SwissCoco2025 - The Swiss Corpora Collection 2025

Mark Cieliebak, Jonathan Gerber, Manuela Hürlimann

Zurich University of Applied Sciences (ZHAW) and
Swiss Association for Natural Language Processing (SwissNLP)
[ciel, gerj, hueu]@zhaw.ch

Abstract

The Swiss Corpora Collection (SwissCoco) gives an overview of language resources that are relevant for Switzerland. It focuses on text, speech, and documents. The Swiss Corpora Collection 2025 is the first iteration of the collection and contains 30 corpora. All corpora are also listed at https://swissnlp.org/swiss-corpora-collection.

1 Introduction

Language resources such as text and speech corpora are fundamental building blocks for research and applications in Natural Language Processing, Computational Linguistics, Large Language Models, Machine Learning and Artificial Intelligence in general. They provide the essential data for training and evaluating models and for conducting linguistic analyses. While a wealth of resources exists for major world languages, finding and accessing corpora for specific linguistic varieties, dialects, and smaller languages remains a significant challenge. This is particularly true for Switzerland, with its four national languages and diverse regional dialects, notably Swiss German, the predominant spoken variety in the German-speaking part of the country.

To address this challenge in the Swiss context, we introduce the **Swiss Corpora Collection (SwissCoco)**, an initiative to systematically identify, catalogue, and provide an overview of language resources relevant to Switzerland. SwissCoco is organized by the Swiss Association for Natural Language Processing (SwissNLP) at https://swissnlp.org/swiss-corpora-collection. This paper presents the inaugural collection, **SwissCoco2025**, which brings together publicly available corpora that are essential for advancing research for Swiss languages and linguistic varieties.

For a corpus to be included in SwissCoco, it must meet at least one of the following criteria that define it as "Switzerland-related":

1. is produced in Switzerland,
2. represents one or more Swiss national or regional languages (German, French, Italian, Romansh, Swiss German dialects),
3. reflects Swiss-specific linguistic varieties,
4. originates from a Swiss entity, or
5. contains content relevant to Swiss contexts.

Furthermore, all corpora in the collection must be publicly available to ensure accessibility for the wider research community. **SwissCoco2025** is the first iteration of this collection and contains 30 corpora. The full list and detailed information on each corpus are also available on the SwissNLP website at https://swissnlp.org/swiss-corpora-collection. New corpora are added throughout the year on the website and will be formally documented in the next annual survey paper in 2026.

2 Corpus Curation

The curation of the corpora followed a methodology centered on two core principles, as outlined in the introduction: corpora must be **Switzerland-related** and **publicly available**.

The collection process for the inaugural Swiss Corpora Collection 2025 leveraged existing knowledge within the SwissNLP community and was complemented by targeted web searches and manual information extraction. More precisely, the initial list of corpora was compiled from resources already known to the authors and other members of SwissNLP. This list was then expanded through

web searches to identify additional publicly available resources. The focus was on identifying a comprehensive set of foundational corpora that have been previously used in research.

The SwissText 2025 Conference hosted a special "Corpora Track", which called for Switzerland-related submissions. Two corpora were published within this track (SwissGPC and SPC_R), which are also included in the corpora list of SwissCoco2025.

The identified corpora were individually reviewed against the inclusion criteria. Each entry in the collection includes metadata such as the corpus name, a brief description, the languages covered, the data type (e.g., text, speech), and a link to the resource, which were extracted manually from the corpus' documentation and references. The resulting collection, as presented in the following section, provides the most comprehensive overview of publicly available Switzerland-related language resources to date.[1]

3 SwissCoco2025

The Swiss Corpora Collection 2025 contains 30 corpora and datasets that are relevant for the Swiss context. Table 1 gives an overview of the corpora. The full details including annotation types, license information, contact addresses and other relevant metadata, are provided in Appendix A.

In SwissCoco2025, 18 corpora contain texts, 12 contain speech recordings, and one contains web pages with text and images. There are 14 corpora with Swiss German data (text or speech) and one specific for Valais German.

4 Conclusion

This paper presents the inaugural edition of the Swiss Corpora Collection (SwissCoco2025), a comprehensive overview of language resources relevant to Switzerland. We have curated a collection of 30 publicly available corpora, spanning various modalities, including text, speech, and documents, and covering the country's national and regional languages. This collection, which is also publicly listed at https://swissnlp.org/swiss-corpora-collection, is a centralized resource to facilitate NLP research for Swiss-related linguistic varieties.

Building on the foundation of SwissCoco2025, we plan to systematically expand the collection by identifying new resources and regularly updating the metadata of existing ones. By providing this curated collection, we hope to lower the entry barrier for new researchers and contribute to the development of robust and effective NLP systems tailored to Switzerland's unique multilingual and multi-dialectal landscape.

