A Survey of Corpora for Germanic Low-Resource Languages and Dialects

Anonymous Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymouser Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymousest Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

Despite much progress in recent years, the vast majority of work in natural language processing (NLP) is on standard languages with many speakers. In this work, we instead focus on low-resource languages and in particular non-standardized low-resource languages. Even within branches of major language families, often considered well-researched, little is known about extent and type of available resources and what the major NLP challenges are for these language varieties. The first step to address this situation is a systematic survey of available corpora (most importantly, annotated corpora, which are particularly valuable for NLP research). Focusing on Germanic low-resource language varieties, we provide such a survey in this paper. Except for geolocation (origin of speaker or document), we find that manually annotated linguistic resources are sparse and, if they exist, mostly cover morphosyntax. Despite this lack of resources, we observe that interest in this area is increasing: there is active development and a growing research community. To facilitate research, we make our overview of over 60 corpora publicly available.[1]

1 Introduction

The majority of current NLP today focuses on standard languages. Much work has been put forward in broadening the scope of NLP (Joshi et al., 2020), with long-term efforts pushing boundaries for language inclusion, for example in resource creation (e.g. Universal Dependencies (Zeman et al., 2022)) and in advancing the frontiers of cross-lingual transfer research (de Vries et al., 2022). However, even within major branches of language families or even single countries, plenty of language varieties are under-researched. For instance, local languages spoken in Italy (Ramponi, 2022), Indonesia (Aji et al., 2022) and Australia (Bird, 2020) are to name but a few.

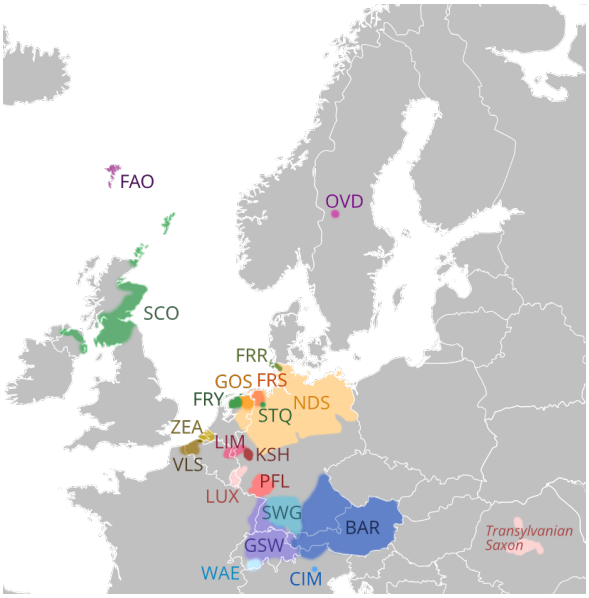

Figure 1: Approximate locations of most of the languages discussed in this article (not pictured: PDC, YID, NOR, SWE, ENG, DEU). Figure based on a map by Marian Sigler, CC BY-SA 3.0.

Current technology lacks methods to handle scarce data and the rich variability that comes with low-resource and non-standard languages. Nevertheless, interest in these under-resourced language varieties is growing. It is a topic of interest not only for (quantitative) dialectologists (Wieling and Nerbonne, 2015; Nerbonne et al., 2021), but also NLP researchers, as evidenced by the participation in specialized workshops like VarDial[2] and

[1] We share a companion website of this overview at *[URL omitted in review version]*.

[2] sites.google.com/view/vardial-2023

special interest groups for endangered[3] and under-resourced languages.[4]

In this paper, we provide an overview of the current state of NLP corpora for Germanic low-resource languages (LRLs) and dialects, with a particular focus on non-standard variants and four dimensions: annotation type, curation profile, resource size, and (written) data representation. We find that the amount and type of data varies by language, with manual annotations other than for morphosyntactic properties or the speaker's dialect/origin being especially rare. With this survey, we aim to support development of language technologies and quantitative dialectological analyses of Germanic low-resource languages and dialects, by making our results publicly available. Finally, based on the experiences we made while compiling this survey, we share recommendations for researchers releasing (or using) such datasets.

