1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.249

Published: 07 Sept. 2023

Swiss-AL: Platform for Language Data in Applied Sciences

On challenges in the field of Language Open Research Data

Julia Krasselt[1][https://orcid.org/0000-0003-1060-2657], Philipp Dreesen[1][https://orcid.org/0000-0001-5291-2798], Peter Stücheli-Herlach[1][https://orcid.org/0000-0002-3560-7182], Dolores Lemmenmeier[1][https://orcid.org/0000-0003-0541-6956], Sooyeon Cho[1][https://orcid.org/0009-0005-4172-7008], Klaus Rothenhäusler[1][https://orcid.org/0000-0003-4744-3362], and Matthias Fluor[1][https://orcid.org/0000-0002-0780-8024]

[1] ZHAW Zurich University of Applied Sciences

Abstract. Open Science is transforming the way researchers collect, process, analyze, and store empirical research data, particularly in the social sciences and humanities, where language data is crucial. This transformation process especially concerns developers and providers of large language corpora and manifests itself in at least three challenges when providing these corpora as Open Research Data (ORD). Challenges concern heterogeneous practices that researchers apply when working with language data, research data lifecycle, and legal and ethical aspect. In this paper, we present Swiss-AL, a language data platform developed in Switzerland that is being transformed into an Open Research Data Resource for Applied Sciences within the Swiss Open Science Strategy. The paper gives an overview over the data contained in Swiss-AL and the infrastructure that is used to process and analyze the data. Furthermore, it presents approaches to the three abovementioned challenges to language ORD.

Keywords: Language Data, Corpus Linguistics, Interdisciplinarity

1. Introduction

Open Science is revolutionizing the way researchers collect, process, analyze, and store empirical research data, particularly in the social sciences and humanities, where language data is crucial. However, sharing large language corpora with a diverse research community presents unique challenges, including structured and FAIR data access, disciplinary research practices, and compliance with copyright and data protection laws. Here, We will introduce Swiss-AL, a language data platform for Applied Sciences developed at ZHAW, Switzerland, that is currently being developed into an Open Research Data (ORD) Resource for the Swiss and European CLARIN Community This paper presents Swiss-AL's approach to these challenges, which is currently funded under the Swiss Open Science Strategy.

2. Swiss-AL: Platform for Language Data

2.1 Corpora and Processing Pipeline

Swiss-AL is a multilingual text collection designed for analyzing public communication and relevant societal discourses in interdisciplinary and transdisciplinary contexts [1]. The corpus currently contains 4.5 billion words, making it the largest Swiss text collection [2].

Swiss-AL provides three types of text collections: Swiss-AL Base, Swiss-AL Media, and Swiss-AL Projects (Fig. 1). Swiss-AL Base contains web-crawled texts published by various public actors, including political and administrative authorities, industry associations, Swiss universities, civil society, and newspapers. Swiss-AL Media focuses exclusively on journalistic media from major and regional newspapers of Swiss publishing houses. Swiss-AL Projects consists of thematically specific corpora compiled for research projects and shared with the research community.

