A Workbench for Corpus Linguistic Discourse Analysis

Julia Krasselt
Zurich University of Applied Sciences, Switzerland

Matthias Fluor
Zurich University of Applied Sciences, Switzerland

Klaus Rothenhäusler
Zurich University of Applied Sciences, Switzerland

Philipp Dreesen
Zurich University of Applied Sciences, Switzerland

Abstract

In this paper, we introduce the *Swiss-AL workbench*, an online tool for corpus linguistic discourse analysis. The workbench enables the analysis of Swiss-AL, a multilingual Swiss web corpus with sources from media, politics, industry, science, and civil society. The workbench differs from other corpus analysis tools in three characteristics: (1) easy access and tidy interface, (2) focus on visualizations, and (3) wide range of analysis options, ranging from classic corpus linguistic analysis (e.g., collocation analysis) to more recent NLP approaches (topic modeling and word embeddings). It is designed for researchers of various disciplines, practitioners, and students.

2012 ACM Subject Classification Computing methodologies → Language resources; Computing methodologies → Discourse, dialogue and pragmatics

Keywords and phrases corpus analysis software, discourse analysis, data visualization

Digital Object Identifier 10.4230/OASIcs.LDK.2021.26

Supplementary Material *InteractiveResource (Online Tool)*: `https://swiss-al.linguistik.zhaw.ch/shiny/dashboard/`

Funding This work was supported by internal funding of the Zurich University of Applied Sciences.

1 Introduction

Linguistic corpora are highly dependent on tools that enable a systematic analysis of primary data, annotations and metadata. Corpora are always approached with the need for specific information, e.g., regarding the frequency of a word form over time or a word's embeddedness in a linguistic context. This dependency relation is reinforced by the variety of research fields that use corpora, such as discourse analysis, lexicography, or language acquisition. An in-depth technical and statistical knowledge of processing annotated language data (e.g., by means of a programming language such as Python or R) is not necessarily part of the core competencies of these research fields. The same holds true for the translation of quantitatively obtained corpus data into diagrammatic representations (e.g., bar and line graphs or networks). Thus, researchers working with corpus data need to rely on appropriate analysis tools.

Furthermore, corpora are not only an invaluable resource in linguistic research, but also in other academic disciplines and in the field of professional communication. From an applied perspective, a good and easy to understand corpus analysis tool is needed because corpus data is approached from an "outsiders" (non-linguistic) perspective, e.g., by professionals developing a communication strategy for a company.

Here, we present the *Swiss-AL workbench*, an online tool for analysing Swiss-AL, a multilingual web corpus for Applied Linguistics [17]. The motivation for the development of the workbench arose from the need for a user-friendly, intuitively accessible, state-of-the-art analysis tool for different user groups. As a consequence, the workbench is characterized by the following key aspects: (1) it is easily accessible with any online browser without prior registration; (2) it has a strong focus on visualizing results in a diagrammatic way; (3) it offers not only traditional corpus linguistic methods (e.g., collocation analysis), but also more recent approaches from natural language processing (topic modeling and word embeddings).

The Swiss-AL workbench is designed for the purpose of applied discourse linguistics, but it can also be used in other research fields. Applied discourse analysis is concerned with identifying the communicative conditions that shape the way a society talks and writes about specific topics [28]. These conditions appear as patterns of language use, i.e. recurring ways of talking or writing about something [7]. As an applied discipline, discourse linguistics pursues the goal of solving communicative problems. The Swiss-AL workbench enables both the corpus-based and corpus-driven identification of patterns of language use by providing different means to analyse the available corpora (e.g., the statistical co-occurrence of words or the distribution of ngrams).

The workbench is designed for a rather heterogeneous audience in order to overcome the difficulties and desiderata outlined in the first two paragraphs. The intended audience ranges from discourse and corpus linguists (who typically have very specific questions, e.g., regarding variants of a specific word or regarding the frequency distribution of a word over a certain period of time) to researchers from other disciplines (who do not normally have linguistic expertise), students and actors of professional communication.

The paper is structured as follows: Section 2 gives a brief overview of related work on corpus analysis software; Section 3 describes the intended audience of the workbench and typical use case scenarios. Section 4 describes the workbench, its architecture and underlying data and individual functionalities in detail. Section 5 contains a conclusion and plans for future work.

2 Related Work

The first digital tools for analysing corpora date back until the 1970s and have since then developed from merely providing concordances for a given search word to web-based or standalone software allowing for quantitative and qualitative analysis of ever-growing corpora (for a historical overview, see [21]).