2 Related surveys

Zampieri et al. (2020) provide a recent overview of research on NLP for closely related language varieties and mention a few data sets. Recently, several surveys focusing on NLP tools and data for regional languages and dialects have been released: for local languages in Italy (Ramponi, 2022) and France (Leixa et al., 2014), Arabic dialects (Shoufan and Alameri, 2015; Younes et al., 2020; Guellil et al., 2021), creole languages (Lent et al., 2022), Irish English (Vaughan and Clancy, 2016), and spoken varieties of Slavic languages (Dobrushina and Sokur, 2022).

3 Language varieties

Our survey contains corpora for more than two dozen Germanic low-resource varieties, selected based on dataset availability.

This includes the languages associated with the ISO 639-3 codes FAO (Faroese), OVD (Elfdalian), SCO (Scots), FRR (North Frisian), FRY (West Frisian), STQ (Saterland Frisian), NDS (Low Saxon), FRS (East Frisian Low Saxon), GOS (Gronings), ZEA (Zeelandic), VLS (West Flemish), LTZ (Luxembourgish), LIM (Limburgish), KSH (Colognian), PFL (Palatine German), PDC (Pennsylvania Dutch), YID (Yiddish), BAR (Bavarian), SWG (Swabian), GSW (Swiss German and Alsatian), WAE (Walser), and CIM (Cimbrian). Our survey also encompasses data for dialects/non-standard varieties of Norwegian (NOR), Swedish (SWE), English (ENG), and German (DEU) that do not have their own ISO codes. We focus on specialized corpora showcasing regional variation, but not necessarily global variation. Figure 1 shows where most of the doculects included in this survey are spoken.

This overview does not include any corpora for Germanic-based creoles like Naija, as those are included in the recent survey by Lent et al. (2022).

4 Methodology

Similarly to Ramponi (2022), we search for corpora on several online repositories for language resources: the CLARIN Virtual Language Observatory (Van Uytvanck et al., 2010), the LRE Map (Calzolari et al., 2012), the European Language Grid (Rehm et al., 2020) OLAC (Simons and Bird, 2003), ORTOLANG (Pierrel et al., 2017), and the Hamburg Centre for Language Corpora.[5] We also search for corpora on Zenodo and on Google Dataset Search, and look for mentions of corpora in articles hosted by the ACL Anthology and on ArXiv.[6] We search for mentions of the word "dialect" and the names of various Germanic low-resource languages.

We use the following inclusion criteria:

- The corpus is accessible to researchers (immediately via a website, or indirectly through a request form or via contact information),[7] and this accessibility is indicated on the corpus website or in a publication accompanying the corpus.
- The corpus can be downloaded easily (does not require scraping a website) and does not require extensive pre-processing (we are interested in file formats like XML, TSV or TXT rather than PDF).
- The data are of a high quality, e.g. we ignore OCR'ed corpora that were not carefully cleaned.

[3] SIGEL, acl-sigel.github.io

[4] SIGUL, www.elra.info/en/sig/sigul

[5] vlo.clarin.eu; lremap.elra.info; live.european-language-grid.eu; www.language-archives.org; www.ortolang.fr/market/corpora; corpora.uni-hamburg.de/hzsk/en/repository-search

[6] zenodo.org; datasetsearch.research.google.com; aclanthology.org; arxiv.org

[7] The latter case is indicated with a lock 🔒 in the overview tables.

- The corpus (mainly) contains full sentences or utterances, and the data were (mainly) produced in the past century.

We base this survey only on the versions of corpora that are easily accessible to the research community; e.g. if a corpus contains audio material, but only the written material is available for download (and thus as a data source for quantitative research), the corpus is treated as a text corpus.[8]

5 Corpora

Most of the language varieties we survey have no or only a very recent written tradition. This unique challenge is reflected in the different written formats used to represent the data (if the corpora contain any written material at all) and concerns both the transcription of audio data (Tagliamonte, 2007; Gaeta et al., 2022) as well as the elicitation of written data (Millour and Fort, 2020). We opted to discern between audio data 🎙 and the following written variants: standard orthographies (of the doculects themselves where existing **A** (e.g. West Frisian orthography), or of a closely related higher-resource language otherwise ¶), ad-hoc pronunciation spelling (by speakers of the doculect) ✎, and phonetic or phonemic transcriptions (by linguists) ✍. Appendix A provides examples.