[1] If you have suggestions for additional corpora, or if you notice any errors, please send an email to info@swissnlp.org

Table 1: All Corpora of SwissCoco2025. Column "Com. Use" indicates whether commercial use is possible. See Appendix for more details.

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
1	**ArchiMob**: Archives de la mobilisation	Speech, Text	Swiss German	Transcripts of interviews on the mobilisation in the Second World War in Switzerland	500k tokens	No	Link
2	**BCMS-MT**: Map Task Corpus of Heritage BCMS	Speech, Text	Swiss German	Spontaneous dialogues in Swiss German. Recordings are annotated with dialogue acts and speaker characteristics.	3 hours	Yes	Link
3	**CEASR**: Corpus for Evaluating Automatic Speech Recognition	Speech, Transcripts	German, English	Audio recordings from nine English and six German speech corpora and accompanying transcriptions generated by seven different ASR systems.	56 hours, 1360 speakers	Yes	Link
4	**CHEU-lex**: CHEU-lex Corpus	Text	German, French, Italian	Parallel and comparable corpus of Swiss and European Union (EU) legislation.	Not specified	Yes	Link
5	**DS21 Corpus**: Corpus of Historical Legal Texts	Text: Documents	German, French, Italian, Romansh, Latin	Historical Swiss legal texts from the early Middle Ages to 1798. Based on the Collection of Swiss Law Sources.	Varies by canton and volume	Yes	Link
6	**GSA-Data**: German Speaking Area Data	Text	German, Swiss German, Austrian	German, Austrian and Swiss German Jodels with geolocations	16.8M posts	Unk.	Link
7	**LEDGAR**: Multi-label Corpus for Text Classification of Legal Provisions in Contracts	Text: Legal Documents	English	Legal judgments from the Swiss Federal Supreme Court, intended for legal document analysis.	60k legal documents, 100k provisions, 12k labels	Yes	Link

Continued on next page

Table 1 – continued from previous page

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
8	**LEX.CH.IT**: Corpus for Micro-Diachronic Linguistic Investigations of Swiss Normative Acts in Italian	Text: Documents	Italian	Monolingual corpus of Swiss normative acts, including 366 federal acts, from 1974 to 2018.	366 acts	No	Link
9	**MediaParl**: MediaParl Bilingual Database	Speech	French, German	Bilingual speech database with recordings from the Valais Parliament.	16k sentences, 210 speakers	No	Link
10	**NOAH's Corpus**: NOAH's Corpus of Swiss German Dialects	Text: Various Sources	Swiss German	Swiss German texts from various genres, including Wikipedia articles, news, blogs, and novels. Manually annotated with Part-of-Speech tags.	73k tokens	No	Link
11	**SB-10k**: German Sentiment Corpus	Text: Tweets	German	German tweets from 2017, each annotated by 3 annotators with sentiment labels "positive", "negative", "neutral", "mixed", or "unknown".	9700 tweets	Yes	Link
12	**SB-CH**: Swiss German Sentiment Corpus	Text: Social Media	Swiss German	Swiss German sentences from Facebook comments and online chats. Includes manual sentiment labels for some sentences.	166k sentences, 2800 with sentiment labels	Yes	Link
13	**SDATS Corpus**: Swiss German Dialects Across Time and Space Corpus	Speech	Swiss German	Spoken Swiss German recordings from 1,000 speakers across 125 localities.	1k speakers, 125 localities, 300 variables	Yes	Link

Continued on next page

Table 1 – continued from previous page

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
14	**SDS-200**: Schweizer Dialeksammlung-200	Speech	Swiss German	Swiss German audio recordings with transcripts in Standard German. Collected via a crowdsourcing web app. Intended for ASR, TTS, and dialect identification.	200 hours, 4000 speakers	Yes	Link
15	**SMG-CH**: Social Media Variety Geolocation - Swiss German	Text	Swiss German	Swiss German Jodels with geolocations	29k Jodel conversations	Unk.	Link
16	**SPC_R**: Swiss Parliaments Corpus Re-Imagined	Speech, Text	Swiss German, Standard German	Enhanced long-form speech-text corpus of Swiss German parliamentary debates, with high-quality, corrected transcriptions.	751 hours	Yes	Link
17	**STT4SG-350**: Speech-to-Text for Swiss German-350	Speech	Swiss German	Swiss German audio recordings with transcripts in Standard German. Balanced distribution across dialects and demographics such as gender. Collected via a crowdsourcing web app. Intended for ASR, TTS, and dialect identification. Dedicated test set with approx. 5 hours of audio of identical sentences spoken in 7 different dialects.	343 hours, 316 speakers	Yes	Link

