Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4145–4151
Marseille, 11–16 May 2020

Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics

Julia Krasselt, Philipp Dreesen, Matthias Fluor, Cerstin Mahlow, Klaus Rothenhäusler, Maren Runte
ZHAW Zurich University of Applied Sciences, School of Applied Linguistics
Theaterstrasse 17, 8400 Winterthur
{krss, dree, fluo, maho, rotk, runm}@zhaw.ch

Abstract

The Swiss Web Corpus for Applied Linguistics (Swiss-AL) is a multilingual (German, French, Italian) collection of texts from selected web sources. Unlike most other web corpora it is not intended for NLP purposes, but rather designed to support data-based and data-driven research on societal and political discourses in Switzerland. It currently contains 8 million texts (approx. 1.55 billion tokens), including news and specialist publications, governmental opinions, and parliamentary records, web sites of political parties, companies, and universities, statements from industry associations and NGOs, etc. A flexible processing pipeline using state-of-the-art components allows researchers in applied linguistics to create tailor-made subcorpora for studying discourse in a wide range of domains. So far, Swiss-AL has been used successfully in research on Swiss public discourses on energy and on antibiotic resistance.

Keywords: web corpus, applied linguistics, annotation, discourse analysis

1. Introduction

Applied Linguistics, such as discourse analysis, relies on corpora. Corpus-based methods, e.g., collocation or keyword analysis, have become standard tools of analysis (Bubenhofer, 2009; Baker, 2006; Gabrielatos and Baker, 2008). Until recently, there was no Swiss corpus that met the requirements of applied discourse analysis. In this paper, we introduce Swiss-AL, a multilingual Swiss web corpus for Applied Linguistics, that fills this research gap.

With Swiss-AL we address four major challenges: a) multilingual discourse (with the major national Swiss languages German, French, and Italian); b) the representation of the Swiss federal structure in the selection of corpus sources; c) following the perspective of applied discourse analysis, it is necessary to maintain a flexible pipeline and corpus structure that can take into account not only mass media but also, for example, texts from political parties, selected business associations, and NGOs; d) in the context of the increasing range of machine and deep learning algorithms (e.g., topic modeling and neural word embeddings), it is important to have a linguistically annotated corpus at hand that allows for basic corpus linguistic methods as well as for the adaption of such innovative NLP methods.

In the following sections, we first present the current state of the art regarding web corpora (Section 2.) and then introduce Swiss-AL and our corpus linguistic processing pipeline (Section 3.) in more detail. In Section 4., we present research based on Swiss-AL as a corpus linguistic resource. In Section 5., we describe future plans for extending the processing pipeline.

2. Related Work on Web Corpora

Today, the web is used as a linguistic resource quite naturally. There are several very large and multilingual corpus collections which gather their texts from the web, e.g., WaCky Wide Web (Baroni et al., 2009), the TenTen Corpus family (Jakubíček et al., 2013), COrpora from the Web (Schäfer and Bildhauer, 2012), or the C4-Corpora (Habernal et al., 2016). These corpora are rather large and aim to contain a representative if not balanced part of contemporary language use tailored to be used for general linguistic research or as resource for the evaluation and development for NLP tools (Ferraresi et al., 2008).

Usually, web corpora are built via crawling all available web pages which later will be cleaned from non-text elements: First, lists of seed URLs are identified which serve as a starting point for the crawling algorithm itself. The most widespread crawling algorithm is the breadth-first crawling strategy, which recursively follows all links on a website. Schäfer (2016) claims that the algorithm is biased, e.g., towards websites to which many other websites contain links. He proposed a random walk algorithm, ClaraX, which is used for the COW-corpora. De Groc (2011) introduced a focused, i.e., topic specific crawler, working with user-provided seed URLs and a lexicon of relevant search terms extracted from the web pages. A different approach for crawling web pages for a web corpus is based on using search engines, e.g., the BootCaT toolkit described by Baroni and Bernardini (2004).

Following the crawling approach to build corpora from the web for purposes of (applied) linguistics requires several post-processing steps: removal of HTML tags and boilerplate (like copyright statements), excluding pages with texts not meeting a minimum length, etc. As Fletcher (2004) noted, texts gathered from the web contain a ”sheer amount” of noise which makes this processing unavoidable. To that end, web corpus projects like the ones mentioned above developed processing methods to circumvent these problems, as described by Schäfer (2017).