The following corpora are sorted by annotation and curation type. For an overview sorted by language, see Appendix B. Some of the corpora share the same data sources. Appendix C lists the cases where we are aware of such overlaps.

5.1 Annotated corpora

This section only includes corpora with manual (or manually corrected) annotations.

5.1.1 Morphosyntactic annotation

Table 1 provides an overview of datasets with morphosyntactic annotations. These mostly contain part-of-speech (POS) tags and/or syntactic dependencies. Such annotations are, for instance, of interest to dialectologists studying morphosyntactic variation (see for example Lybaert et al., 2019). Automatically generating high-quality morphosyntactic annotations for non-standard and/or low-resource data is not trivial, and the more annotated data are available for training, the better the results (Schulz and Ketschik, 2019; Scherrer et al., 2019a).

The annotation standards tend to either be general and cross-linguistically applicable (inviting comparisons between languages), or to be very specific to the language variety at hand. In the former case, annotations follow the guidelines from the Universal Dependencies project (Zeman et al., 2022) (UD, UPOS).[9] In the latter case, tag sets created for a (usually closely related) higher-resource language are modified so that they capture the lower-resource language variety's idiosyncrasies. These specialized tag sets are based on: the annotations of the Giellatekno project (Wiechetek et al., 2022), the annotations developed for the Penn Parsed Corpora of Historical English (Penn-h),[10] the tag set of the Norwegian Dependency Treebank (NDT) (Solberg et al., 2014) (based on the Oslo-Bergen Tagger's tag set (OBT) (Johannessen et al., 2012)) and the Stuttgart-Tübingen tag set (STTS) (Schiller et al., 1999).

Most of the annotated corpora are presented only in one written form, typically either written in a standard orthography or pronunciation spelling. Some cases (marked with an asterisk* in the table) require further explanation: The Norwegian LIA and NDC treebanks (Øvrelid et al., 2018; Kåsen et al., 2022) use normalized orthographies (Nynorsk and Bokmål, respectively), but aligned versions of the original phonetic and orthographic transcriptions can be downloaded from the Tekstlab links in the table. The sentences in the UD Low Saxon LSDC treebank (Siewert et al., 2021) are presented both in the original ad-hoc pronunciation spelling and in a recently proposed orthography for Low Saxon, *Nysassiske Skryvwyse*. The Yiddish corpus (Santorini, 2021) is romanized, partially according to the YIVO transliteration system, and partially in a non-systematic manner.

5.1.2 Semantic annotation and parallel sentences

Very few resources with other types of annotations exist; we were only able to find four (Table 2). These all have very different kinds of annotations:

[8] This is not a rare scenario, as the audio versions might intrinsically contain more personally identifying information (like the voice of someone who is part of a small speaker population), and it also requires more work to censor e.g. locations or personal names in audio data than in text data.

[9] universaldependencies.org/guidelines.html

[10] ling.upenn.edu/hist-corpora/annotation/index.html

Corpus	Langs	Annotation	Size	Rep.
UD Faroese OFT (Tyers et al., 2018) `github.com/UniversalDependencies/UD_Faroese-OFT`	FAO	POS (UPOS, Giellatekno-FAO), dep (UD), morpho (UD) lemmas	1.2k sents	A

FarPaHC / UD Faroese FarPaHC
(Ingason et al., 2012; Rögnvaldsson et al., 2012)
`hdl.handle.net/20.500.12537/92`
`github.com/UniversalDependencies/UD_Faroese-FarPaHC`
	FAO	POS (mod. Penn-h, UPOS),
		phrase struc.(mod. Penn-h),
		dep (UD), morpho (UD)
			53k (FarP.) /
			40k (UD.) toks
				A

LIA Treebank / UD Nor. NynorskLIA
(Øvrelid et al., 2018)
`tekstlab.uio.no/LIA/norsk/index_english.html` `github.com/UniversalDependencies/UD_Norwegian-NynorskLIA`
`github.com/textlab/spoken_norwegian_resources/tree/master/treebanks/Norwegian-NynorskLIA`
	NOR 📍	POS (UPOS, mod. NDT),
		dep (UD, mod. NDT), lemmas,
			77k toks (LIA),
			55 toks / 5.3k sents (UD)
				¶ ✎*