Continued on next page

Table 1 – continued from previous page

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
18	**Swiss Politics Corpus**: Swiss Politics Corpus	Text: Documents	German, French, Italian	A database of who said what and when in both chambers of the Swiss parliament over the past 127 years, based on digitized proceedings with oldest documents being from 1891	40k documents	No	Link
19	**Swiss SMS Corpus**: Swiss SMS Corpus	Text: SMS	Swiss German, German, French, Italian, Romansh	SMS messages crowdsourced from the Swiss public.	26k SMS messages, 650k tokens	No	Link
20	**SwissCrawl**: SwissCrawl Web Corpus	Text: Web	Swiss German	Large-scale, multilingual web crawl of the .ch domain.	560k sentences	No	Link
21	**SwissDial**: Parallel Multidialectal Corpus of Spoken Swiss German	Speech	Swiss German	Audio recordings of eight major Swiss German dialects and corresponding transcripts in Swiss German and Standard German	26 hours, 8 speakers	No	Link
22	**SwissGPC**: Swiss German Podcasts Corpus	Speech	Swiss German	Links to Swiss podcast from Swiss TV broadcasters	5000 hours	No	Link
23	**TEVOID**: Temporal Voice Idiosyncrasy	Speech, Text	Swiss German (Zurich)	Recordings of read and spontaneous speech of different speakers of Zurich German for the research on idiosyncratic differences	16 sentences	Yes	Link
24	**TRANSLIT**: Large-scale Name Transliteration Resource	Text	180 languages	Variations of person and geolocation names in various languages.	1.6 million entries	Yes	Link

Continued on next page

Table 1 – continued from previous page

ID	Name	Modalities	Languages	Description	Size	Com. Use	Link
25	**VarDial_GDI17**: German Dialect Identification at VarDial 2017	Text	Swiss German	Text containing dialects from the cantons of Basel (BS), Bern (BE), Lucerne (LU), and Zurich (ZH)	18k sentences	Unk.	Link
26	**Walliserdeutsch ASR Corpus**: ASR and Translation Corpus for Walliserdeutsch	Speech	Walliserdeutsch, German	Broadcast news from a local radio station in Walliserdeutsch dialect.	8.1 hours	Yes	Link
27	**WebClasSeg25**: WebClasSeg-25: A Dual-Classified Webpage Segmentation Dataset	Text, Image	25 languages	Webpages from public sector websites of Europe	2580 webpages	Yes	Link
28	**What's up, Switzerland?**: Swiss Chat Corpus	Text: Messages	Swiss German, German, French, Italian, Romansh, Spanish, Slavic languages	WhatsApp chat messages, gathered in Summer 2014	760k messages, 5.1mio tokens	No	Link
29	**ZHCORPUS**: Zurich Corpus of Vowel and Voice Quality	Speech, Text	Swiss German (Zurich)	Focused on sounds of the long Standard German vowels produced with varying basic production parameters	34k utterances, 70 speaker	Yes	Link
30	**ZTC_BAS**: Zurich Tangram Corpus	Speech	Swiss German (Zurich)	Recordings of Swiss German dialect from Zurich, including transcriptions.	2 hours	Unk.	Link

APPENDIX: Full Corpora Details

This appendix provides the full details for each of the corpora listed in SwissCoco2025. The data provides detailed information about each resource, including its full name, authors/creators, languages, modalities, size, publication year, licensing details, and a direct link to the data or its corresponding publication.

1: ArchiMob - Archives de la mobilisation

Reference: (Samardžić et al., 2016)
Modalities: Speech, Text
Languages: Swiss German
Description: Transcripts of interviews on the mobilisation in the Second World War in Switzerland
Size: 500k tokens
Annotations: Transcripts, Audio, metadata
Publication Year: 2016
License(s): CC NC 4.0
Commercial Use: No
Website: Link
Reference: Tanja Samardzic, Yves Scherrer, Elvira Glaser: *ArchiMob - A corpus of Spoken Swiss German*. LREC 2016.
Paper Link: Link
Contact: tanja.samardzic@uzh.ch

2: BCMS-MT - Map Task Corpus of Heritage BCMS

Reference: (Lemmenmeier-Batinić et al., 2023)
Modalities: Speech, Text
Languages: Swiss German
Description: Spontaneous dialogues in Swiss German. Recordings are annotated with dialogue acts and speaker characteristics.
Size: 3 hours
Annotations: Dialogue acts, speaker characteristics
Publication Year: 2023
License(s): Not specified
Commercial Use: Yes
Website: Link
Reference: Dolores Lemmenmeier-Batinić, Josip Batinić, Anastasia Escher: *Map Task Corpus of Heritage BCMS spoken by second-generation speakers in Switzerland*. Language Resources and Evaluation 57.4, 2023.
Paper Link: Link
Contact: dolores.lemmenmeier@uzh.ch