However, starting with a rather random selection of seed URLs and following links without assessing their relevance results in large amounts of web pages which later have to be evaluated and will probably disregarded. Although this can be done automatically during post-processing, it is costly and time-consuming.

Issues related to copyright are a significant problem when building web crawled corpora which should be used and shared widely by interested linguists. There are various approaches to overcome theses issues. Habernal et al. (2016) built a corpus only containing texts published under a CreativeCommons (CC) license by including a license identification algorithm in the processing pipeline. Lyding et al. (2014) built a web crawled corpus of contemporary Italian by only using websites that publish texts under a CC licence (like Wikipedia). Similarly, Barbaresi and Würzner (2014) manually identified blogs on computer mediated communication that were published under a CC license and built a corpus from them. Otherwise, only consulting of corpora would be allowed, but not the general distribution. With Swiss-AL we present a corpus that acknowledges the copy right situation in Switzerland, primarily by providing aggregated analysis data to the end user (cf. 3.4.).

3. Swiss-AL

The following sections describe the general characteristics and the processing pipeline of Swiss-AL, a multilingual Swiss Web Corpus for Applied Linguistics. In contrast to other large web corpus projects (see Section 2.), Swiss-AL is not designed as a large scale corpus for NLP purposes, but as a corpus for data-based and data-driven linguistic analyses of specific discourses (understood in a Foucauldian sense as the virtual sum of all written and oral utterances on a specific topic (Spitzmüller and Warnke, 2011)).

Swiss-AL has been generated by a bottom-up data collection: Swiss-AL consists primarily of sources with selected topics that are relevant to our project partners (see Section 4.). The general aim is to use Swiss-AL to model discourses in which agents such as officials or NGOs position themselves through communication on the Web.

The corpus has several annotation layers beyond the usual ones like POS tagging and lemmatization. For example, it contains named entities (e.g., toponyms, organizations) and syntactic dependencies (used for example to find agents and actions associated with them). Since applied discourse linguistics is mainly interested in semantics and pragmatics, it is of great importance to have annotations at hand that go beyond a purely grammatical level.

3.1. Database

Swiss-AL is built via web crawling and the subsequent extraction of texts from the downloaded websites using XPATH expressions (so called scraping). Unlike other web-based corpora that depend on a list of seed URLs (e.g., the WaCky corpora, see Section 2.), Swiss-AL consists of a curated list of domains (e.g., www.admin.ch), which can be extended at any time depending on the research topic. This is made possible by a flexible processing pipeline, that easily allows the processing of varying data formats (such as HTML, XML, and json). The crawler follows all internal links and thus gathers the whole website. Currently, Swiss-AL mainly consists of texts from Swiss domains (top-level domain .ch, see Section 3.2.). It also includes a sample of Twitter Data, user forum discussions, and a small number of news websites from Germany. The domains are manually categorized as either politics, industry, science, mass media or social media.

The corpus represents multilingualism in Switzerland: it contains texts in the three official Swiss languages German, French, and Italian, because the web pages from the top level domain .ch are multilingual per se. Additionally, the three large language regions of Switzerland were taken into account when initially sampling the domains. Currently, also media texts in Rhaeto-Romanic, spoken in the smallest language region of Switzerland, are acquired. With the latest release in June 2019, Swiss-AL contains 1.55 billion tokens in 8 million texts (see Tab. 3.1.). Since web pages in Switzerland are often offered in multiple languages, the corpus contains parallel texts. This is especially the case with websites from Swiss authorities, which consistently offer information at least in German, French, and Italian. Currently, parallel texts are not marked as such but will be in the near future.

3.2. Data Sampling

The sources contained in Swiss-AL are the result of careful sampling processes which took place and are still ongoing in specific research projects (cf. Section 4.).[1] All projects belong to the discipline of applied linguistics, because they analyze and model specific discourses together with agents from practice (like federal offices or business companies). In consideration of Switzerland's multilingual, federal, and direct-democratic structure, the sampling process is not only a practical research challenge, but also a political one.

The initial sampling was done for the project "Energy Discourses in Switzerland (2016–2019)" (Stücheli-Herlach et al., 2018).[2] The aim of the project was the corpus-driven and corpus-based analysis of language use in the public discourse about energy policies in Switzerland. The sampling criteria were carefully constructed in order to adequately model publicly accessible Swiss discourses on energy. To that end, linguistic, geographic, thematic, and situative criteria were taken into account. All sampled sources represent agents from politics (e.g., the Swiss government and federal offices, cantons, and political parties), industry (e.g., consulting companies, NGOs, and business partners of public authorities), science (e.g., universities) and mass media (e.g., daily and weekly newspapers and specialist newspapers).

