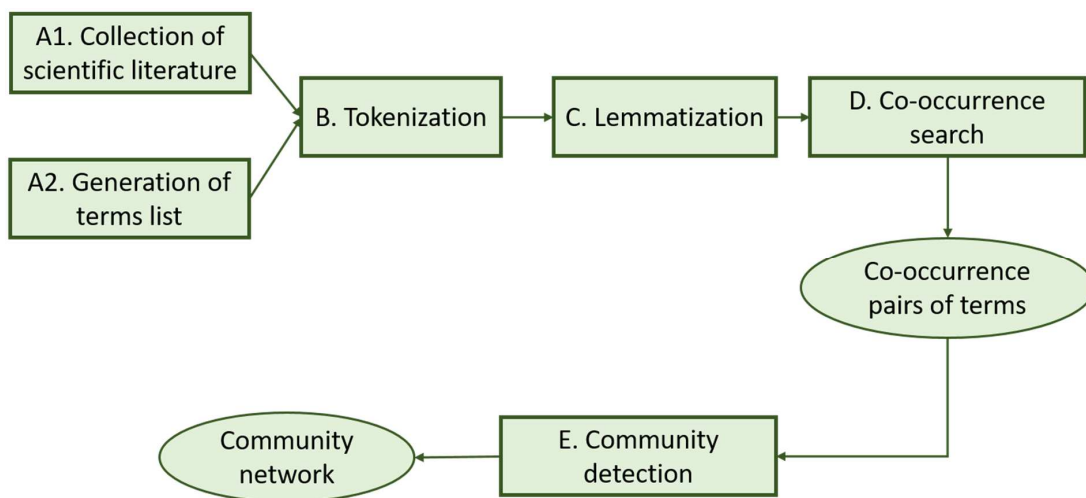


Figure 1: The schematic of the methodology.

A1. Collection of scientific literature represents the compilation of 25 scientific karstology texts, including papers, doctoral dissertations, and the glossary of cave and karst terminology. This corpus was compiled as part of the TermFrame project and is an extended version of earlier work (Vintar & Grčić Simeunović, 2016).³

A2. Generation of terms list was performed as a two-phase process. First, relevant terms were automatically extracted from the TermFrame corpus using the LUIZ-CF term extractor (Pollak et al., 2012), which is a variant of LUIZ (Vintar, 2010) refined with scoring and ranking functions. The terms were validated by the domain expert and were used to compile a term list along with the previously acquired terms from the QUIKK termbase⁴. This process of term extraction and evaluation is presented in more detail in Pollak et al. (2019).

³ We used the corpus version v1.0.

⁴ <http://islovar.ff.uni-lj.si/karst>

B. Tokenization was performed using the NLTK Tokenizer for Python.

C. Lemmatization was performed using the Lemmagen tool (Juršič et al., 2010).

D. Co-occurrence search was performed automatically by the Python script, which stores in a separate file the co-occurring term pairs and the number of their co-occurrences in the whole TermFrame corpus.

E. Community detection was performed using the Louvain algorithm (Blondel et al., 2008), which works by decreasing the modularity of the network, a function that measures the density of links inside communities compared to links between communities. The modularity of a network is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where A_{ij} denotes the weight of the edge between nodes i and j (in our case, the number of co-occurrences), k_i denotes the degree (sum of all adjacent edge weights) of node i , and m denotes the total sum of weights in the network. The term c_i denotes the community to which node i is assigned, meaning the sum above runs over all pairs of i, j where i and j belong to the same community.

4. Results and discussion

For the purposes of this research, we compiled a list of 452 karst terms drawing from a corpus of karstology texts which contained 108,769 sentences in total. Both the list and the literature were tokenized and lemmatized prior to the co-occurrence search, which yielded a list of 10,990 unique co-occurrence pairs using 426 unique lemmatized terms, as well as the data regarding co-occurrence frequency.

The initially obtained co-occurrence pairs would result in a complex network that would be difficult to represent in a comprehensible manner. To simplify the visualization, co-occurrence pairs with frequencies of ten or less were removed from the subsequent analysis. This left us with 1,247 co-occurrence pairs (see Table 1).

	Initial co-occurrence list	Filtered co-occurrence list
Number of co-occurrence pairs	10,990	1,247
Number of unique terms	426	309

Table 1: The summary of the initially obtained co-occurrence list and the filtered version, which contains only the co-occurrence pairs with frequencies of 10 or more.

The 20 most frequent co-occurrence pairs extracted from the karst corpus are listed in Table 2.

ID	Term 1	Term 2	Frequency of appearing	ID	Term 1	Term 2	Frequency of appearing
1	cave	karst	1688	11	limestone	dolomite	368
2	cave	passage	1482	12	cave	karren	349
3	cave	limestone	739	13	solution	karren	319
4	cave	spring	735	14	karren	limestone	311
5	cave	speleothem	664	15	cave	pit	288
6	cave system	cave	597	16	limestone	marble	282
7	cave	gypsum	512	17	karst	spring	270
8	cave	calcite	468	18	karst	term	261
9	karst	limestone	464	19	cave	canyon	261
10	calcite crust	cave	381	20	karst	doline	259

Table 2: The list of common co-occurrence pairs extracted from the karst corpus sorted from most to least frequent.

The filtered co-occurrence pairs served as input for the Louvain algorithm for community detection. Starting with each node in its own community, the algorithm iteratively works in two stages. In the first stage, it searches for the optimum pairs or groups of communities to merge into a larger community and thus increase the modularity of the partition. In the second stage, the algorithm reduces the network to a coarser network based on the discovered communities. The two-stage procedure is then repeated until no increases in modularity can be made. This results in a hierarchy of network node clusters, which can then be cut at any level to produce a clustering of the network nodes. In our case, the algorithm resulted in a three-layer hierarchy. The top level consisted of only two communities and the bottom level of single-node communities. The middle layer was the only layer containing non-trivial information about the structure of the co-occurrence network, and it was therefore subject to further analysis.