3: CEASR - Corpus for Evaluating Automatic Speech Recognition

Reference: (Ulasik et al., 2020)
Modalities: Speech, Transcripts
Languages: German, English
Description: Audio recordings from nine English and six German speech corpora and accompanying transcriptions generated by seven different ASR systems.
Size: 56 hours, 1360 speakers
Annotations: Transcripts from different ASR engines, meta-data such as gender, accent etc.
Publication Year: 2020
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Malgorzata Anna Ulasik, Manuela Hürlimann, Fabian Germann, Esin Gedik, Fernando Benites, Mark Cieliebak: *CEASR: A Corpus for Evaluating Automatic Speech Recognition*. LREC 2020.
Paper Link: Link
Contact: ciel@zhaw.ch

4: CHEU-lex - CHEU-lex Corpus

Reference: (Felici, 2025)
Modalities: Text
Languages: German, French, Italian
Description: Parallel and comparable corpus of Swiss and European Union (EU) legislation.
Size: Not specified
Annotations: Structural, morphosyntactic, and content-related information
Publication Year: 2020
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Annarita Felici: *CHEU-lex: a parallel multilingual corpus of Swiss and EU legislation*. Applied Corpus Linguistics 2025.
Paper Link: Link
Contact: Annarita.Felici@unige.ch

5: DS21 Corpus - Corpus of Historical Legal Texts

Reference: (Höfler and Piotrowski, 2011)
Modalities: Text: Documents
Languages: German, French, Italian, Romansh, Latin
Description: Historical Swiss legal texts from the early Middle Ages to 1798. Based on the Collection of Swiss Law Sources.
Size: Varies by canton and volume
Annotations: Transcribed, annotated, commented
Remarks: Dataset link does not work
Publication Year: 2011
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Alternative Names: Collection of Swiss Law Sources
Reference: Stefan Höfler, Michael Piotrowsk: *Building Corpora for the Philological Study of Swiss Legal Texts*. 2011.
Paper Link: Link
Contact: Not available

6: GSA-Data - German Speaking Area Data

Reference: (Hovy and Purschke, 2018)
Modalities: Text
Languages: German, Swiss German, Austrian
Description: German, Austrian and Swiss German Jodels with geolocations
Size: 16.8M posts
Annotations: Coordinates
Publication Year: 2018
License(s): not specified
Commercial Use: Unk.
Website: Link
Reference: Dirk Hovy, Christoph Purschke: *Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting*. EMNLP 2018.
Paper Link: Link
Contact: dirk.hovy@unibocconi.it

7: LEDGAR - Multilabel Corpus for Text Classification of Legal Provisions in Contracts

Reference: (Tuggener et al., 2020)
Modalities: Text: Legal Documents
Languages: English
Description: Legal judgments from the Swiss Federal Supreme Court, intended for legal document analysis.
Size: 60k legal documents, 100k provisions, 12k labels
Annotations: Labeled provisions in contracts and legal texts
Remarks: Created by ZHAW and a Swiss Startup
Publication Year: 2020
License(s): MIT License
Commercial Use: Yes
Website: Link
Reference: Don Tuggener, Pius von Däniken, Thomas Peetz, Mark Cieliebak: *LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts*. LREC 2020.
Paper Link: Link
Contact: tuge@zhaw.ch

8: LEX.CH.IT - Corpus for Micro-Diachronic Linguistic Investigations of Swiss Normative Acts in Italian

Reference: (Canavese, 2019)
Modalities: Text: Documents
Languages: Italian
Description: Monolingual corpus of Swiss normative acts, including 366 federal acts, from 1974 to 2018.
Size: 366 acts
Annotations: -
Remarks: No download link found
Publication Year: 2019
License(s): CC BY 4.0
Commercial Use: No
Website: Link
Reference: Paolo Canavese: *LEX.CH.IT: A Corpus for Micro-Diachronic Linguistic Investigations of Swiss Normative Acts in Italian*. Comparative Legilinguistics 40.1, 2019.
Paper Link: Link
Contact: Paolo.Canavese@unige.ch

9: MediaParl - MediaParl Bilingual Database

Reference: (Imseng et al., 2012)
Modalities: Speech
Languages: French, German
Description: Bilingual speech database with

recordings from the Valais Parliament.
Size: 16k sentences, 210 speakers
Annotations: Transcripts, speaker metadata, language tags
Remarks: Split German-French approx. 50:50
Publication Year: 2012
License(s): Non-commercial research only
Commercial Use: No
Website: Link
Reference: David Imseng, Hervé Bourlard, Holger Caesar, Philip N. Garner, Gwénolé Lecorvé, Alexandre Nanchen: *MediaParl: Bilingual mixed language accented speech database*. Spoken Language Technology 2012.
Paper Link: Link
Contact: Not available