Modern corpus analysis software can be categorized in (1) ready-made corpus analysis tools, i.e. tools already equipped with corpora and a set of functionalities to analyse these corpora, (2) corpus analysis software designed for the import of own corpora and (3) software allowing for both approaches. Table 1 gives an overview over existing corpus analysis tools and a comparision with the Swiss-AL workbench.

Regarding (1) and with a focus on German, the Institut für Deutsche Sprache and the Berlin-Brandenburgische Akademie der Wissenschaften provide online tools to the reference corpora DeReKo and DWDS [3, 18, 10]. Similar to the Swiss-AL workbench, these tools provide only limited access to the full texts in the corpus due to copyright reasons. [29] introduce the cOWIDplus Viewer, allowing to analyze the vocabulary of German online media during the COVID-19 pandemic on a regularly updated data base. Similar to the Swiss-AL workbench, the cOWIDplus Viewer has a focus on visualizing results and is aimed at non-linguistic experts. For English, the BYU corpus analysis tools offer access to a broad variety of corpora.

Table 1 Comparison of corpus analysis tools, with a focus on available methods and intended audience (some of the tools offer additional methods, not all can be mentioned here for pragmatical reasons).

		concordance	keywords	collocations	tm[5]	we[6]	ngrams	frequency lists	distribution analysis	text view	intended audience[1]
pre-installed corpora	DWDS[8] [3]	✓		✓				✓[3]	✓		scientific and non-scientific audience
	Cosmas II[8] [18]	✓		✓				✓	✓		researchers, translators, students, linguistic laypeople
	english-corpora.org[8]	✓	✓	✓			✓	✓	✓	✓	researchers, students
	Swiss-AL workbench[8]	✓	✓	✓[2]	✓	✓	✓	✓	✓	(✓)[4]	researchers, students, practitioners, linguistic laypeople
import of own corpora	AntConc[7] [1]	✓	✓	✓			✓	✓		✓	students
	CorpusExplorer[7] [25]	✓	✓	✓			✓	✓	✓	✓	corpus linguists, data mining experts
	CQPweb[8] [15]	✓	✓	✓				✓	✓	✓	non-technical users
	WMatrix[7] [24]	✓	✓	✓			✓	✓		✓	academic researchers and students
	Wordsmith[7] [26]	✓	✓	✓			✓	✓	✓	✓	lexicographers, researchers, students
pre-installed corpora/import of own corpora	LancsBox[7] [6]	✓	✓	✓[2]			✓	✓		✓[3]	anyone interested in language
	Sketch Engine[8] [16]	✓	✓	✓			✓	✓	✓	✓	linguists, lexicographers, translators, students, teachers

1) According to self-description in publications/on website
2) including collocation networks
3) only for specific corpora
4) on request (due to copyright restrictions)
5) topic models
6) word embeddings
7) desktop application (installed locally)
8) server based application

With regard to (2), the *Corpus Workbench* (CWB) and its webserver based GUI *CQPweb* is one of the most flexible tools for indexing and analysing own corpora [14, 15]. The Swiss-AL workbench heavily relies on the CWB architecture (see Section 4.1). Other freely available corpus tools are *AntConc* [1], *Wordsmith* [26] and *CorpusExplorer* [25].

One of the leading corpus tools that enables the import of own corpora but that is also equipped with a large variety of corpora in multiple languages is *Sketch Engine* [16]. It is a proprietary software and became a standard tool especially in lexicography. Another proprietary corpus analysis software is *WMatrix* [24], accessible online via Lancaster University. Recently, *LancsBox* has been published by the University of Lancaster as a standalone software package [6]. It is designed for importing own corpora but is also equipped with a range of preinstalled corpora. Similar to the Swiss-AL workbench, it also has a strong focus on visualizing results.

The Swiss-AL workbench presented here belongs to the first group of software tools, since it enables access to a variety of corpora from the Swiss-AL corpus family. While a wide range of tools for corpus linguistic analysis exists, the Swiss-AL workbench fills a noticeable gap: it is easily accessible (without a user account or a prior installation of software), targets at a very heterogeneous audience and offers a wide range of analysis methods.