The middle layer of the hierarchy, discovered by the Louvain algorithm, consisted of eight communities. Next, we visualized the network using the Barnes-Hut approximation of the force-directed layout to calculate optimal node positions (Jacomy et al., 2014). The discovered communities were then displayed on the network visualization by colouring nodes corresponding to the communities they belong to (see Figure 2).

The karst domain experts analysed the resulting network and found the network visualization particularly interesting, as the communities (listed below) were found to correspond to different types of karst phenomena.

- Community 0: Exokarst landforms ('kamenitza', 'grike', 'stone forest'), which are the result of direct effects of dissolution of bedrock exposed on the surface;
- Community 1: Subsurface landforms, speleogenetic features, and cave environments (e.g. 'passage', 'flowstone deposit', 'cave system'). This community comprises all types of underground voids typical for karst environments regardless of their morphogenesis, including characteristic mechanical and chemical fills within.
- Community 2: Surface karst landforms and environments (e.g. 'uvala', 'doline', 'karst terrain') which are a product of surface and subsurface karst processes, materialising as relief forms or terrain types.
- Community 3: Karst hydrologic processes, environments, and methods (e.g. 'karst recharge', 'groundwater basin', 'tracer test') incorporate all karst aquifer types, the processes within them, and methods concerning their research.
- Community 4: Karst geology representing terms related to karst lithology (e.g. dolomite), minerals ('calcite') and processes affecting them (e.g. 'dissolution')
- Community 5: Includes only two terms (karrenfield, phreatic-cave), which is not enough to define the topic field.
- Community 6 includes only two terms ('turbulent flow', 'laminar flow'), which is not enough to define the topic field.
- Community 7 includes only two terms ('vadose zone', 'phreatic zone'), which is not enough to define the topic field.

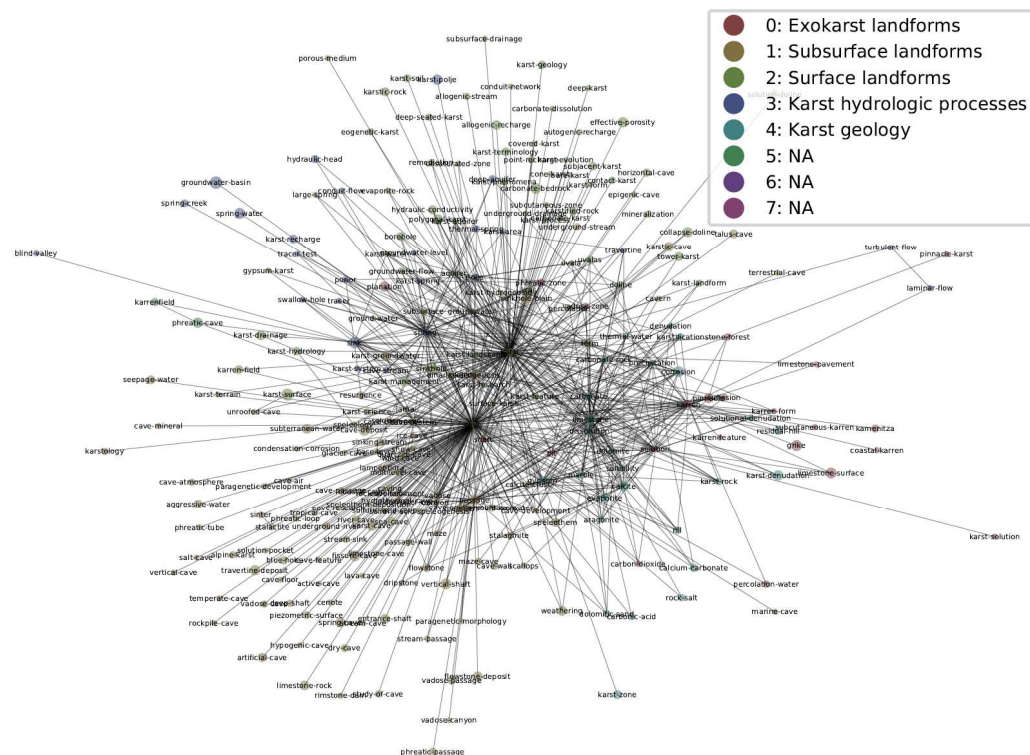


Figure 2: The co-occurrence network, visualized using a force-directed layout, showing the communities discovered within the network. The colours of the nodes correspond to the communities the nodes belong to.

5. Topic modelling experiments

As graph-based modelling is a relatively novel field for harvesting knowledge from specialized corpora, this section discusses our results with respect to more standard topic modelling approaches. For the purpose of this research, we used LDA and NMF algorithms, implemented within a Scikit-learn Python module. The algorithms searched through the complete corpus of 25 documents (described above) containing 108 769 lemmatized sentences, presenting the domain expert with the 25 most important words for each topic. The domain expert subsequently evaluated whether the derived topic words adequately represent specific subfields of karstology. In Table 3, we list the topics and the topic words identified by the NMF and LDA algorithms, which were estimated as meaningful groups by the domain expert. To enable further comparison of results with the community detection experiment, the number of topics was set to eight for both algorithms.