10: NOAH's Corpus - NOAH's Corpus of Swiss German Dialects

Reference: (Hollenstein and Aepli, 2014)
Modalities: Text: Various Sources
Languages: Swiss German
Description: Swiss German texts from various genres, including Wikipedia articles, news, blogs, and novels. Manually annotated with Part-of-Speech tags.
Size: 73k tokens
Annotations: Part-of-Speech tags
Versions: V1.0 from 2014, NOAH 3.0 contains 114k tokens
Publication Year: 2014
License(s): CC Attribution 4.0
Commercial Use: No
Website: Link
Reference: Nora Hollenstein, Noemi Aepli: *Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging*. VarDial 2014.
Paper Link: Link
Contact: Not available

11: SB-10k - German Sentiment Corpus

Reference: (Cieliebak et al., 2017)
Modalities: Text: Tweets
Languages: German
Description: German tweets from 2017, each annotated by 3 annotators with sentiment labels "positive", "negative", "neutral", "mixed", or "unknown".
Size: 9700 tweets
Annotations: Sentiment labels
Publication Year: 2017
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Mark Cieliebak, Jan Deriu, Fatih Uzdilli, Dominic Egger: *A Twitter Corpus and Benchmark Resources for German Sentiment Analysis*. SocialNLP 2017.
Paper Link: Link
Contact: info@spinningbytes.com

12: SB-CH - Swiss German Sentiment Corpus

Reference: (Grubenmann et al., 2018)
Modalities: Text: Social Media
Languages: Swiss German
Description: Swiss German sentences from Facebook comments and online chats. Includes manual sentiment labels for some sentences.
Size: 166k sentences, 2800 with sentiment labels
Annotations: Sentiment labels
Publication Year: 2018
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Ralf Grubenmann, Don Tuggener, Pius von Daniken, Jan Deriu, Mark Cieliebak : *Towards a Corpus of Swiss German Annotated with Sentiment*. LREC 2018.
Paper Link: Link
Contact: info@spinningbytes.com

13: SDATS Corpus - Swiss German Dialects Across Time and Space Corpus

Reference: (Leemann et al., 2020)
Modalities: Speech
Languages: Swiss German
Description: Spoken Swiss German recordings from 1,000 speakers across 125 localities.
Size: 1k speakers, 125 localities, 300 variables
Annotations: Sociolinguistic and psycholinguistic metadata, phonetic variables
Publication Year: 2020
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Adrian Leemann, Péter Jeszenszky Carina Steiner, Jan Messerli, Melanie Studerus:

SDATS Corpus – Swiss German dialects across time and space. 2020.
Paper Link: Link
Contact: carina.steiner@phbern.ch

14: SDS-200 - Schweizer Dialeksammlung-200

Reference: (Plüss et al., 2022)
Modalities: Speech
Languages: Swiss German
Description: Swiss German audio recordings with transcripts in Standard German. Collected via a crowdsourcing web app. Intended for ASR, TTS, and dialect identification.
Size: 200 hours, 4000 speakers
Annotations: Transcripts, dialect information, age group, gender.
Remarks: Same data format as STT4SG-350
Publication Year: 2022
License(s): META-SHARE NonCommercial NoRedistribution
Commercial Use: Yes
Website: Link
Reference: Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, Manfred Vogel: *SDS-200: A Swiss German Speech to Standard German Text Corpus*. LREC 2022.
Paper Link: Link
Contact: info@swissnlp.org

15: SMG-CH - Social Media Variety Geolocation - Swiss German

Reference: (Gaman et al., 2020)
Modalities: Text
Languages: Swiss German
Description: Swiss German Jodels with geolocations
Size: 29k Jodel conversations
Annotations: Sentences, Coordinates
Remarks: Dataset used in VarDial 2020
Publication Year: 2020
License(s): Not specified
Commercial Use: Unk.
Website: Link
Reference: Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, Marcos Zampieri: *A Report on the VarDial Evaluation Campaign 2020*. VarDial 2020.
Paper Link: Link
Contact: yves.scherrer@helsinki.fi

16: SPC_R - Swiss Parliaments Corpus Re-Imagined

Reference: (Timmel et al., 2025)
Modalities: Speech, Text
Languages: Swiss German, Standard German
Description: Enhanced long-form speech-text corpus of Swiss German parliamentary debates, with high-quality, corrected transcriptions.
Size: 751 hours
Annotations: Transcriptions, LLM-based correction, predicted BLEU scores
Remarks: An extension of the original Swiss Parliaments Corpus
Publication Year: 2025
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Vincenzo Timmel, Manfred Vogel, Daniel Perruchoud, Reza Kakooee: *Swiss Parliaments Corpus Re-Imagined (SPC_R): Enhanced Transcription with RAG-based Correction and Predicted BLEU*. arXiv 2025.
Paper Link: Link
Contact: info@swissnlp.org