3 Intended Audience and Use Case Scenarios

The intended audience includes three main groups. (1) Corpus linguistic laypeople use the workbench especially as project partners in discourse-related research. Typically, as practitioners of professional communication, they have very specific questions, e.g., regarding the frequency distribution of a word referring to their organization. Often for the first time, the workbench provides these practitioners with access to data from the discourse that affects them and enables a much wider perspective, i.e. extrospection [11]. It is planned to offer the workbench also for actors from the civil society like NGOs and citizen science initiatives. The workbench enables practitioners to change their perspective from introspection to extrospection.

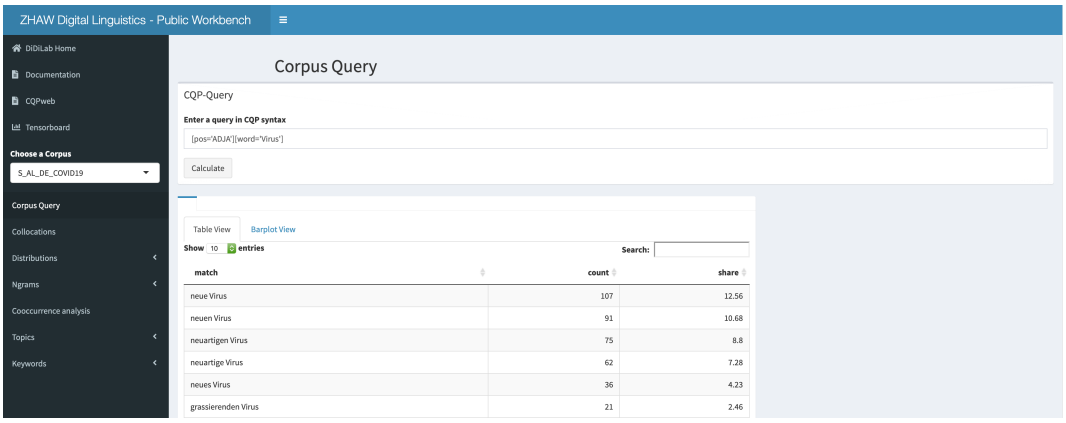

Figure 1 Workbench interface.

(2) After an introduction, undergraduate to PhD students of Applied Linguistics can use the workbench to learn the variety of corpus linguistics analysis almost independently. As such, the workbench can be used for exercises, seminar papers, bachelor, and master theses. (3) The group of corpus data experts uses the workbench especially in inter- and transdisciplinary research projects. The workbench makes it possible to quickly explain and show how corpus linguistic analysis works and how results can be visualized.

The main advantage of the workbench is the macro perspective (distant reading) on discourses, including visualizations. The workbench offers the possibility to aggregate discourses in form of distributions, e.g., of frequency and co-occurrence data (see Section 4.3). As the intended audience of the workbench is so heterogeneous, the workbench has a tidy surface (cf. Figure 1) and can be used without prior registration. The available functions are non-nested. Instead, they are available from the main interface in order for users to always be well oriented.

A typical use case scenario could proceed in three steps (of course, depending on the competencies of the user group). For example, if a user wants to analyze the communication about pandemic measures in the German and Italian COVID-19 discourse in Switzerland, a first step would be to use the word embedding model to find semantically similar words referring to pandemic measures. Alternatively, topic modeling can be used for a first overview to get hints on discoursive thematicity of known measures. As a second step, frequency and distribution over time can be analyzed for the words identified in the word embedding model. Finally, the most interesting/frequent words can be used for collocation, co-occurrence, and ngram analysis.

4 Workbench Description

The workbench is available under the following URL: `https://swiss-al.linguistik.zhaw.ch/shiny/dashboard/`. Figure 1 shows the general layout: on the left navigation pane, users can choose a corpus from a drop-down menu. The workbench provides various functions which will be performed for a selected corpus. The results will be displayed in the right window pane. For each function, different visualization options are available, e.g., a tabular view or a graph view. Additionally, the workbench is equipped with a documentation giving an overview over the available corpora and implemented corpus linguistic functions.

4.1 Workbench Architecture

The main workbench is built on top of R Shiny by RStudio [9]. R Shiny allows to create a visually pleasing web app which triggers R code on the fly and allows for adjustment and manipulation of parameters. This principle of separating the code from the visuals allows us to create a workbench that is easy to use for laypersons and linguists alike. The visualisations and queries are done in real time. The majority of the corpus-related functions are processed with the help of the polmineR package [4] which uses the underlying Corpus Workbench (CWB, [14]) for accessing the corpus data. For a more detailed description of the implemented functionalities, please see Section 4.3.