NDC Treebank
(Kåsen et al., 2022; Johannessen et al., 2009)
`tekstlab.uio.no/scandiasyn/download.html`
`github.com/textlab/spoken_norwegian_resources/tree/master/treebanks/Norwegian-BokmaalNDC`
	NOR 📍	POS (mod. NDT),
		dep (mod. NDT), lemmas,
		morpho (mod. NDT)
			66k toks	¶ ✎*

NorDial (subset) (Mæhlum et al., 2022) Contact authors 🔒	NOR	POS (UPOS)	35+ tweets	✎
TwitterAAE-UD (Blodgett et al., 2016) `slanglab.cs.umass.edu/TwitterAAE`	ENG (AAVE)	Dep (UD)	250 tweets	✎
UD Low Saxon LSDC (Siewert et al., 2021) `github.com/UniversalDependencies/UD_Low_Saxon-LSDC`	NDS 📍	POS (UPOS), dep (UD), morpho (UD), glosses (GML) lemmas	95 sents	✎ ¶*
UD Frisian/Dutch Fame (Braggaar and van der Goot, 2021; Yilmaz et al., 2016) `github.com/UniversalDependencies/UD_Frisian_Dutch-Fame`	FRY/NLD	POS (UPOS), dep (UD)	400 sents code-switching	A
Stemmen uit het verleden (annotated subset) (Lybaert et al., 2019; Van Keymeulen et al., 2019)	VLS 📍	V2 variation `doi.org/10.18710/NSFN2B`	1.4k sents	✎
Annotated Corpus for the Alsatian Dialects (Bernhard et al., 2018, 2019) `zenodo.org/record/2536041#.YvIZrvFByKc`	GSW (Alsatian)	POS (UPOS, mod. UPOS), lemmas, glosses (FRA)	798 sents	✎
BISAME (STIH, 2020; Millour and Fort, 2018) `hdl.handle.net/11403/bisame_gsw/v1`	GSW (Alsatian)	POS (mod. UPOS)	382 sents	✎
NOAH's corpus (Hollenstein and Aepli, 2015) `noe-eva.github.io/NOAH-Corpus`	GSW	POS (mod. STTS, subset: UPOS and STTS)	115k toks	✎
UD Swiss German UZH (Aepli and Clematide, 2018) `github.com/UniversalDependencies/UD_Swiss_German-UZH`	GSW	POS (UPOS, mod. STTS)	100 sents	✎
WUS_DIALOG_GSW (subset of *What's up, Switzerland?*) (Stark et al.; Ueberwasser and Stark, 2017)	GSW 📍	POS (mod. STTS) `whatsup.linguistik.uzh.ch` 🔒	34.7k toks	✎ ¶
Kontatto (Dal Negro and Ciccolone, 2020) `kontatti.projects.unibz.it` 🔒	BAR (South Tyrol)	POS (unknown), lemmas (DEU)	147k toks	🎤 ✎
Penn Parsed Corpus of Historical Yiddish (Santorini, 2021) `github.com/beatrice57/penn-parsed-corpus-of-historical-yiddish`	YDD	POS (Penn-h), phrase struc. (Penn-h)	ca. 200k toks	*

Table 1: **Morphosyntactically annotated corpora.** Abbreviations for the annotation tag sets are explained in Section 5.1.1, as are the orthographies of entries with an asterisk (*). Other abbreviations and symbols: *Rep.* = 'data representation,' *dep* = 'syntactic dependencies,' *phrase struc* = 'phrase structure,' *morpho* = 'morphological features,' *mod.* = 'modified,' *AAVE* = 'African-American Vernacular English,' *GML* = 'Middle Low Saxon,' *NLD* = 'Dutch,' *FRA* = 'French,' 🔒 = access is not immediate, 📍 = fine-grained dialect distinctions, ✎ = phonetic/phonemic transcription, ✎ = pronunciation spelling, A = LRL orthography, ¶ = normalized orthography.