17: STT4SG-350 - Speech-to-Text for Swiss German-350

Reference: (Plüss et al., 2023)
Modalities: Speech
Languages: Swiss German
Description: Swiss German audio recordings with transcripts in Standard German. Balanced distribution across dialects and demographics such as gender. Collected via a crowdsourcing web app. Intended for ASR, TTS, and dialect identification. Dedicated test set with approx. 5 hours of audio of identical sentences spoken in 7 different dialects.
Size: 343 hours, 316 speakers
Annotations: Transcripts, dialect information, age group, gender.
Versions: Extension of Swiss Parliament Corpus SPC

Remarks: Same data format as SDS-200
Publication Year: 2023
License(s): META-SHARE NonCommercial NoRedistribution
Commercial Use: Yes
Website: Link
Reference: Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, Mark Cieliebak: *STT4SG-350: A Speech Corpus for All Swiss German Dialect Regions*. ACL 2023.
Paper Link: Link
Contact: info@swissnlp.org

18: Swiss Politics Corpus - Swiss Politics Corpus

Reference: (Salamanca, 2018)
Modalities: Text: Documents
Languages: German, French, Italian
Description: A database of who said what and when in both chambers of the Swiss parliament over the past 127 years, based on digitized proceedings with oldest documents being from 1891
Size: 40k documents
Annotations: labeled text lines and paragraphs
Publication Year: 2018
License(s): MIT License
Commercial Use: No
Website: Link
Reference: Luis Salamanca, Lilian Gasser, Laurence Brandenberger, Frank Schweitzer: *A trip through Swiss politics and history*. Blog Post 2018.
Paper Link: Link
Contact: Not available

19: Swiss SMS Corpus - Swiss SMS Corpus

Reference: (Stark et al., 2009-2015)
Modalities: Text: SMS
Languages: Swiss German, German, French, Italian, Romansh
Description: SMS messages crowdsourced from the Swiss public.
Size: 26k SMS messages, 650k tokens
Annotations: Language, PoS tags
Remarks: 41% Swiss German, 28% German, 18% French, 6% Italian, 4% Romansh
Publication Year: 2009
License(s): CC-NY-NC
Commercial Use: No
Website: Link
Reference: Elisabeth Stark, Simone Ueberwasser, Beni Ruef: *Swiss SMS Corpus*. 2009-2015.
Contact: sms@cl.uzh.ch

20: SwissCrawl - SwissCrawl Web Corpus

Reference: (Linder et al., 2020)
Modalities: Text: Web
Languages: Swiss German
Description: Large-scale, multilingual web crawl of the .ch domain.
Size: 560k sentences
Annotations: Crawling data
Remarks: 89% of sentences in Swiss German
Publication Year: 2020
License(s): CC BY-NC 4.0
Commercial Use: No
Website: Link
Reference: Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, Andreas FischerLucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, Andreas Fischer: *Automatic Creation of Text Corpora for Low-Resource Languages from the Internet: The Case of Swiss German*. LREC 2020.
Paper Link: Link
Contact: andreas.fischer@hefr.ch

21: SwissDial - Parallel Multidialectal Corpus of Spoken Swiss German

Reference: (Dogan-Schönberger et al., 2021)
Modalities: Speech
Languages: Swiss German
Description: Audio recordings of eight major Swiss German dialects and corresponding transcripts in Swiss German and Standard German
Size: 26 hours, 8 speakers
Annotations: Transcripts and dialect information
Versions: V1.0 from 2021, V1.1 contains additional 7726 recorded GR sentences.
Publication Year: 2021
License(s): Research use only, commercial use restricted
Commercial Use: No
Website: Link
Reference: Pelin Dogan-Schonberger, Julian

Mäder, Thomas Hofmann: *SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German*. arXiv 2021.
Paper Link: Link
Contact: Not available

22: SwissGPC - Swiss German Podcasts Corpus

Reference: (Stucki et al., 2025)
Modalities: Speech
Languages: Swiss German
Description: Links to Swiss podcast from Swiss TV broadcasters
Size: 5000 hours
Annotations: Transcripts, audio, language tags
Publication Year: 2025
License(s): CC BY 4.0 for Link Collection
Commercial Use: No
Website: Link
Reference: Samuel Stucki, Jan Deriu, Mark Cieliebak: *SwissGPC v1.0 - The Swiss German Podcasts Corpus*. SwissText 2025.
Paper Link: Link
Contact: deri@zhaw.ch