NMF	Topic 0: SPELEOLOGY	cave passage entrance long know study world km large deep map exploration bat sediment mammoth example explore stream important contain river site animal speleothem state
	Topic 1: KARST HYDROLOGY	water flow spring table level aquifer zone high groundwater discharge surface underground stream sea conduit phreatic supply resource fresh mix air rise sink temperature time
	Topic 5: KARST GEOMORPHOLOGY	rock form figure surface limestone large develop small carbonate karren passage process area 10 soil solution high occur dissolution doline lower feature cover sediment deposit
	Topic 6: SPELEOBIOLOGY	species family subterranean know troglobitic habitat include genera number genus group population troglomorphic bat fauna large occur troglobite terrestrial aquatic marine represent small order environment
	Topic 7: GENERAL METHODOLOGY (KARST)	use method data term model technique land date tracer place study time site widely approach human dye analysis test trace map measure determine source work
LDA	Topic 0: SPELEOLOGY	cave sediment passage type channel wall 20 place contain small like 12 width speleothem vertical significant 100 2001 possible figure direction rillenkarren floor stream scale
	Topic 2: KARST GEOLOGY	rock large limestone carbonate cover deposit upper surface gypsum forest dissolution area stone protect calcite earth line layer bed joint various material analysis salt fracture
	Topic 5: KARST HYDROLOGY	water flow spring zone soil deep high aquifer karst surface occur groundwater slope natural condition table value depression low erosion increase result point temperature climate

Table 3: Topic modelling results with Non-negative Matrix Factorization (NMF) and Latent Dirichlet distribution (LDA) applied to karst literature

From a karstologic point of view, the following topics extracted by means of the NMF method describe various aspect of karstology, i.e. different scientific fields regarding karst research:

- Topic 0: Speleology incorporates topic words that are directly referring to cave processes, cave-related landforms, or toponyms regarding to research of caves (i.e. speleology).
- Topic 1: Karst hydrology topic words comprise a variety of terms describing karst aquifers and their study.
- Topic 5: Karst geomorphology topic words correspond to a variety of surface landforms and processes, as well as words labelling their properties.
- Topic 6: Speleobiology topic words are related to cave biota and habitats.
- Topic 7: General karst methodology topic words incorporate a combination of various terms describing research methods from different karst research fields.

LDA identified only three topic groups meaningful to the domain expert, compared to the five identified by NMF:

- Topic 0: Speleology (see NMF Topic 0).
- Topic 2: Karst geology words regarding karst rocks, minerals, and processes concerning them.
- Topic 5: Karst hydrology (see NMF Topic 1).

NMF and community detection experiments have some overlaps in results, such as karst hydrologic processes and karst surface landforms and environments, as well as a partial topic overlap with terms related to speleology.

The results of our proposed community detection methodology have identified several specific topics as evaluated by the expert; however, it can be hard to determine to which extent this is to be attributed to term pre-selection, the community detection algorithm, or to the visualization of results. A detailed study of the role of each component is beyond the scope of this paper, but we believe that graph-based methods coupled with visualization offer great opportunities for investigating terminology as dynamic systems.

An overview of the number of meaningful communities identified by the proposed community detection approach and topic modelling methods (NMF and LDA) is presented in Table 4. All of the topics listed in this paper were manually evaluated by a domain expert. Community detection differs from the topic modelling approaches in that it takes pre-specified terms as input, while topic modelling approaches take as

input all words in the corpus documents. For this reason, a deeper quantitative comparison between these approaches is not feasible.

Number of meaningful topics		
Community detection algorithm	Topic modelling (LDA)	Topic modelling (NMF)
5	3	5

Table 4: Quantitative overview of the discovered topics with topic modelling and graph-based methods.

6. Conclusions and future work

In this work, we used a list of terms extracted from karst scientific literature and then performed a network analysis of karst terminology, wherein the network was constructed from co-occurring karst terms. The community detection algorithms described in this paper grouped specialized terms into semantically related topics, which were also visually presented as coloured nodes in the graphs. In addition, we approached the same corpus from the viewpoint of more standard topic modelling techniques, using LDA and NMF as our main tools.

In future work we plan to include the exploration of second-order co-occurrences, embedding-based topic modelling, and combining graph-based term and community detection methods. In addition, we consider performing a systematic comparison of graph-based community detection and topic modelling approaches, as well as evaluating if term extraction can contribute to these approaches.

Furthermore, we plan to use network representation in the form of triplets {subject, predicate, object}, which can also be a source of identifying novel semantic relations. Within the scope of the TermFrame project, a multi-layer semantic annotation has been performed and the most frequent conceptual frames for specific semantic categories explored. By combining information from manual annotations and the proposed network-based techniques, new knowledge about conceptual frames, semantic relations, and topics could be observed. The potential of graph-based topological analysis lies also in its power to explore structural information, which could reveal potential language and culture-driven differences if, for example, applied to larger comparable corpora of karst texts in different languages.

7. Acknowledgements

This work was financed by Slovenian Research Agency grants J6-9372 (Terminology and Knowledge Frames across Languages - TermFrame) and P2-0103 (Knowledge Technologies). This paper is supported by European Union's Horizon 2020 research

and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this work reflect only the authors' views, and the European Commission is not responsible for any use that may be made of the information it contains.

8. References

- Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M., & Homouz, D. (2017). Constructing Genetic Networks using Biomedical Literature and Rare Event Classification. *Scientific Reports*, 7(1), pp. 2045-2322.
- Becker, J., & Kuropka, D. (2003). Topic-based vector space model. *Proceedings of the 6th International Conference on Business Information Systems*. Colorado Springs, USA.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, pp. 993-1022.
- Blondel, V., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008.
- Brewer, N. T., Gilkey, M. B., Lillie, S. E., Hesse, B. W., & Sheridan, S. L. (2012). Tables or Bar Graphs? Presenting Test Results in Electronic Medical Records. *Medical Decision Making*, 32(4), pp. 545-553.
- Cabré, M. T. (1999). *Terminology: Theory, methods, applications*. Amsterdam; Philadelphia: J. Benjamins Publishing Company.
- Capocci, A., Servedio, V. D., Caldarelli, G., & Colaioni, F. (2005). Detecting communities in large networks. *Physica A: Statistical Mechanics and its Applications*, 352(2), pp. 669-676.
- Correia, R. B., Li, L., & Rocha, L. M. (2016). Monitoring potential drug interactions and reactions via network analysis of Instagram user timeliness. *Pacific Symposium on Biocomputing*. 21. Kohala Coast, Hawaii, USA: World Scientific, pp. 492-503.
- Donetti, L., & Muñoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(10), P10012.
- Duch, J., & Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical review E*, 72(2), pp. 027104.
- Duque, A., Stevenson, M., Martinez-Romo, J., & Araujo, L. (2018). Co-occurrence graphs for word sense disambiguation in the biomedical domain. *Artificial Intelligence in Medicine*, 87, pp. 9-19.
- Eronen, L., & Toivonen, H. (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13(1), pp. 119.
- Faber, P. (ed.). (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin, Boston: De Gruyter Mouton.