Due to the fact that Swiss-AL is built by crawling web domains (see Section 3.3.), all accessible subpages are crawled and the texts contained on them enter into the corpus when certain criteria are met (e.g., minimum length). So, despite the fact that the initial sampling of sources was done for the purpose of analyzing the Swiss energy discourse, Swiss-AL is rather unspecific with regard to topics. The corpus contains a majority of sources potentially relevant for various discourse linguistic research questions (e.g., federal web pages, newspapers, political parties). Ev-

[1] A complete list of sources can be found at the following website: `https://swiss-al.linguistik.zhaw.ch/docs/ord/`

[2] The project was funded by the Swiss Federal Office of Energy within the Energy-Economy-Society research program.

	(sub)corpus	tokens	texts
.ch	Swiss-AL-DE-CHE	700,789,708	1,299,510
	Swiss-AL-FR-CHE	342,728,211	612,286
	Swiss-AL-IT-CHE	150,688,050	386,935
.de	Swiss-AL-DE-DEU	215,253,374	707,600
social-media	Swiss-AL-twitter	128,714,764	4,882,407
	Swiss-AL-forums	9,813,273	105,186
	Swiss-AL (total)	**1,547,987,380**	**7,993,924**

Table 1: Size of Swiss-AL subcorpora in tokens and texts (June 2019). Cf. footnote 3 for naming conventions.

idence for this is offered by topic models trained for all CHE-corpora (see Section 4.).[3]

Swiss-AL is not conceptualized as a reference corpus for contemporary use of German, French, and Italian in Switzerland, since no considerations about balancing the corpus for genre, register, or text length were made. Instead, the corpus is compiled from the perspective of applied linguistics, which considers research questions relevant for agents from practice, such as: How do others write about my organization? Who mentions my organization in publications? Which sub-topics dominate the discussions? Are there differences in attitudes towards political processes between German-, French-, and Italian-speaking Switzerland?

Since the initial sampling process in 2016, new sources are added to the corpus depending on research projects that work with the corpus (see Section 4.).

3.3. Pipeline and Linguistic Processing

The data from the web is crawled through a python-based polite crawler, which saves all HTML files locally to ensure future compatibility, expansion, and adaptation of the pipeline. If initially identified as relevant, the whole website is being crawled and every link followed to create the highest possible yield of usable data (by respecting robots.txt files if available). This especially becomes important once news websites hide their content after the news is not “new” anymore, therefore making it harder to access. In such cases, we observe that those now “old” articles often reappear as suggested reading which through premature boilerplate removal would be out of sight, but by following those links it is possible to retrieve this data as well. The data extraction happens via customized scrapers which are carefully adjusted to each website individually. This is done to circumvent potential problems in boilerplate removal and heuristics and to ensure the best possible precision on extracting the content of the website properly. This means we can separate the content into distinct parts for further analysis and extract the proper metadata (e.g., author, date, and source).

Before feeding the scraped texts into our processing pipeline we perform near duplicate detection for which we use an implementation of the SpotSigs algorithm (Theobald et al., 2008). While being weakly linguistic grounded through antecedent chains anchored at function words it can easily be adapted to new languages by manually specifying a short list of language specific antecedents. At runtime the algorithm is parameterized by a single value specifying the maximum document similarity in terms of the Jaccard similarity, which we experimentally fixed at 0.85. The algorithm identifies sets of documents that are considered near duplicates of one another. We randomly choose one document from these sets and discard the rest before running them through the automatic linguistic processing.

We built our processing pipeline on top of the UIMA framework (Ferrucci and Lally, 2004) which allows for flexible configurations of processing modules. We maintain a number of basic linguistic processors for all three languages currently contained in the corpus, these include:

1. Part of speech tagger and lemmatizer: As a default we use the TreeTagger (Schmid, 1994) but we included OpenNLP[4] and Mate tools (Björkelund et al., 2010) as alternatives.

2. Named entity recognizer: We rely on Stanford NER (Finkel et al., 2005) and include the trained model from Tint[5] (Palmero Aprosio and Moretti, 2016) for Italian. The identified location mentions are disambiguated using the geo resolution system CLAVIN[6].