23: TEVOID - Temporal Voice Idiosyncrasy

Reference: (Dellwo et al., 2012)
Modalities: Speech, Text
Languages: Swiss German (Zurich)
Description: Recordings of read and spontaneous speech of different speakers of Zurich German for the research on idiosyncratic differences
Size: 16 sentences
Annotations: Transcripts, audio
Publication Year: 2012
License(s): Not specified
Commercial Use: Yes
Website: Link
Reference: Volker Dellwo, Adrian Leemann, Marie-José Kolly: *Speaker idiosyncratic rhythmic features in the speech signal*. Interspeech 2012.
Paper Link: Link
Contact: volker.dellwo@uzh.ch

24: TRANSLIT - Large-scale Name Transliteration Resource

Reference: (Benites et al., 2020)
Modalities: Text
Languages: 180 languages
Description: Variations of person and geolocation names in various languages.
Size: 1.6 million entries
Annotations: Sentiment labels
Publication Year: 2020
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, Mark Cieliebak: *TRANSLIT: A Large-scale Name Transliteration Resource*. LREC 2020.
Paper Link: Link
Contact: Not available

25: VarDial_GDI17 - German Dialect Identification at VarDial 2017

Reference: (Zampieri et al., 2017)
Modalities: Text
Languages: Swiss German
Description: Text containing dialects from the cantons of Basel (BS), Bern (BE), Lucerne (LU), and Zurich (ZH)
Size: 18k sentences
Annotations: Dialects
Publication Year: 2017
License(s): MIT License
Commercial Use: Unk.
Website: Link
Reference: Marcos Zampieri, Shervin Malmasi, Nikola Ljubesi, Preslav Nakov, Ahmed Ali, Jorg Tiedemann, Yves Scherrer, Noemi Aepli: *Findings of the VarDial Evaluation Campaign 2017*. VarDial 2017.
Paper Link: Link
Contact: simon.clematide@cl.uzh.ch

26: Walliserdeutsch ASR Corpus - ASR and Translation Corpus for Walliserdeutsch

Reference: (Garner et al., 2014)
Modalities: Speech
Languages: Walliserdeutsch, German
Description: Broadcast news from a local radio

station in Walliserdeutsch dialect.
Size: 8.1 hours
Annotations: Transcribed speech and translated text
Publication Year: 2014
License(s): Not specified
Commercial Use: Yes
Website: Link
Alternative Names: Walliserdeutsch ASR; Walliserdeutsch speech corpus
Reference: Philip N. Garner, David Imseng, Thomas Meyer: *Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch*. Interspeech 2014.
Paper Link: Link
Contact: dimseng@idiap.ch

27: WebClasSeg25 - WebClasSeg-25: A Dual-Classified Webpage Segmentation Dataset

Reference: (Gerber et al., 2025)
Modalities: Text, Image
Languages: 25 languages
Description: Webpages from public sector websites of Europe
Size: 2580 webpages
Annotations: Sentiment labels, Crawling Data, Screenshots
Publication Year: 2025
License(s): CC BY 4.0
Commercial Use: Yes
Website: Link
Reference: Jonathan Gerber, Jasmin Saxer, Kimia Rabishokr, Bruno Kreiner, Andreas Weiler: *WebClasSeg-25: A Dual-Classified Webpage Segmentation Dataset - Integrating Functional and Maturity-Based Analysis*. SIGIR 2025.
Paper Link: Link
Contact: jonathan.gerber@zhaw.ch

28: What's up, Switzerland? - Swiss Chat Corpus

Reference: (Ueberwasser and Stark, 2017)
Modalities: Text: Messages
Languages: Swiss German, German, French, Italian, Romansh, Spanish, Slavic languages
Description: WhatsApp chat messages, gathered in Summer 2014
Size: 760k messages, 5.1mio tokens
Annotations: Language tags
Publication Year: 2014
License(s): Non-commercial research only
Commercial Use: No
Website: Link
Reference: Simone Ueberwasser, Elisabeth Stark: *What's up, Switzerland? A corpus-based research project in a multilingual country*. Linguistik online 84/5, 2017.
Paper Link: Link
Contact: estark@rom.uzh.ch

29: ZHCORPUS - Zurich Corpus of Vowel and Voice Quality

Reference: (Maurer et al., 2018)
Modalities: Speech, Text
Languages: Swiss German (Zurich)
Description: Focused on sounds of the long Standard German vowels produced with varying basic production parameters
Size: 34k utterances, 70 speaker
Annotations: Transcripts, audio
Publication Year: 2018
License(s): Research use only, commercial use restricted
Commercial Use: Yes
Website: Link
Reference: Dieter Maurer, Christian d'Heureuse, Heidy Suter, Volker Dellwo, Daniel Friedrichs, Thayabaran Kathiresan: *The Zurich Corpus of Vowel and Voice Quality, Version 1.0*. Interspeech 2018.
Paper Link: Link
Contact: dieter.maurer@zhdk.ch