- Faber, P., Montero Martínez, S., Castro Prieto, M. R., Senso Ruiz, J., Prieto Velasco, J. A., León Arauz, P. & Vega Expósito, M. (2006). Process Oriented Terminology Management in the Domain of Coastal Engineering. *Terminology*, 12(2), pp. 136.
- Figueiredo, F., Rocha, L., Couto, T., Salles, T., André Gonçalves, M., & Meira Jr, W. (2011). Word co-occurrence features for text classification. *Information Systems*, 36(5), pp. 843-858.
- Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), pp. 7821-7826.
- Grygiel, M. (2017). *Cognitive Approaches To Specialist Languages*. (M. Grygiel, Ed.) Cambridge Scholars Publishing.
- Guimerà, R. & Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028), pp. 895-900.
- Hughes, L. M. (2012). Using ICT methods and tools in arts and humanities research. In L. M. Hughes (ed.) *Digital Collections: Use, Value and Impact*. London, UK: Facet Publishing, pp. 123-134.
- Hughes, L., Constantopoulos, P., & Dallas, C. (2015). Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines. In S. Schreibman, R. Siemens, & J. Unsworth (eds.) *A New Companion to Digital Humanities*. John Wiley & Sons, pp. 150-170.
- Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*, 9(6), e98679.
- Juršič, M., Mozetič, I., Erjavec, T., & Lavrač, N. (2010). LemmaGen: multilingual lemmatisation with induced Ripple-Down rules. *Journal of Universal Computer Science*, 16(9), pp. 1190-1214.
- Kageura, K. (2002). *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Krek, S., Laskowski, C., & Robnik-Šikonja, M. (2017). From Translation Equivalents to Synonyms: Creation of a Slovene Thesaurus Using word co-occurrence Network Analysis. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Leiden, the Netherlands: Lexical Computing CZ s.r.o., Brno, Czech Republic, pp. 93-109.
- Krestel, R., & Chen, L. (2008). Using co-occurrence of tags and resources to identify spammers. *ECML PKDD*. Antwerp: Springer, pp. 38-46.
- Li, T., Bai, J., Yang, X., Liu, Q., & Chen, Y. (2018). Co-Occurrence Network of High-Frequency Words in the Bioinformatics Literature: Structural Characteristics and Evolution. *Applied Sciences*, 8, 1994.
- Li, Y., Sha, C., Huang, X., & Zhang, Y. (2018). Community detection in attributed graphs: an embedding approach. *Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, Louisiana, USA: AAAI .

- Liu, Y., McInnes, B. T., Pedersen, T., Melton-Meaux, G., & Pakhomov, S. V. (2012). Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. *2nd ACM SIGHIT International Health Informatics Symposium*. Miami, Florida, USA: ACM, pp. 363-372.
- Luo, X., & Zincir-Heywood, A. N. (2004). Combining word based and word co-occurrence based sequence analysis for text categorization. *Machine Learning and Cybernetics*. Shanghai: IEEE, pp. 1580-1585.
- Maldonado, A., & Emms, M. (2012). First-order and second-order context representations: geometrical considerations and performance in word-sense disambiguation and discrimination. *JADT 11th International Conference on the Statistical Analysis of Textual Data*. Liege, France, pp. 676-686.
- Meyer, P., & Eppinger, M. (2018). fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts, pp. 1017-1022.
- Newman, M. E. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3), pp. 036104.
- Newman, M. E. (2006b). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), pp. 8577-8582.
- Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2), pp. 026113.
- Paatero, P., & Tapper, U. (1994). Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics*, 5(2), pp. 111-126.
- Pollak, S., Repar, A., Martinc, M., & Podpečan, V. (2019). Karst exploration: extracting terms and definitions from karst domain corpus. In I. Kosem et al. (eds.) *Proceedings of eLex 2019*. Sintra, Portugal, pp. 934-956.
- Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N., & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. *Proceedings of KONVENS 2012*. Vienna, Austria: ÖGAI, pp. 53-60.
- Rodríguez Redondo, A. L. (2004). Aspects of cognitive linguistics and neurolinguistics: conceptual structure and category-specific semantic deficits. *Estudios ingleses de la Universidad Complutense*, 12, pp. 43-62.
- Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation 178.1. *The European Physical Journal Special Topics*, 178(1), pp. 13-23.
- Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 952-961.

- Škrlj, B., & Pollak, S. (2019). Language comparison via network topology. *Proceedings of the 7th International Conference on Statistical Language and Speech Processing*. LNCS 11816. Springer.
- Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive-approach*. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Vintar, Š. (2010). Bilingual Term Recognition Revisited. The Bag-of-Equivalents Term Alignment Approach and Its Evaluation. *Terminology*, 16(2), pp. 141-158.
- Vintar, Š., & Grčić Simeunović, L. (2016). Definition frames as language-dependent models of knowledge transfer. *Fachsprache: internationale Zeitschrift für Fachsprachenforschung, - didaktik und Terminologie*, 39(1-2), pp. 43-58.
- Vintar, Š., Saksida, A., Stepišnik, U., & Vrtovec, K. (2019). Knowledge frames in karstology: the TermFrame approach to extract knowledge structures from definitions. In *Proceedings of eLex 2019*, Sintra, Portugal.
- Wakita, K., & Tsurumi, T. (2007). Finding community structure in mega-scale social networks:[extended abstract]. *16th international conference on World Wide Web*. Banff, Alberta, Canada: ACM, pp. 1275-1276.
- White, S., & Smyth, P. (2005). A spectral clustering approach to finding communities in graph. *SIAM International Conference on Data Mining*. 5. Newport Beach, California, USA: SIAM, pp. 76-84.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Appendix H: An Example of Cross-Language Text Summarization

An Example of Cross-Language Text Summarization

This Appendix details an example of Slovene-to-English Cross-Language Text Summarisation task. This example is part of the MultiLing Pilot 2011 dataset. We selected a topic of this dataset and we automatically translated all source documents from English into Slovenian¹. In this task, the human annotators and our CLTS approach analysed a cluster of 10 documents in Slovenian about the same topic and then generated summaries (up to 250 words) in English. We list below all source documents in Slovenian and the cross-lingual summaries in English of a human annotator and our approach.