At its core, the so called shiny app triggers R functions, which in turn retrieve and process data and send them back to be rendered on the website. The data can be manipulated in various ways in order to create a useful plot or table for further investigation by the app users. In terms of manipulation, shiny can be used to give the user a choice of parameters to take into account. E.g., it is possible to offer the user a simple slider to limit the year in which the texts in a corpus were created. This allows for a better investigation and data exploration. Further, due to the usage of the polmineR package, the power of the CWB syntax can be used to create a detailed analysis of the underlying data.

All corpora available on the workbench belong to the family of corpora subsumed under the label *Swiss-AL* ([17], compare Section 4.2). The texts in these corpora are crawled from a predefined set of web pages and annotated linguistically by an automated pipeline. Since Swiss-AL mainly contains texts that are subject to Swiss copyright restrictions, the workbench currently does not offer access to the full texts in the corpora.

Since the workbench is currently in an early stage of development and due to the copyright restrictions of the underlying corpus data, the code is not open source.

4.2 Available Corpora

The workbench is equipped with a variety of corpora from the Swiss-AL family of corpora [17]. The corpora are web-based, i.e. texts are crawled from a curated list of websites from politics, media, industry, science and civil society. All corpora are processed with a linguistic pipeline (described in detail in [17]). Due to the multilingualism in Switzerland, most corpora are available in German, French, and Italian. The workbench also serves as a tool to make research data publicly available in order to follow an open research data policy. E.g., we recently published a corpus on Swiss COVID-19 discourses.

4.3 Functionalities

The workbench provides access to standard linguistic methods for discourse analysis (corpus query, distribution analysis, collocations/co-occurrences, keywords, ngrams, cf. [2, 7]) and also to approaches that have become relevant for the analysis of public discourses more recently, coming from natural language processing (topic modeling, word embeddings).

Corpus Query

This mode of analysis allows to query for a word or a sequence of words in a selected corpus by using CQP-syntax [14] and to get the frequency for this query. Strings can be specific (combinations of) word forms, lemmas, part-of-speech or even dependency relations, depending on the token level annotations available in the selected corpus (so called positional attributes). CQP-syntax allows for the use of regular expressions. To that end, users can

search for strings matching a specific pattern (see example 1). Frequencies will be reported for all matches of a query (e.g., the second search string in the example below will report all individual frequencies for words beginning with the morpheme {Vir-}).

Example 1 corpus queries using CQP-syntax.

```
[pos = "ADJA"][lemma = "Virus"] # sequence of adjective plus 'Virus'
[word = "Vir.*"] # word forms beginning with {Vir-}
[depRel = "SB" & pos = "NN"] # nouns in subject position
```

Distribution Analysis

By entering up to five word forms, users can analyse the relative frequency of these words in a user-defined time period and/or a user-defined set of sources. For example, users can get the frequencies per month for the word forms *Lockdown* and *Shutdown* in Swiss media since January 2020 in order to see wether there is a preference for one of these words and wether these preferences change over time. Results will be visualized as a line graph or barplot.

ngrams

Since a considerable amount of language consists of conventionalized chunks of words (cf. [12]), an analysis above the level of single words is an important tool in discourse analysis ([7]). By using the ngrams function, a user can calculate sequences of up to four words. The user needs to define the length of the ngram and a word and/or a part-of-speech tag that needs to be part of the ngram. E.g., a user can search for 4grams containing the word form *wir* ("we") and compare sources from media and politics, in order to identify similarities and differences in the use of the pronoun. The results will be displayed as a table or visualized as a bar chart.

Context Sensitive Analysis

In discourse analysis, but also in other fields like lexicography and language learning, context sensitive methods are crucial for analyzing the semantics of a word by its co-occurrence with other linguistic units. The workbench allows for two context sensitive modes of analysis, which differ in the size of context that is taken into account: collocation analysis and co-occurrence analysis.

- Collocation Analysis: By entering a specific word or phrase, the workbench will calculate words (so called *collocates*, cf. [13]) that occur significantly often within the immediate context of the given search string. The size of the context window can be adjusted individually, ranging from one to ten words to the right and left of the given search word, respectively. Log Likelihood is used as a measure of statistical association. Collocates can be either displayed as a table, as a bar chart or as a treemap. Collocations are especially useful to analyse the meaning of words in a given discourse, since meaning is mainly constructed by a word's immediate context.
- Co-occurrence Analysis: In contrast to the previous function, users can also identify words that correlate with a given word on a *textual level*. We use the term co-occurrence analysis to distinguish this approach from the classical window approach described for collocations. Pearson correlation is used as a statistical measure. Co-occurring words (i.e. words often appearing in the same text) can be either visualized in a bar chart or in a