Corpus	Langs	Annotation	Size	Rep.
TaPaCo (subset) (Scherrer, 2020) zenodo.org/record/3707949	NDS, GOS	paraphrases	1107 sents (NDS), 122 sents (GOS)	✎
SwissDial (Dogan-Schönberger et al., 2021) projects.mtc.ethz.ch/swiss-voice-data-collection 🔒	GSW 📍	topic, translations (across dialects and into DEU)	2.5–4.6 hrs × 8 dialects	🎙✎¶
SB-CH (subset) (Grubenmann et al., 2018) github.com/spinningbytes/SB-CH	GSW	sentiment	2.8k sents	✎
xSID (subset) (van der Goot et al., 2021) bitbucket.org/robvanderg/xsid	BAR (South Tyrol)	slot and intent detection, translations (12 langs)	800 sents	✎

Table 2: **Corpora with semantic annotations or parallel sentences.** Abbreviations and symbols: *Rep.* = 'data representation,' 🔒 = access is not immediate, 📍 = fine-grained dialect distinctions, 🎙 = audio, ✎ = pronunciation spelling, ¶ = standard orthography.

sentiment or topic classification, intent detection and slot-filling, and sentence-level translations or paraphrases.

5.1.3 Dialect annotation

Many corpora contain detailed annotations on the dialect area (or more precise location) an utterance's speaker or the author of a document is from. Such information is important for linguistic research comparing related dialects (Wieling and Nerbonne, 2015), for comparing the results of traditional and quantitative dialectological approaches (e.g. Heeringa et al., 2009) and for evaluating whether NLP systems perform differently on different closely related language varieties (Ziems et al., 2022). Since corpora with such annotations cross-cut all of the categories of curated datasets in this survey, they are not presented on their own, but instead marked with a pin symbol 📍 elsewhere.

5.2 Other curated corpora

5.2.1 Text corpora

Table 3 presents written corpora of low-resource languages like Elfdalian or Faroese, or corpora that showcase dialectal variation through phonetic transcriptions or pronunciation spelling. (While variation also occurs on linguistic levels encoded in normalized text written in standard orthographies – lexical, syntactic or pragmatic variation – we focus on phonological variation, as this is where specialized corpora are required.)

5.2.2 Audio corpora

In this survey, our focus lies on written resources, and as such, this selection of audio corpora is not exhaustive. However, many of the language varieties surveyed in this article are predominately spoken rather than written. Creating language technology for unwritten languages is a topic of interest for NLP researchers (Scharenborg et al., 2020), and this is also reflected by the number of recently created speech corpora for Germanic low-resource languages.

Many of the audio corpora (Table 4) fall into one of two categories: recordings created for dialectological research, and post-hoc collections of already existing audio data (like radio broadcasts or public recordings of council meetings). Most of the audio corpora are transcribed, typically according to a standard orthography.

5.3 Uncurated text corpora

A final type of corpus are uncurated text collections (Table 5). This includes data coming from community-based data collection efforts unrelated to research projects (Wikipedia, Tatoeba) and web-crawled data. It is important to note that there are quality issues with web-crawled corpora, especially for low-resource languages (Kreutzer et al., 2022).[11] Both CC-100 (Wenzek et al., 2020) and OSCAR (Abadji et al., 2022) are cleaned versions of CommonCrawl[12] – and Abadji et al. (2022) specifically remark on the low quality of the low-resource language data in that dataset.

Wikipedia has editions in many Germanic low-resource languages (see Appendix D), and several projects that extend wiki dumps with automatically inferred annotations have been released (Pan et al., 2017; Schwenk et al., 2021). However, the linguistic quality of the LRL wikis is not always very high. For instance, the Scots Wikipedia made the news in 2020, when attention was brought to

[11] However, see Artetxe et al. (2022) for an argument that the linguistic quality of a corpus might not be the most important factor for all downstream applications.