30: ZTC_BAS - Zurich Tangram Corpus

Reference: (Kalmanovitch, 2016)
Modalities: Speech
Languages: Swiss German (Zurich)
Description: Recordings of Swiss German dialect from Zurich, including transcriptions.
Size: 2 hours
Annotations: Transcripts, audio
Publication Year: 2019
License(s): Not specified
Commercial Use: Unk.
Website: Link
Reference: Yshai Kalmanovitch, Wolfgang Kesselheim: *The Zurich Tangram Corpus - BAS*

Edition. 2019.
Paper Link: Link
Contact: bas@bas.uni-muenchen.de

References

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak. 2020. TRANSLIT: A large-scale name transliteration resource. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3265–3271, Marseille, France. European Language Resources Association.

Paolo Canavese. 2019. LEX.CH.IT: A corpus for microdiachronic linguistic investigations of swiss normative acts in italian. *Comparative Legilinguistics*, 40(1):43–65.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

Volker Dellwo, Adrian Leemann, and Marie-José Kolly. 2012. Speaker idiosyncratic rhythmic features in the speech signal. Interspeech Conference Proceedings.

Pelin Dogan-Schönberger, Julina Mäder, and Thomas Hofmann. 2021. SwissDial: Parallel multidialectal corpus of spoken Swiss German.

Annarita Felici. 2025. CHEU-lex: a parallel multilingual corpus of Swiss and EU legislation. *Applied Corpus Linguistics*, 5(3):100151.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Philip N Garner, David Imseng, and Thomas Meyer. 2014. Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch. In *INTERSPEECH*, pages 2118–2122.

Jonathan Gerber, Jasmin Saxer, Kimia Rabishokr, Bruno Kreiner, and Andreas Weiler. 2025. Webclasseg-25: A dual-classified webpage segmentation dataset - integrating functional and maturity-based analysis. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3792–3801, New York, NY, USA. Association for Computing Machinery.

Ralf Grubenmann, Don Tuggener, Pius von Däniken, Jan Deriu, and Mark Cieliebak. 2018. SB-CH: A Swiss German corpus with sentiment annotations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stefan Höfler and Michael Piotrowski. 2011. Building corpora for the philological study of Swiss legal texts. *Journal for Language Technology and Computational Linguistics*, 26(2):77–89.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 85–94, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

David Imseng, Hervé Bourlard, Holger Caesar, Philip N Garner, Gwénolé Lecorvé, and Alexandre Nanchen. 2012. MediaParl: Bilingual mixed language accented speech database. In *2012 IEEE spoken language technology workshop (SLT)*, pages 263–268. IEEE.

Yshai Kalmanovitch. 2016. Speech in interaction–the Zurich Tangram corpus. *Tagungsband der Tagung Phonetik und Phonolotie im Deutschsprachigen Raum*, 12:79–81.

Adrian Leemann, Péter Jeszenszky, Carina Steiner, Jan Messerli, and Melanie Studerus. 2020.

Dolores Lemmenmeier-Batinić, Josip Batinić, and Anastasia Escher. 2023. Map Task Corpus of Heritage BCMS spoken by second-generation speakers in Switzerland. *Language Resources and Evaluation*, 57(4):1607–1644.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the Internet: The case of Swiss German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2706–2711, Marseille, France. European Language Resources Association.

Dieter Maurer, Christian d'Heureuse, Heidy Suter, Volker Dellwo, Daniel Friedrichs, and Thayabaran Kathiresan. 2018. The Zurich Corpus of vowel and voice quality, version 1.0.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 3250–3256, Marseille, France. European Language Resources Association.

Luis Salamanca. 2018. A trip through swiss politics and history. `https://www.datascience.ch/articles/trip-swiss-politics-history`. Accessed: 2025-10-17.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).

Elisabeth Stark, Simone Ueberwasser, and Beni Ruef. 2009-2015. www.sms4science.ch.

Samuel Stucki, Mark Cieliebak, and Jan Deriu. 2025. SwissGPC v1.0 – The Swiss German Podcasts Corpus. *arXiv preprint arXiv:2509.19866*.

Vincenzo Timmel, Manfred Vogel, Daniel Perruchoud, and Reza Kakooee. 2025. Swiss Parliaments Corpus Re-Imagined (SPC_R): Enhanced transcription with RAG-based correction and predicted BLEU. *arXiv preprint arXiv:2506.07726*.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

Simone Ueberwasser and Elisabeth Stark. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik online*, 84(5).

Malgorzata Anna Ulasik, Manuela Hürlimann, Fabian Germann, Esin Gedik, Fernando Benites, and Mark Cieliebak. 2020. CEASR: A corpus for evaluating automatic speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6477–6485, Marseille, France. European Language Resources Association.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.