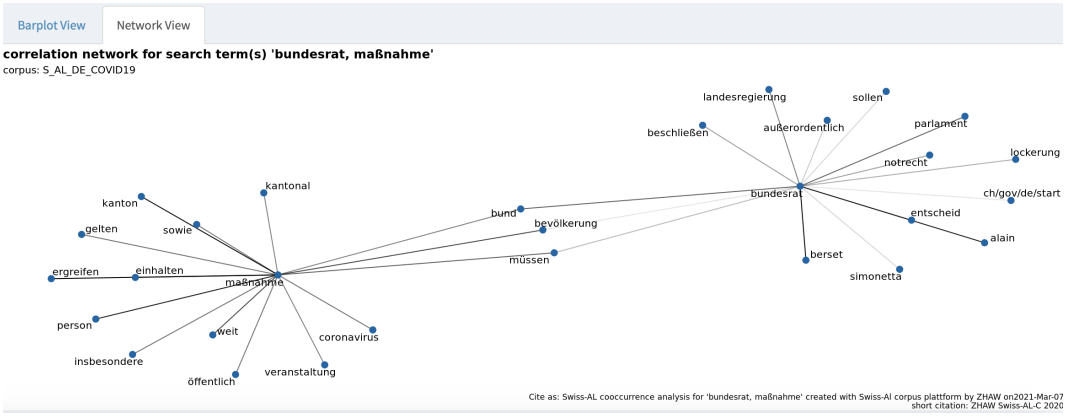

Figure 2 Co-occurrence analysis: network view. For the words *Maßnahme* ("measure") and *Bundesrat* ("Federal Council") the fifteen most correlating words on a textual level (i.e. co-occurrences) are visualized as a network. The two words share three co-occurrences (*Bund* "federation", *Bevölkerung* "population", and *müssen* "need to"), indicating a discoursive association between both words.

network. A network visualization is especially useful when co-occuring words of more than one given word should be displayed to reveal associations within a discourse (cf. Figure 2).

Keyword Analysis

Keyword analysis is one of the most established methods in corpus linguistic discourse analysis since it identifies typical vocabulary for specific discourses (or sub-discourses). The workbench allows (1) the comparison of specific years for the whole corpus (e.g., by comparing the vocabulary of 2019 with that of 2020) or (2) the comparison of specific actors for specific years (e.g., by comparing the vocabulary of a newspaper for 2019 with the same newspaper's vocabulary for 2020).

Topic Modeling

For all corpora on the workbench, separate topic models are available which can be used to get an overview over the thematic structure of the corpus. The models are precalculated by using an LDA algorithm with a prior removal of stopwords [5].[1]. Users can choose between a tabular view (showing the top 25 word of each topic) and an interactive, web-based visualization (LDAvis, [27]) to get an overview over all topics in the model and their distribution in the corpus. Furthermore, the development of topics over time can be visualized as line graphs, in order to see wether a topic is especially prominent at certain points in time.

Word Embeddings

Semantic vector space models [19] have recently become of interest in domains outside NLP. E.g., [8] shows the potential of word embeddings for the data-driven reconstruction of narrations in texts and for the analysis of public discourse. The workbench provides access to a variety of word embedding models based on the *word2vec* algorithm introduced by [22]. Models can be visualized with TensorBoard[2].

[1] Topic models were precalculated with the R wrapper for the machine learning software *Mallet* [23, 20].
[2] `https://www.tensorflow.org/`

In discourse analysis, word embeddings are especially useful for identifying semantically related words which refer to an overarching concept. E.g., users interested in the discoursive construction of fear in COVID-19 discourses could start by identifying words semantically related to *Bedrohung* ("threat") (a word from which we know that it is related to the concept of fear). The word embedding model for the Covid-19 corpus would on the one hand reveal expectable next neighbors like *Angst* ("fear") and *Panik* ("panic"), but also words that one might not think of initially but that are connected with the concept of fear in Covid-19 discourse (e.g., *Trauma* ("trauma")).

5 Conclusion

We introduced the Swiss-AL workbench as a tool for discourse analysis with a strong focus on the visualization of aggregated data and the combination of traditional corpus methods and recently developed machine and deep learning methods. The workbench is designed for a rather heterogeneous audience, i.e. researchers, practitioners and students. As such, it complements existing tools for corpus linguistic analysis. Consequently, the possibility of importing other corpora is not implemented at the moment since this would require corpus and computer linguistic expertise on the part of the user (e.g., preparing an annotated XML version of the corpus or precalculating a topic model). This scenario does not match with the expertise and needs of the intended audience of the workbench.