3. Dependency parsing is not among the standard components run over all the corpus data. But Stanford Dependency Parser (Manning et al., 2014) as well as MaltParser (Nivre et al., 2006) and the dependency parser included in the Mate tools are available for pars-

[3] Subcorpora of Swiss-AL are named with ISO-3166-1 country codes: Two-letter country codes indicate the language of the texts contained in the subcorpus (e.g., DE for German); three-letter-codes indicate the top-level domain, from which the texts have been taken (e.g., CHE for .ch).

[4] `https://opennlp.apache.org/`

[5] `https://github.com/dhfbk/tint`

[6] `https://clavin.bericotechnologies.com`

ing sub-corpora. For Italian, we use a trained model from Tint for the Stanford parser.

Additionally, we have a number of specialized components which are only available for German, either because there are currently no equivalent resources for the other languages (Named Entity Linking for which we adapted code from AmbiverseNLU[7]) or because they are designed to deal with linguistic features peculiar to or more needed when processing German texts such as topological parsing (meaning the disambiguation of automatically recognized toponymes; in-house development) or morphological analysis (RFTagger (Schmid and Laws, 2008)).

Our pipeline might be run with a language preset but it can also be configured to automatically determine the language of texts and recognize the switch of one language to another within a document which is not uncommon in the Swiss context. The annotated corpus is stored in the Corpus Workbench (Evert and Hardie, 2011) to afford detailed manual linguistic queries and analysis. In addition, the processing results are also fed into an ElasticSearch[8] index to offer an alternative programmatic interface.

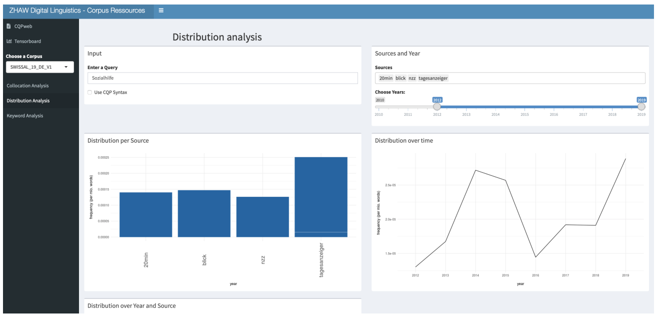
3.4. Access & Rights

As a special feature, Swiss-AL enables non-linguistic end users (i.e., project partners, the general public) to access easy-to-understand data analyses. The corpus is accessible for basic corpus linguistic analysis under the domain `https://swiss-al.linguistik.zhaw.ch`. Users can calculate collocations and distributions (over time and sources) for specific words or phrases they are interested in. Furthermore, the platform offers the functionality of calculating keywords by comparing sub-corpora. Users have access to a corpus documentation, offering an overview over the different sources that feed into Swiss-AL. Users can transparently choose the sources which they would like to include in their analysis.

The corpus linguistic analysis is based on R-Shiny-Apps (Chang et al., 2019), whereby the underlying data is provided through the Corpus Workbench, a standard infrastructure when working with large corpora (Evert and Hardie, 2011). Communication between R/R-Shiny and the Corpus Workbench is established via the R-package PolmineR (Blätte, 2016), which allows the interactive analysis of corpora indexed in the Corpus Workbench in R. As a consequence, CQP syntax can be used on the Swiss-AL platform to search for words and phrases. The platform offers different visualizations of the search results (word clouds, barplots, line graphs, see Figure 1).

Furthermore, the platform offers neural word embedding models for various sub-corpora of Swiss-AL (such as the Swiss-AL-CHE corpora for German, French and Italian). The models are calculated with the word2vec-algorithm (Mikolov et al., 2013) and are currently used for semantic and pragmatic analysis (cf. Section 4.).

Swiss-AL mainly contains texts that are subject to copyright protection according to Swiss law. Exceptions are official documents published by the Swiss government and public authorities. As a consequence, the texts contained in Swiss-AL and their linguistics annotations are not itself provided on the Swiss-AL platform in order not to infringe copyright regulations. There is currently one exception to that: the platform offers access to the complete records of the debates of the Swiss National Council, the Swiss Council of States, and the United Federal Assembly since they are not subject to Swiss copyright regulations.



