Swiss-AL: Language Data Platform for Applied Sciences

Julia Krasselt, Philipp Dreesen, Dolores Lemmenmeier-Batinić, Sooyeon Geckeler, Klaus Rothenhäusler, Matthias Fluor

Zurich University of Applied Sciences, Institute of Language Competence
Theaterstrasse 17, 8400 Winterthur, Switzerland
{krss, dree, leme, chos, rotk, fluor}@zhaw.ch

Abstract

This paper introduces Swiss-AL, a language data platform designed for the multilingual, comparative analysis of public discourse in Switzerland. Swiss-AL is an open research data resource providing browser-based access to a variety of corpora in all four of Switzerland's official languages. Corpora contain journalistic, organisational, and parliamentary discourse. The platform supports research in applied linguistics as well as neighbouring disciplines (e.g., social sciences, communication and media studies).

Keywords: multilingual discourse corpus, public discourse, corpus analysis platform, applied sciences

1. Swiss-AL: Purpose and Composition of the resource

Swiss-AL is a language data platform for applied sciences, hosted and developed by the Digital Discourse Lab at ZHAW University of Applied Sciences (http://swiss-al.zhaw.ch). It is designed for the analysis and comparison of multilingual public discourses, with a focus on Switzerland. Swiss-AL is part of the Swiss linguistic research infrastructure landscape CLARIN-CH and a key component of the CLARIN knowledge centre for applied comparative discourse analysis (CLARIN-APPLIED, www.clarin-applied.zhaw.ch), hosted at ZHAW.

Swiss-AL contains a collection of linguistic corpora comprising publicly available documents from political, industrial, civil society, scientific and journalistic actors from all four language regions of Switzerland. Swiss-AL is not intended as a reference corpus, e.g. for Swiss High German, but rather as an empirical basis for analysing communicative practices in discursively constructed communication contexts. Swiss-AL is composed of three main corpus types:

- Journalistic corpora containing news articles published in Swiss daily and weekly newspapers and magazines by the country's leading publishing houses. The data covers a period from 2010 onwards and is available in all four Swiss national languages. It is provided by the Swiss Media Database via Swissdox@LiRi (provided by Zurich University).
- Organizational corpora containing press releases, news items, and blog posts from the official websites of over 360 actors in politics, administration, science, industry, and civil society. For example, these include the websites of all 26 Swiss cantons, the websites of parties represented in the Swiss National Council, and the websites of all Swiss universities. The data is available in three languages (German, French and Italian), covers the period from 2010 onwards, and is collected using a web-crawling and -scraping pipeline developed at ZHAW (Krasselt et al., 2023).
- Parliamentary corpora containing transcripts of speeches given by politicians in national parliamentary debates. The data is available in German, French and Italian, covers the period from 1999, and is provided via the parliamentary service's API.

In addition, Swiss-AL contains corpora compiled in the context of specific research projects conducted by the ZHAW Digital Discourse Lab (e.g, on covid-19 and vaccination discourses).

In the context of the Swiss Open Science Strategy, Swiss-AL has been developed into an Open Research Data resource since 2022 (Krasselt et al., 2023). This includes the implementation of FAIR-principles, the development and provision of a browser-based workbench for corpus analysis and the systematic consideration of legal issues concerning the publication of language data.

2. Access and Functionalities

A browser-based workbench is available for researchers to access and analyse the corpora. As the central interface for Swiss-AL, it provides access to corpus data in accordance with legal requirements, particularly with regard to copyright restrictions and data protection. This means that corpora cannot be downloaded and only short extracts of documents are displayed directly on the workbench. Where possible, a link to the original document on an external website is provided (e.g. Swissdox Essentials for all Journalistic corpus documents).

The workbench enables users to access a wide range of Swiss-AL corpora (see Table 1 for a selection) and create custom subcorpora based on criteria such as source, search terms and time spans. In the dedicated workspace, users can open multiple tabs and choose between the

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 104–105. Palma, Mallorca. May 11, 2026.

following data-driven and search-term-based analysis modes.

- In data-driven exploratory mode, users can analyse the corpus without entering a specific search term. This mode provides word embedding models and LDA-based topic models for all corpora (with predefined parameters such as number of topics), as well as keyword lists for all user-created subcorpora, currently based on Log Likelihood (Figure 1).

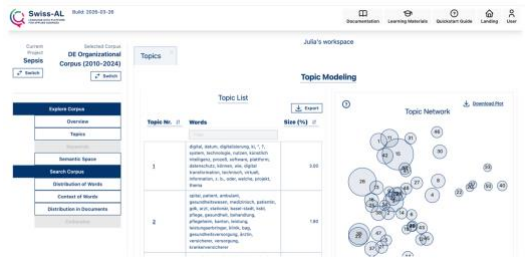

Figure 1: Screenshot of the topic modeling module on the Swiss-AL workbench

- In the keyword-based analysis mode, users can enter a specific search term and choose from three different search modes: simple, basic, or a CQP-based advanced search. Users can analyse the frequency and distributional patterns of a given search term, view classical and document-based concordances, calculate collocations (currently based on LogDice) and access the full text containing the search term on platform external websites (Figure 2).

	Corpus	Size (in token)
Journalistic Corpora	DE Journalistic Corpus (high reach media, 2010-2025)	1.2 billion
	IT Journalistic Corpus (high reach media, 2010-2025)	32 million
	FR Journalistic Corpus (high reach media, 2010-2024, 20%)	1.3 billion
	RM Journalistic Corpus (high reach+regional media, 2018-2025)	22 million
Organizational Corpora	DE Organizational Corpus (2010-2024)	66 million
	FR Organizational Corpus (2010-2024)	29 million
	IT Organizational Corpus (2010-2024)	18 million
Parliamentary Corpora	DE Swiss Federal Parliament Debates Corpus (1999-2024)	10 million
	FR Swiss Federal Parliament Debates Corpus (1999-2024)	4 million
	IT Swiss Federal Parliament Debates Corpus (1999-2024)	150,000

Table 1: Selection of corpora available on the Swiss-AL workbench (DE = German, FR = French, IT = Italian, RM = Romansh)

3. Bibliographical References

Krasselt, J., Dreesen, P., Fluor, M., & Rothenhäusler, K. (2023). Swiss-AL. Korpus und Workbench für mehrsprachige digitale Diskurse. In M. Kupietz & T. Schmidt (Eds.), *Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022*, pp. 127-142

Krasselt, J., Dreesen, P., Stücheli-Herlach, P., Lemmenmeier, D., Cho, S., Rothenhäusler, K., & Fluor, M. (2023). Swiss-AL: Platform for Language Data in Applied Sciences: On Challenges in the Field of Language Open Research Data. *Proceedings of the Conference on Research Data Infrastructure*, *1*. https://doi.org/10.52825/cordi.v1i.249