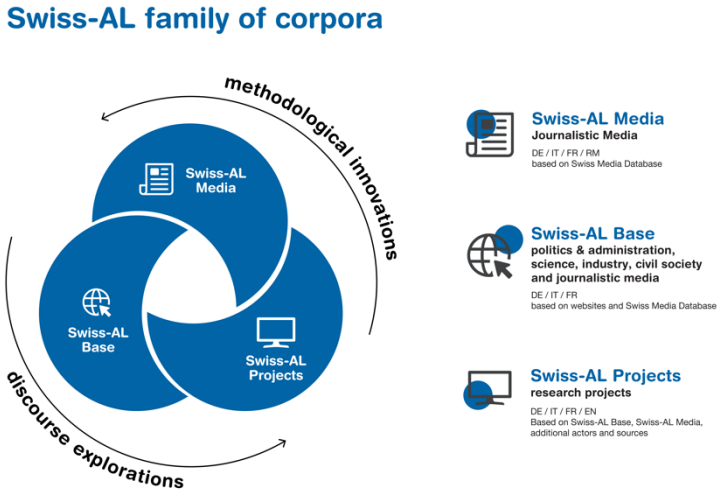

Figure 1. Swiss-AL family of corpora

Swiss-AL corpora are compiled using a computational linguistic pipeline that can adapt to different types of input data. The focus is on dynamic parts of websites (subpages covering news reports, media releases, blogs), using a web crawler and web page-specific Xpath scrapers. The data is loaded into an Elastic Search database in a structured form. Filters are applied to process texts in language-specific ways and to recognize and sort out near-duplicates [3]. The linguistic processing is based on the UIMA framework [4] and various modules such as TreeTagger for PoS tagging [5], Stanford NER for named entity recognition [6]. the Stanford Dependency Parser [7] or in-house developed tools for additional annotation layers.

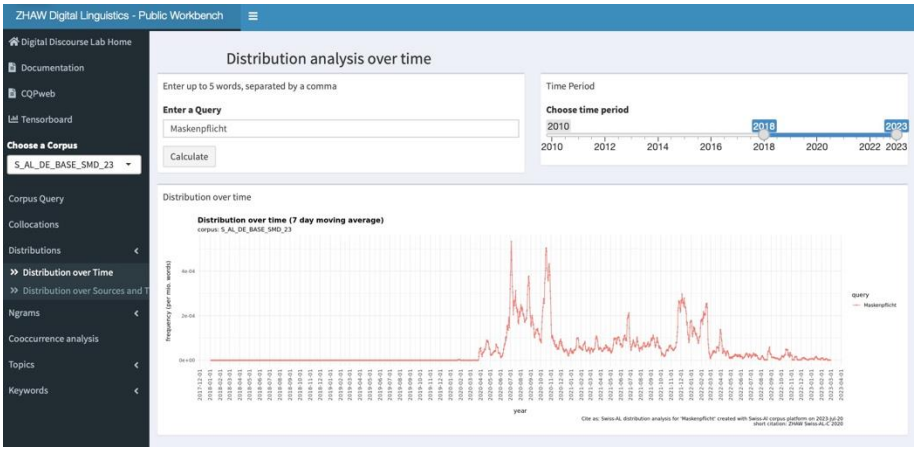

Figure 2. Swiss-AL Workbench. Left: Selection of a sub-corpus and a desired analysis method, access to the documentation. Right: Display of the results. Here, the frequency of a word (*Maskenpflicht*) over time (2018-2023) in the Swiss-AL-Base corpus is shown.

2.2 Access: Swiss-AL Workbench

Swiss-AL corpora can be accessed via an in-house developed public web app, the *Swiss-AL Workbench* (Fig. 2, [8]). The focus is on aggregating methods of data analysis (*distant reading*), in which phenomena on the language surface are summarized quantitatively. The workbench enables word-based query methods (e.g., word distributions over time) and analyses using machine and deep learning methods (topic modelling, word embeddings). In the course of developing Swiss-AL into an ORD resource, the workbench will be reimplemented by the end of 2024 in order to approach the challenges mentioned in the following section.

3. Challenges related to Linguistic Open Research Data

Text data not only form the empirical basis in linguistics, but in many other research disciplines. This makes language data like Swiss-AL different from other data like genome sequences and particularly valuable as an ORD resource. At the same time, however, this also gives rise to special challenges, three of which we will discuss in the following sections.

3.1 Heterogeneous research practices

Humanities, social sciences, communications sciences, law, and architecture studies are examples for disciplines interested in language data, each employing diverse research practices with labels such as *qualitative*, *quantitative*, *data-driven*, *hypothesis-driven*, *close reading*, and *distant reading*. When developing Swiss-AL into an ORD resource, it is crucial to consider these varied user groups and research methods. These users, labelled semi-professionals from a corpus-linguistic perspective, are experienced empirical researchers in their fields but unfamiliar with corpus linguistic methods and linguistic surface analysis. Keeping this target user group in mind, we developed user stories, which answer who wants to do what for which purpose. Acceptance criteria derived from these user stories guide the re-implementation of the existing workbench.