Choose a Corpus: SWISSAL_19_DE_V1

Enter a Query: Sozialhilfe

☐ Use CQP Syntax

Sources: 20min blick nzz tagesanzeiger

Choose Years: 2012 – 2019

Figure 1: Screenshot from the Swiss-AL online search platform

4. Application

Swiss-AL is conceptualized as a corpus for applied linguistics with a strong focus on discourse linguistic research questions. The theoretical framework and research design is presented in detail by Dreesen and Stücheli-Herlach (2019). It describes how to study specific thematic discourses with involved agents in a data-driven and transdisciplinary way. Swiss-AL has been used as a resource in two linguistic projects following that framework: on public discourses on energy in Switzerland (Stücheli-Herlach et al., 2018) and on antibiotic resistance (Borghoff et al., 2019). Both projects did a careful sampling of discourse-relevant sources which finally entered into the corpus (cf. Section 3.2.). There was a large overlap of sampled sources, which indicates that Swiss-AL is thematically rather unspecific, although it was initially sampled for specific discourses.

Both projects followed a novel approach to obtain discourse specific corpora, i.e., corpora exclusively containing texts on energy and on antibiotic resistance, respectively. Discourse specific texts were identified via topic modeling on the Swiss-AL-CHE language corpora. In the case of the energy discourse, 500 topics were calculated with an LDA algorithm (Blei, 2012) for each language specific corpus. In a first step, all topics were coded according to the principles of grounded theory and marked for their relevance for energy discourse. In a second step, all texts containing the relevant topics above a specific threshold were extracted which in the end led to the energy specific project corpus. A comparison of the underlying corpus and the resulting discourse specific corpus regarding the number of tokens and texts reveals that Swiss-AL is indeed discourse-unspecific, despite the fact that it was initially sampled for the analysis of the public discourse on energy in Switzerland (cf. Table 4.).

[7] `https://github.com/ambiverse-nlu/ambiverse-nlu`

[8] `https://www.elastic.co/de/`

	energy discourse corpora		Swiss-AL-CHE	
	tokens	texts	tokens	texts
German	15,704,867	18,609	700,789,708	1,299,510
French	9,125,584	12,778	342,728,211	612,286
Italian	2,735,511	5433	150,688,050	386,935

Table 2: Size of energy discourse corpora (EDC) for German, French, and Italian in comparison to the main corpora of which they form subsets (June 2019).

Bubenhofer et al. (2019) used Swiss-AL for the linguistic analysis of right-wing populism by comparing word embedding models computed for two different sub-corpora of Swiss-AL. The word embedding models were calculated using the word2vec-algorithm as presented in Mikolov et al. (2013). Comparing models that were trained on different corpora support the hypothesis, that word embedding models reveal different semantic patterns associated with a word. The first model was computed for the leading German magazine and newspapers Spiegel Online, Bild Online, and ZEIT Online (shortened ORIENT). The second model was computed for the right-wing newspapers PI-News and COMPACT (shortened PINES). From 100 calculated nearest neighbours to a word in PINES, these can be identical to a value between 0 and 100 with the corresponding word's neighbours in ORIENT. The number of nearest neighbours found is therefore between 0 and 100. If there are hardly any matches between the nearest neighbours (e.g. 0 or 1 of 100), this may be caused by the fact that the calculated nearest neighbours in PINES are out of vocabulary in ORIENT. But if all of the 100 words are part of the vocabulary of ORIENT but nevertheless do not show up as nearest neighbors, this points to different semantic patterns associated with the word under consideration in the two corpora. Words in PINES fulfilling these criteria (0 out of 100 out of vocabulary and 0 out of 100 shared next neighbors in ORIENT) show Islamophobic and hostility towards media and elite. Different semantic patterns were then verified by collocation analysis. Swiss-AL and word embeddings are also used to detect semantic-functional equivalents in meta-pragmatic-discourses (Dreesen and Bubenhofer, 2020).

Last but not least, Swiss-AL serves as a database for the annual election of the word of the year in Switzerland[9] and is successfully applied in linguistic bachelor and masters courses.

5. Conclusion

In this paper, we introduced Swiss-AL, a multilingual Swiss corpus for applied linguistics. The corpus is built to support corpus-based and corpus-driven linguistic discourse analysis. As a consequence, Swiss-AL consists of carefully sampled sources and is easily extensible due to a flexible processing pipeline. In order to achieve a corpus of high linguistic quality, scrapers are used for extracting relevant texts from websites. The corpus is accessible for corpus linguistic analysis on a publicly available online platform. We presented two successful use cases for researching discourse in Switzerland with respect to energy and antibiotic resistance.

As a next step, we will complete the processing of French and Italian with the missing components. In particular we are developing models for entity linking in the two languages which is desired in many application tasks. Apart from that we are building a new interface to streamline the process of defining subcorpora that can be enriched with specialized annotations. Furthermore, we will work on a solution to give users access to the texts contained in Swiss-AL eventually.