Source document 1 90 dni trdega dela za Abu Ghraib vodnika. Petek 2. junij, 2006. Obsojen na prvi od junija vojaški žirija za sodelovanje v zlorab zapornikov v Abu Ghraib, bivši narednik Santos Cardona, 32, a vodnika v vojski Združenih držav, je bil danes obsojen na devetdeset dni trdega dela in znižajo na čin specialist. Cardona je bil obtožen in obsojen na dve stvari: hujši napad, za uporabo belgijski ovčar ogrožajo pripornikov z ukrepi lahko povzročijo smrt ali posebno hude telesne poškodbe, in zanemarjanja dolžnosti v zvezi s tem aktom. Bil je izbil sedem drugih dajatev. Cardona soočajo do tri leta in pol v zaporu na teh obtožb, kljub prepričanju je njegova attourney izrazil olajšanje ob stavka. On ne bo potekala v zaprtem prostoru v času njegovega stavka.

Source document 2 Najdenih med zlorab v zaporu Abu Ghraib in Guantanamo Bay vezi. Četrtek, 28. julij, 2005. V pričanju na včerajšnji vojaški obravnavi zlorab v taborišču ZDA v Iraku, nekdanji upravnik Abu Ghraib, Maj. David Dinenna, je dejal, se je udeležil septembra 2003, je na srečanju z Maj. General Geoffrey D. Miller, ki je bil takrat poveljnik Guantanamo Bay zapora. Dinenna dejal general Miller priporočamo uporabo psov, zaradi svoje učinkovitosti. Dva psa vojaki v Abu Ghraib obtoženi v obravnavi. Santos A. Cardona, 31, in Sgt. Michael J. Smith, 24, so trdili, da so uporabili pse groziti in ustrahovati zaporniki. Med obtoženci pričevanja v torek, so rekli tehnike zasliševanja, ki jih uporabljajo na zapornikih je naučil iz ekipe zasliševalci, ki je bil odposlan v Iraku iz vojaškega oporišča Guantanamo Bay na Kubi. Člen 32 vojaško sodišče postopek v, ki je sklenila sredo v Fort Meade, Maryland, je predhodno zaslišanje slišati tožilstvom in obrambo argumente v zadevi. Tožilstvo želi vojno sodišče z zahtevkom, da toženi stranki delovala kazensko. Obrambna trdi, vojaki so bili po naročil, in da so stroški naj bi padla. Preiskovalni uradnik vojaškega sodišča, Maj. Glenn Simpkins, ima dva tedna časa, da tehtanje dokazov, da je bil predstavljen. Nekatere ali vse stroške, ki bi lahko padla, če pa stojijo nekateri stroški, bo pripravi priporočilo o tem, kako Sgts. Cardona in Smith je treba obravnavati, ko gre za sojenje. Oba obtoženca je dejal na včerajšnji pričevanja, da polkovnik Thomas M. Pappas, top vojaški obveščevalni častnik v Abu Ghraib, odobrila uporabo psov. Pričanje je bilo slišati tudi iz Pvt. Ivan L. »Chip« Frederick, zdaj služi 8-letno zaporno kazen je v Fort Leavenworth po svoji vlogi vodje pri zlorabi, ki je pričala po telefonu iz zapora, ki je bil odobren za uporabo pse, in da je civilna zasliševalec včasih sodeluje pri usmerjanju, ki zapornik celice so, da jih obišče psov. Poleg uporabe psov, so bili tudi del serije zlorab agresivne tehnike zasliševanja, kot so oblačila odstranjevanje in pomanjkanje spanja. James Vincent Lucas je že prej povedal preiskovalcem vojske v Guantanamo, ki je zapustil Kubo leta 2003, da gredo v Irak, kjer je, kot član 6-man team, učil se "lekcije", v Guantanamo, in služil za "zagotoviti smernice", da zasliševalce v Abu Ghraib. Zakonodajca več senatu republikanci sponzorira želi posebej ureja ravnanje z zaporniki v Guantanamo

¹<https://translate.google.com>

in drugih vojaških zaporih. Pomožna sponzor zakona, Lindsey Graham (R-SC), je pred kratkim izdal razgaljenimi interne beležke, ki segajo od leta 2003 in vrhunskih vojaških odvetnikov napisal. So opozorili na Pentagon o agresivne taktike v Guantanamo. V beležke opozoriti, da bi se povečujejo nevarnosti za ameriških vojakov, ki jih je sovražnik ujetih. Army polnjenja listi obtožujejo Cardona in Smith z maltreating pripornikov od 15. novembra 2003 do 15. januarja 2004, ki ga usmerja, spodbuja ali dopušča [njihovo] unmuzzled vojaški delovni pes [i], da lubje in Rezanje na zapornikov, da bi nezakonito Nadlegovanje in ogrožajo zapornikov in da bi zaporniki uriniranju ali blato na sebi. Cardona, za Fullerton, z 42. vojaške policije odreda v Ft. Bragg, N.C., je bil obtožen devetih šteje. Smith, Fort Lauderdale, Fla., Z 523rd vojaške policije odreda v Ft. Riley, Kan., Je bil obtožen 14 grofov.