Next steps include the further development of the individual modes of analysis (e.g., by providing different statistical measures for keyword and collocation analysis) and the presentation of use cases and exemplary discourse analyses in order to give a user an even better understanding of how to apply corpus data to discourse analytical questions.

References

1 L. Anthony. Antconc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005.*, pages 729–737, 2005. doi:10.1109/IPCC.2005.1494244.

2 Paul Baker. *Using Corpora in Discourse Analysis*. Continuum, London, New York, 2006.

3 Berlin-Brandenburgischen Akademie der Wissenschaften. DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart.

4 Andreas Blaette. *polmineR: Verbs and Nouns for Corpus Analysis*, 2020. R package version 0.8.2. doi:10.5281/zenodo.4042093.

5 David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77, 2012. doi:10.1145/2133806.2133826.

6 Vaclav Brezina, P. Weill-Tessier, and A. McEnery. *#LancsBox*, 2020. v. 5.x.

7 Noah Bubenhofer. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Number 4 in Sprache und Wissen. De Gruyter, Berlin, New York, 2009.

8 Noah Bubenhofer, Selena Calleri, and Philipp Dreesen. Politisierung in rechtspopulistischen Medien: Wortschatzanalyse und Word Embeddings. *Osnabrücker Beiträge zur Sprachtheorie (OBST)*, 95:211–241, 2019.

9 Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. *shiny: Web Application Framework for R*, 2021. R package version 1.6.0. URL: https://CRAN.R-project.org/package=shiny.

10 Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. KorAP architecture – diving in the Deep Sea of Corpus Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3586–3591, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL: https://www.aclweb.org/anthology/L16-1569.

11 Philipp Dreesen and Julia Krasselt. Exploring and analyzing linguistic environments. In François Cooren and Peter Stücheli-Herlach, editors, *Handbook of Management Communication*, number 16 in Handbooks of Applied Linguistics. De Gruyter, Berlin, Bostom, to appear. doi:10.1515/9781501508059-021.

12 Britt Erman and Beatrice Warren. The idiom principle and the open choice principle. *Text*, 20(1):29–62, 2000.

13 Stefan Evert. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 1212–1248. De Gruyter, Berlin, 2008.

14 Stefan Evert and Andrew Hardie. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, 2011.

15 Andrew Hardie. CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics*, 17(3):380–409, 2012. doi:10.1075/ijcl.17.3.04har.

16 Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. The Sketch Engine: Ten years on. *Lexicography*, 1(1):7–36, 2014. doi:10.1007/s40607-014-0009-9.

17 Julia Krasselt, Philipp Dreesen, Matthias Fluor, Cerstin Mahlow, Klaus Rothenhäusler, and Maren Runte. Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4138–4144, Marseille, France, 2020.

18 Leibniz-Institut für Deutsche Sprache. COSMAS I/II (Corpus Search, Management and Analysis System). URL: https://cosmas2.ids-mannheim.de/.

19 Alessandro Lenci. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1):151–171, 2018. doi:10.1146/annurev-linguistics-030514-125254.

20 Andrew Kachites McCallum. *MALLET: A Machine Learning for Language Toolkit*, 2002. URL: http://mallet.cs.umass.edu.

21 Tony McEnery and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, New York, 2012.

22 Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, 2013. Association for Computational Linguistics.

23 David Mimno. *mallet: A wrapper around the Java machine learning tool MALLET*, 2013. R package version 1.0. URL: https://CRAN.R-project.org/package=mallet.

24 Paul Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549, 2008. doi:10.1075/ijcl.13.4.06ray.

25 Jan Oliver Rüdiger. CorpusExplorer, 2018. URL: http://corpusexplorer.de.

26 Mike Scott. Developing wordsmith. *International Journal of English Studies*, 8(1):95–106, 2008.

27 Carson Sievert and Kenneth Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics. doi:10.3115/v1/W14-3110.

28 Jürgen Spitzmüller and Ingo H. Warnke. *Diskurslinguistik. Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse*. De Gruyter, Berlin, Boston, 2011.

29 Sascha Wolfer, Alexander Koplenig, Frank Michaelis, and Carolin Müller-Spitzer. *cOWIDplus Viewer*, 2020. URL: https://www.owid.de/plus/cowidplusviewer2020/.