6. Acknowledgements

The work presented in this paper was funded by the Swiss Federal Office of Energy, the Swiss Federal Office of Public Health, and the Zurich University of Applied Sciences (ZHAW). We thank our colleagues, especially Noah Bubenhofer (Zurich University), Selena Calleri and Peter Stücheli-Herlach (ZHAW), and our project partners.

7. Bibliographical References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum, London, New York.

Barbaresi, A. and Würzner, K.-M. (2014). For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 2–10, Hildesheim, Germany.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *COLING '10. Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36, Beijing, China.

Blätte, A. (2016). polmineR. v0.6.2.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77.

Borghoff, B., Stücheli-Herlach, P., Schwarz, N., and Bilat, L. (2019). Antibiotikaresistenzen auf die Agenda! L'antibiorésistance à l'agenda : Schlussbericht zur

[9] `https://www.zhaw.ch/de/linguistik/wort-des-jahres-schweiz/`

situativen Analyse öffentlicher Diskurse über Antibiotikaresistenzen mittels digitaler Daten. Technical report, ZHAW Zürcher Hochschule für Angewandte Wissenschaften, Winterthur.

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

Bubenhofer, N. (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode Der Diskurs- Und Kulturanalyse*. Number 4 in Sprache Und Wissen. De Gruyter, Berlin, New York.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2019). Shiny: Web Application Framework for R. R package version 1.3.2.

De Groc, C. (2011). Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 497–498, Lyon, France. IEEE.

Dreesen, P. and Bubenhofer, N. (2020). Das Konzept "Übersetzen" in der digitalen Transformation. *Germanistik in der Schweiz*, 16:26–49.

Dreesen, P. and Stücheli-Herlach, P. (2019). Diskurslinguistik in Anwendung. Ein transdisziplinäres Forschungsdesign für korpuszentrierte Analysen zu öffentlicher Kommunikation. *Zeitschrift für Diskursforschung*, pages 123–164.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham,United Kingdom. University of Birmingham.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *In Proceedings of the 4th Web as Corpus Workshop (WAC-4). Can We Beat Google?*, Marrakech, Morocco.

Ferrucci, D. and Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, Ann Arbor, Michigan.

Fletcher, W. H. (2004). Making the web more useful as a source for linguistic corpora. In Ulla Connor et al., editors, *Corpus Linguistics in North America 2002*, pages 191–205. Rodopi, Amsterdam.

Gabrielatos, C. and Baker, P. (2008). Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005. *Journal of English Linguistics*, 36(1):5–38.

Habernal, I., Zayed, O., and Gurevych, I. (2016). C4Corpus: Multilingual Web-size Corpus with Free License. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference CL*, pages 125–127, Lancaster, United Kingdom.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., and Pirrelli, V. (2014). The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in neural information processing systems*, 26:3111–3119.

Nivre, J., Hall, J., and Nilsson, J. (2006). MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 2216–2219, Genoa, Italy. European Language Resources Association (ELRA).

Palmero Aprosio, A. and Moretti, G. (2016). Italy goes to Stanford: A collection of CoreNLP modules for Italian. *ArXiv e-prints*, September.

Schäfer, R. and Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Schäfer, R. (2016). On Bias-free Crawling and Representative Web Corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, Berlin, Germany.

Schäfer, R. (2017). Accurate and efficient general-purpose boilerplate detection for crawled web corpora. *Language Resources and Evaluation*, 51(3):873–889.

Schmid, H. and Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In *COLING 2008*, Manchester.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, United Kingdom.

Spitzmüller, J. and Warnke, I. H. (2011). *Diskurslinguistik. Eine Einführung in Theorien Und Methoden Der Transtextuellen Sprachanalyse*. De Gruyter, Berlin and Boston.

Stücheli-Herlach, P., Ehrensberger-Dow, M., and Dreesen, P. (2018). *Energiediskurse in der Schweiz : anwendung-*

sorientierte Erforschung eines mehrsprachigen Kommunikationsfelds mittels digitaler Daten. Number 16 in Working Papers in Applied Linguistics. ZHAW Zürcher Hochschule für Angewandte Wissenschaften, Winterhur.

Theobald, M., Siddharth, J., and Paepcke, A. (2008). SpotSigs: Robust and Efficient Near Duplicate Detection in Large Web Collections. In *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2008 (SIGIR 2008)*, Singapore.