Source document 3 US Army na vojaškem sodišču Abu Ghraib častnika. Sobota 27 januar, 2007. Steven Lee Jordan se bo sodilo vojaško sodišče za njegove vpletenosti v zlorabo zapor primeru Abu Ghraib je US Army napovedal včeraj. Jordan bodo sodili na osmih šteje. Stroški vključujejo krutost, trpinčenju zapornikov, neupoštevanja nadrejeni častnik, namerno zanemarjanje dolžnosti in dajanje lažnih izjav. Jordan je samo v ZDA častnik zaračuna v primeru -eleven Abu Ghraib vojaki so bili že obsojeni do 10 let zaporu. 50-letni Jordan, A rezervist iz Virginije, vodil zaslišanje center v Abu Ghraib, zloglasnem zaporu samo zunaj Bagdada. Pod njegovim poveljstvom, vojaki fotografirali gole zapornike v ponižujoče pozah. Pooblaščenec sam ni bil prisoten v kateri koli od slik. Jordan se sooča z najvišjo kazen 22 let za osem stroškov skupaj.

Source document 4 Human Rights Watch: ZDA zlorabe iraških zapornikov nadaljuje. Nedelja, 23 julij, 2006. Poročilo je objavljeno s Human Rights Watch o ravnanju z zaporniki v Iraku s strani ameriških vojakov po zaporu škandal Abu Ghraib. Trdi, mučenja in drugih zlorab proti zapornikov v ameriškem priporu v Iraku nadaljuje in je pooblaščen in rutinsko. Poročilo vsebuje podrobne račune zlorabe iz centrov za pridržanje v celotnem Iraku, in obtožbe iz izpraševalnika vojske nameščene na Camp Nama na mednarodnem letališču v Bagdadu. Poročilo o 55-stran z naslovom Ne v krvi, Vojaki Računi pripornika zlorabe v Iraku zahtevkov, ki so bili, da so ostre tehnike zasliševanja odobrila poveljnikov. Vojaki opisujejo, kako so zaporniki rutinsko podvrženi hudim pretepanju, bolečih položajih stres, hudo pomanjkanje spanja in izpostavljenost ekstremnem mrazu in vročini. Računi prihajajo iz intervjujev s Human Rights Watch, ki se izvajajo, dopolnjujejo jih memorandumov in zaprisežene izjave iz odpravi tajnost dokumentov. Marc Garlasco, Human Rights Watch vojaški analitik, je dejal, do zdaj, veliko obtožb in dokazov je bila plava okoli o dovoljenju do verigi poveljevanja neprimernih tehnik, poslanci na sojenju poskušala dvigniti to v svojo obrambo in dobil nikjer. Prvič, smo bili pod pogojem, jasne informacije o obsegu odobritve neprimernih tehnik, in opozarja neposredno častnikom in Pentagon. John Sifton, raziskovalec pri Human Rights Watch je dejal Vojaki so povedali, da so Ženevske konvencije ne uporabljajo, in da zasliševalce lahko uporabite zlorabe tehnike, da bi dobili pripornike in govori. Ministrstvo za obrambo zanika kakršno koli Pentagon soglasje za kakršno koli zlorabo. Polkovnik Mark Ballesteros, tiskovni predstavnik Pentagona pravi DoD politika je in vedno bo humano ravnanje s priporniki v njegovem priporu. Rekel je, da je naloga, sila v Iraku, ki nadzoruje pripornika operacije in je dosegla ducat pregled pripornika politik. Nobena od pregledov ugotovilo, da je ministrstvo za obrambo doslej odredil ali dopušča pripornika zlorabe. poročilo Human Rights Watch priporoča imenovanju Dvostranački komisijo, da razišče obseg priprt zlorab v Iraku, prenovi vojaškega pravosodnega sistema, in imenovanju neodvisnega tožilca za preiskovanje in pregon odgovornih. Marc Garlasco dejal V sedanjih razmerah dolgih poklicev v Iraku in Afganistanu z vrtenjem čet, ni nobenega razloga, neodvisen tožilec ne moremo se ukvarjajo z zlorabo obtožb, bi bilo narediti razliko, če je E-3 (zasebni prvi razred) v polju žage nekateri višji uradniki ali zastave, ki so pooblašteni zlorabo preizkušajo na sodiščih-borilne namesto napredovali.

Source document 5 ZDA zapušča Abu Ghraib. Petek 10 marec, 2006. ZDA častniki izjavili, da se bo v Abu Ghraib zaprt v nekaj mesecih, in zapornikom preselil v drugih zaporih in taboriščih v Iraku. Nekatere od 4500 zapornikov bodo poslani novi Camp Cropper, ki je vgrajen v ta poseben namen. Delovanje Abu Ghraib se bo prenesla na iraške vlade. Podpolkovnik Keir-Kevin Curry je dejal: so niz ni natančni datumi, vendar je načrt, da bi to dosegli v naslednjih dveh do treh mesecih. Polkovnik Barry Johnson, vojaški predstavnik, pravi Jasno theres veliko čustev, pritrjeno na Abu Ghraib. Za Iračane, da sega veliko dlje od zlorab, ki jih ameriški vojaki zagrešili, in ga naredi to magnet za napade. Kompleks zapor je bila zelo bali center za mučenje in izvedbo pod Sadamom Huseinom, in toliko kot 4000 ljudi je bilo usmrčenih v zaporu leta 1984. Stavbe v Abu Ghraib so prvotno zgradili britanskih izvajalcev v 1960. Slike obsežne zlorabe pripornikov v Abu Ghraib, sprejeti znotraj zapora novembra 2003 je bilo v letošnjem letu je izšla v začetku. V januarju 2005 je sodišče ZDA vojska borilna postopek našel Army SPC. Charles Graner kriv za zlorabo zapornikov v zaporu Abu Ghraib. Žirija obsojen GRANER do deset let zapora. Skupno devet ameriških vojakov so bili spoznani za krive v obtožbe, ki izhajajo iz Abu Ghraib zlorabe zapor škandal.