[12] commoncrawl.org

Corpus	Langs	Size	Rep.
Føroyskur talumálsbanki (Jacobsen, 2022) clarino.uib.no/corpuscle-classic/corpus-list 🔒	FAO	599.9k toks	A
BLARK 1.0 (Background text corpus) (Simonsen et al., 2022) (incl. FTS (Språkbanken and Fróðskaparsetur Føroya) and Faroese Korp (Giellatekno)) maltokni.fo/en/resources	FAO	25M toks	A
Nordic Dialect Corpus (Johannessen et al., 2009) tekstlab.uio.no/nota/scandiasyn	NOR 📍, OVD 📍	1.9M toks (NOR), 15.7k toks (OVD)	¶ (NOR: ✎) (OVD: A)
Talemålsundersøkelsen i Oslo (TAUS) (Tekstlab, 2020) tekstlab.uio.no/nota/taus/	NOR (East/West Oslo) 📍	388k toks	✎ ¶
NorDial (Barnes et al., 2021) github.com/jerbarnes/nordial	NOR	348 tweets	✎
American Nordic Speech Corpus (CANS) (Johannessen, 2015) tekstlab.uio.no/norskiamerika/korpus.html	NOR (US/Canada) 📍, SWE (US) 📍	773k toks (NOR), 46k toks (SWE)	✎ ¶
Scottish Corpus of Texts & Speech (SCOTS) (subset) (Anderson et al., 2007) scottishcorpus.ac.uk	SCO	(unknown how many of 4.6M toks in SCO)	mix of ✎ ¶
Low Saxon Dialect Classification (LSDC) (Siewert et al., 2020) github.com/Helsinki-NLP/LSDC/	NDS 📍	88.9k sents	✎
LuxId (Lavergne et al., 2014) lrec2014.lrec-conf.org/en/shared-lrs/current-list-shared-lrs	LTZ/DEU/FRA code-switching	924 sents (most with LTZ content)	A
DiDi (subset) (Frey et al., 2019) hdl.handle.net/20.500.12124/7	BAR (South Tyrol)	unknown	✎
What's up, Switzerland? (Stark et al.; Ueberwasser and Stark, 2017) whatsup.linguistik.uzh.ch 🔒	GSW 📍	507k msgs / 3.6M toks	✎
Text+Berg (subset) (Bubenhofer et al., 2015) http://textberg.ch/site/en/corpora 🔒	GSW	156 sents / 3.1k toks	✎
ArchiWals / CLiMAlp (Angster et al., 2017; Gaeta, 2020) climalp.org 🔒	WAE 📍	80+k tokens	✎

Table 3: **Other curated text corpora.** Abbreviations and symbols: *Rep.* = 'data representation,' 🔒 = access is not immediate, 📍 = fine-grained dialect distinctions, ✎ = phonetic/phonemic transcription, ✎ = pronunciation spelling, A = LRL orthography, ¶ = normalized orthography.

the fact that half of that wiki's articles had been created or edited by a non-Scots speaker writing in a parody of Scots (Brooks and Hern, 2020). Such quality issues should be taken into account when working with data from small wikis without much oversight, especially with data or tools based on the Scots Wikipedia before any clean-up could be started in fall 2020.[13]

6 Outlook

Creating NLP resources and technology for LRLs is an active field. Several additional resources are currently under construction or revision: *UD Frisian Frysk*, a treebank for West Frisian (Heeringa et al., 2021),[14] *Boarnsterhim Corpus*, a West Frisian audio corpus (Sloos et al., 2018),[15] *Schweizerdeutsches Mundartkorpus*, a Swiss German text corpus (Weibel and Peter, 2020),[16] and the *parsed corpus of Southern Dutch Dialects* (Breitbarth et al.).[17] Community-based projects are also being actively developed: many of the small Wikipedias have active editors (Appendix D), as do many of the Tatoeba collections.

Speaker populations of LRLs are not a monolith. Accordingly, different speaker communities have different interests in terms when it comes to the development of language technologies (Lent et al., 2022), and the creation of downstream technologies made for public use should be made in accordance of the wishes and needs of the relevant speaker communities (see also Bird, 2022).