3.2 Research Data Lifecycle

Billion-word corpora like Swiss-AL need special infrastructures for structured access to primary data, metadata, and documentation, as conventional repositories risk violating copyright law and serve only a limited expert community. Swiss-AL promotes reuse by integrating curated data, storage, and analysis tools to support interdisciplinary communities via a dynamic FAIR infrastructure. Traditional repositories, found at the data lifecycle's end, are less conducive to reuse due to the divide between storage and analysis tools. FAIR principles are essential for re-implementing the Swiss-AL workbench. E.g., it will enable users to perform not only word-based analysis, but also to download corresponding data frames, corpus information, and code for reproducing visualizations.

3.3 Legal and ethical aspects

Publishing language data as an ORD resource requires considering legal and ethical aspects. Texts from SMD and crawled web data in Swiss-AL are protected by Swiss copyright law. However, under §24d of the Swiss Copyright Act, reproducing work for scientific research is permissible if it involves a technical process and the copied work is lawfully accessible. Thus, crawling web data for research is allowed. However, representing full texts, a common researcher need, warrants caution. A scientific legal opinion is currently being prepared in collaboration with lawyers to examine possibilities in this domain.

Secondly, language corpora such as Swiss-AL contain personal data (e.g., journalistic media articles mentioning real persons), i.e., the identification of an individual person is potentially possible. Thus, data protection law needs to be considered when obtaining, saving, and

publishing corpus data. The topic is well known from other empirical research fields, e.g., when conducting qualitative interviews or doing field work. However, the practice of anonymisation typically used there is not a practicable solution for large linguistic corpora. In particular, an envisaged solution is to formulate a purpose for which the data contained in Swiss-AL will be collected, stored and analysed.

4. Conclusion

Language data are not prototypical ORD, and they are not an exclusive data resource for linguistics. However, if language data are to be made available as ORD for other disciplines, a variety of challenges arise. These can only be solved in an interdisciplinary way, taking into account technical, legal and ethical aspects on a societal and international level.

Data availability statement

The corpora described in the article can be accessed under www.swiss-al.linguistik.zhaw. A documentation of the corpora is available under https://swiss-al.linguistik.zhaw.ch/docs/ord/.

Author contributions

Conceptualization: JK, PD, PSH; Writing – original draft: JK; Writing – review & editing: JK, PD, PSH, DL, SC; Software: KR, MF

Competing interests

The authors declare that they have no competing interests.

Funding

The project presented in this paper is currently funded by swissuniversities within the programme Swiss Open Research Data Grants (CHORD) and the Zurich University of Applied Sciences (internal funding).

References

[1] P. Dreesen and P. Stücheli-Herlach, "Diskurslinguistik in Anwendung. Ein transdisziplinäres Forschungsdesign für korpuszentrierte Analysen zu öffentlicher Kommunikation", Zeitschrift für Diskursforschung, vol. 7, no. 2, pp. 123–162, 2019, doi: https.doi.org/10.3262/ZFD1902123

[2] J. Krasselt, P. Dreesen, M. Fluor, C. Mahlow, K. Rothenhäusler, and M. Runte, "Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics", in Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 2020, pp. 4138--4144. https://aclanthology.org/2020.lrec-1.510/ [26.04.2023]

[3] M. Theobald, J. Siddharth, and A. Paepcke, "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, Singapore, 2008.

[4] D. Ferrucci and A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment", Natural Language Engineering, vol. 10, no. 3–4, pp. 327–348, 2004, doi: https.doi.org/10.1017/S1351324904003523.

[5] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees", in Proceedings of the international conference on new methods in language processing, Manchester, United Kingdom, 1994, pp. 44–49. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1139

[6] J. R. Finkel, T. Grenager, and C. Manning, "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), Ann Arbor, Michigan, 2005, pp. 363–370.

[7] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit", in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland USA, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[8] J. Krasselt, M. Fluor, K. Rothenhäusler, and P. Dreesen, "A workbench for corpus linguistic discourse analysis", in 3rd conference on language, data and knowledge (LDK 2021), D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, and B. Heinisch, Eds., in Open access series in informatics (OASIcs), vol. 93. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, p. 26:1-26:9. doi: https.doi.org/10.4230/OASIcs.LDK.2021.26.