Source document 6 Nove fotografije zlorabe Abu Ghraib. Sreda 15 februar, 2006. Avstralska televizijska postaja je pokazala nove fotografije domnevno prikazuje več zlorab zapornikov v zaporu Abu Ghraib. Dateline, trenutni program, zadeve oddaja na SBS javnega omrežja, je trdil, slike so prikazane bili primeri razponu od zlorabe izvaja v zaporu Abu Ghraib v Iraku in vključiti prej nevidni material, kot so ubijanje, mučenje in spolno ponižanje. Program je prav tako trdil, da so zaporniki v Abu Ghraib ubit, ko Ameriški vojaki zmanjkalo gumijastih nabojev, kot so poskušali obvladovati zapor izgrediv, ter začeli uporabljati v živo krogih. Izvršni producent Mike Carey je dejal Dateline je pridobila datoteko, ki vsebuje na stotine slik, nekatere znane in drugih, ki kažejo nove zlorabe. Vendar pa se je znižala na reči, kako se slike je prišel v njihovo posest. Nekatere slike se zdi, da pokažejo US vojak Charles Graner, ki je imel vodilno vlogo v prejšnje Abu Ghraib zlorabe škandal. Veliko novih slik so bolj grafika kot predhodno objavljeno - te nove fotografije vključujejo zaporniki opravljajo seksu deluje in ranjene / mrtvih zapornikov. SBS je trdil, da fotografije trupel je bilo ljudi, ki so umrli v Abu Ghraib med zasliševanjem. The Guardian navaja neimenovani obrambe uradnika v ZDA pravi, da je vojska pregledal slike, ki so jih objavili SBS in potrdil, da so bili med tistimi, ki so predmet zakona zahtevo prostem dostopu do informacij, ki jih je ACLU. Ameriški državljsanske svoboščine unije je bil omogočen dostop do fotografij v lasti vojske, ki ni bila javno objavljena z ameriškem zveznem sodišču v septembru, vendar US vlada pritožila na odločitev. Pentagon predstavnica Bryan Whitman je poudaril, da je v ZDA politika obravnavati vse zapornike humano. Dodal je še dejal, da so bili zlorab v Abu Ghraib temeljito preiskati, in da ko je prišlo do zlorab, je ta služba delovala na njih takoj, jih temeljito raziskati, in kjer je to primerno preganja posameznike. Slike so pogosto prikazane na televiziji po vsem Bližnjem vzhodu, tudi v Iraku. Saleh al-Humaidi, Jemna novinar, je povedal Reuters To je resnično ameriški grdega, da nobena druga država na svetu tekmujejo z ... Američani bi morali opravičiti, da človeštvo za njihove vlade ležijo na svetu, ki se borijo za svobodo in da je prišel v Irak, da ga shranite iz Sadama Huseins zatiranja. US obrambni minister Donald Rumsfeld je pričal, da niso bile vse znane fotografije zlorab v Abu Ghraib javno objavilo na preiskavo senat Armed Services odbora v maju 2004. Gospod Rumsfeld je dejal, onkraj zlorabe zapornikov, obstajajo tudi druge fotografije, ki prikazujejo primere fizičnega nasilja proti zapornikom, akti, ki se lahko opiše le kot očitno sadistični, kruto in nečloveško.

Source document 7 US vlada ukazala, da javnost več slik, povezanih z primeru Abu Ghraib. Sreda, oktober 5, 2005. Alvin K. Hellerstein, v Združenih državah zvezni sodnik je odredil izpustitev dodatnih fotografij in video posnetkov, ki se nanašajo na Abu Ghraib iraški zlorabi zapor primeru v teku. Mediji, ki so sestavljeni iz 74 slikami in 3 video posnetkov, je bilo odrejeno za javnost z dne 29. septembra, do 20 dni možnost za

pritožbo. Ameriški državljanske svoboščine unije vložil svobodi informiranja iz Zahteva v gibanju za sprostitev fotografij. Busheva administracija je trdila, da bi naročeno mediji povzročijo terorističnim organizacijam v škodi več ameriških vojakov in državljanov v tujini. US Okrožni sodnik Alvin K. Hellerstein močno nasprotovali, češ da zatirati slike znaša predložiti izsiljevanja. V izjavi, Hellerstein dejal: Naš narod [Združene države Amerike] ne preda do izsiljevanja in strah izsiljevanja ni pravno zadostni argument, da nam preprečujejo opravljanje obvezne ukaz.

Source document 8 Graner spoznan za krivega zlorabami v Abu Ghraib. Sobota, 15. januar, 2005. TEXAS, US - United States Army vojaško sodišče ugotovljeni vojske Spc. Charles Graner kriv za zlorabo zapornikov v Iraqs Abu Ghraib. Žirija obsojen GRANER do deset let zapor. Graner, 36, je bil opisan kot vodje v času svojega štiridnevnega sojenju pred vojaškim žirije. Bil je obtožen napada na zapornike za zabavo. On se sklicuje nedolžen do pet obtožb proti njemu, toda deset oseba Žirija je pet ur, da ga spoznali za krivega. Med sojenjem so video in fotografije, sprejeti znotraj zaporu novembra 2003 predstavljen na sodišču. Fotografije so bile objavljene v začetku leta 2004 prinaša svetove pozornost zlorab dogaja v zaporu. Domneva se, da visoki obrambni ameriški uradniki vedeli za zlorabe, vključno sekretar za obrambo Donald Rumsfeld. Garners obramba trdila, da je bil v skladu z nalogi za blažijo do zapornikov pred zaslišanjem.

Source document 9 Ameriški General uveljavlja pravica zoper samoobtožbo v primeru Abu Ghraib. Četrtek 12 januar, 2006. General Geoffrey Miller, vrh ZDA poveljnika, ki nadzoruje pridržanje in zasliševanje pripornikov v Guantanamo zalivu in Abu Ghraib objektov zavrnil pričanje na sodišču-borilna postopka, s sklicevanjem na pravico, da sam, poroča Washington Post, ne implicirajo. To je verjel, da je prvič, da se je vloga visokih uradnikov v zlorabe ujetnik škandal formalno gladini. General Miller zavrnil pričanje v sodni-borilna sojenja, ki vključuje dva psa-viličarji, ki so obtoženi, ki je priprt zlorabe. Odvetnik enega od obdolženca želi vprašanje Miller o tem, ali je odredil uporabo psov prestrašijo zapornikov med zaslišanji. General Millers odvetnik je izjavil, da je bila odločitev, da padec pričanjem sprejeti, saj je bil Miller večkrat intervjuvali v zadnjih nekaj letih, in da ostaja pri svojih prejšnjih izjav na kongresu, armade preiskovalci in odvetniki.

Source document 10 Pristnost novih Abu Ghraib fotografije potrjena. Četrtek, 16. februar, 2006. Avstralski je danes poročal, da je Pentagon je potrdil pristnost slike avstralskih televizijskih omrežju SBS sporedu. V nasprotju trditve nekaterih medijev, SBS Datelines izvršni producent Mike Carey je trdil, v četrtek, da so programi Raziskovalci so ugotovili primere zlorab v Abu Ghraib zaporniki, ki niso bila obravnavana s strani organov ZDA. nove slike, ki so na voljo na SBS je opisal kot kvantni preskok v primerjavi z letom prej oddajajo slike. Carey je napovedal, da naslednji teden bo program prikazal več zlorab slik, vendar je ugotovil, da je nekaj slik featuring zapornike spolno ponižujočih dejanj ne bo oddaja, saj so bili šteje tudi grafično. Naslednji tedni program bo vključeval tudi intervjuje z ameriškega vojaka, ki je bil obsojen zaradi zlorabe zapornikov in drugega nekdanjega vojaka, ki je bil priča zločine. Washington Post je opozoriti, kot ga SBS je predlagal, da je več mediji so tudi sami ni objavljeno veliko število fotografij, ki jih imajo v posesti. Washington Post nadalje pojasnjuje, da so časopisi, ki so v posesti te slike omejena, v veliki meri, ki ga je čisto grafično narave njih, zlasti golote. Washington Post navaja tudi SBSs možnost zaobiti Združenih državah vlade bolj prizadevala, da bo Abu Ghraib slike iz oči javnosti, kot je še en razlog, zakaj slike niso v drugih medijih prvič objavljena. ZDA internet novicami Salon je danes objavil nekaj svojih prej zadržane slike in potrdila, da imajo datotek in drugih elektronskih dokumentov iz notranje preiskave vojske v Abu Ghraib ujetnik zlorab škandal, ki vključuje tudi prvotno objavljene slike, kot tudi kot tiste, ki jih SBS objavljeni. Salon ugotavlja, da nekatere datoteke iz poveljstva kriminalistične nanašajo na agentov CIA,

da zasliševali zapornike v Abu Ghraib, ampak da niso bili uradniki Cie preganja kljub smrti vsaj enega iraškega med zaslišanjem Cie tam, kar potrjuje trditev, ki jo SBS, da so nekateri izmed fotografij dokumentira prej preganjanih zlorabe. Ameriška vlada je izrazila zaskrbljenost zaradi novih abuse pictures, ki se objavi. John Bellinger zunanjega ministrstva je povedal BBC, Menili smo, da je šlo za vdor v zasebnost pripornikov samih, da imajo te fotografije pridejo ven ... (in da je objava lahko tudi), ventilator plameni po vsem svetu in ker lahko še naprej nasilje.

Summary (human annotator) Pictures of extensive detainee abuse in Abu Ghraib were made public in early 2004. A year later, a United States Army court martial sentenced Army Spc. Graner to ten years in prison. A total of eleven American soldiers were found guilty in allegations stemming from the Abu Ghraib prison abuse scandal. Among them, two dog handler soldiers. During their testimony, they said they learned the interrogation techniques that included the use of dogs from a team of interrogators that was dispatched to Iraq from Guantanamo military base. On September 2005, a U.S. federal judge ordered the release of additional photos and video relating to the case. The Bush administration claimed that the ordered media could provoke terrorist attacks. Five months later, the Australian TV station SBS showed new photos claiming that some of them document previously unprosecuted abuse. According to a report published on June 2006 by Human Rights Watch, torture and other abuses against detainees in US custody in Iraq continued and were authorized. The Department of Defense denied any Pentagon approval for any abuse. On January 2007, the U.S. Army announced that U.S. Lt. Col. Steven Lee Jordan was going to be tried by a military court. Jordan is the only U.S. officer charged in the Abu Ghraib case. A year earlier, a top US commander, who supervised the detention and interrogation of detainees at Guantanamo Bay and Abu Ghraib facilities, had declined to testify in a court-martial proceeding, by invoking his right to not implicate himself.

Summary (our CLTS approach) Convicted on the first of June by a military jury for participating in prisoner abuse at Abu Ghraib, former sergeant Santos Cardona, 32, a dog handler in the United States Army, was today sentenced to ninety days of hard labour and demoted to the rank of specialist. The Guardian cites an unnamed US defense official saying that the Army had reviewed the pictures posted by SBS and confirmed that they were among those that are subject to the Freedom of Information Act request made by the ACLU. Army charge sheets accuse Cardona and Smith with maltreating detainees from November 15, 2003, to January 15, 2004 by directing, encouraging, or permitting [their] unmuzzled military working dog[s] to bark and growl at detainees in order to unlawfully harass and threaten the detainees and in order to make the detainees urinate or defecate on themselves. A report is published by Human Rights Watch on treatment of prisoners in Iraq by US soldiers after the Abu Ghraib prison scandal. SBS alleged the photographs of the dead bodies were of people who had died at Abu Ghraib during interrogation. United States army officers stated that the Abu Ghraib prison will be closed within months, and its prisoners moved to other prisons and camps in Iraq. The two accused said in yesterday's testimony that Col. Thomas M. Pappas, the top military intelligence officer at Abu Ghraib, approved the use of the dogs. Charles Graner guilty of abusing prisoners at the Abu Ghraib prison.