[13] E.g., Scots is included in the language list of mBERT (Devlin et al., 2019), which was trained on Wikipedia data in 2019: github.com/google-research/bert/blob/master/multilingual.md

[14] github.com/UniversalDependencies/UD_Frisian-Frysk

[15] taalmaterialen.ivdnt.org/download/tstc-boarnsterhimcorpus1-0

[16] chmk.ch/de/info_all

[17] gcnd.ugent.be

Corpus	Langs	Size	Rep.
BLARK 1.0 (Transcribed recordings) (Simonsen et al., 2022) maltokni.fo/en/resources	FAO 📍	100 hrs	🎙 **A** (partially ✎)
Faroese Danish Corpus Hamburg (FADAC Hamburg) (subset) (Debess, 2019) corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:fadac-0.2.dan	FAO 📍	31 hrs	🎙 **A**
NB Tale – Speech Database for Norwegian (Språkbanken) nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-31/	NOR 📍	365 × 2 mins (spon.), 7.6k sents (reading)	🎙 ✎ ¶
Norwegian Parliamentary Speech Corpus (NPSC) (Solberg and Ortiz, 2022) nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-58/	NOR 📍	540 hrs	🎙 ¶
Diachronic Electronic Corpus of Tyneside English (DECTE) (Corrigan et al., 2012) research.ncl.ac.uk/decte/index.htm 🔒	ENG (UK: Tyneside)	72 hrs / 804k toks	🎙 ¶ (partially ✎)
Intonational Variation in English (IViE) (Nolan and Post, 2014) phon.ox.ac.uk/files/apps/IViE/	ENG (UK, Ireland) 📍	36 hrs	🎙 ¶
Crowdsourced high-quality UK and Ireland English Dialect speech data set (Demirsahin et al., 2020) openslr.org/83	ENG (UK, Ireland) 📍	31 hrs	🎙 ¶
Helsinki Corpus of British English Dialects (University of Helsinki, 2006) varieng.helsinki.fi/CoRD/corpora/Dialects/ 🔒	ENG (UK) 📍	1M toks	🎙 ¶
Nationwide Speech Project (NSP) (Clopper and Pisoni, 2006) u.osu.edu/nspcorpus	ENG (US) 📍	60 × 1 hr	🎙 (partially ¶)
Common Voice Corpus 12.0 (Frisian subset) (Ardila et al., 2020) commonvoice.mozilla.org/en/datasets	FRY	150 hrs	🎙 **A**
Frisian AudioMining Enterprise (FAME) (Yilmaz et al., 2016) ru.nl/clst/tools-demos/datasets/ 🔒	FRY (partially 📍)	18.5 hrs	🎙 **A**
Recordings of Dutch-Frisian council meetings (Bentum et al., 2022) frisian.eu/dutchfrisiancouncilmeetings	FRY	26 hrs / 281k toks	🎙 **A**
Corpus Spoken Frisian (Frisian Academy) www1.fa.knaw.nl/ksf.html 🔒	FRY	200 hrs (total); 65 hrs (transcribed)	🎙 (partially **A**)
Sprachvariation in Norddeutschland (SiN, Hamburg collection) (Schröder, 2011; Elmentaler et al., 2015) hdl.handle.net/11022/0000-0000-7EE3-3 🔒	NDS, FRS, DEU	300 hrs	🎙
Regional Variants of German 1 (RVG1) (Burger and Schiel, 1998) hdl.handle.net/11022/1009-0000-0004-3FF4-3	DEU* 📍	500+ × 1 min	🎙 ✎ ¶
The Texas German Sample Corpus (TGSC) (Blevins, 2022) doi.org/10.18738/T8/IOX9ZA	DEU (Texas)	13.5 hrs / 75k toks	🎙 ¶
Audioatlas Siebenbürgisch-Sächsischer Dialekte (University of Munich) hdl.handle.net/11022/1009-0000-0001-27B9-3 🔒	DEU (Trans. Saxon)**	360 hrs / 450k toks	🎙 ¶ (partially ✎)
Kontatti (Ghilardi, 2019) kontatti.projects.unibz.it 🔒	BAR (South Tyrol), CIM	unknown	🎙 ¶
ArchiMob (Scherrer et al., 2019b) spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html (audio files: 🔒)	GSW 📍	70 hrs	🎙 ✎ ¶
SDS-200 (Plüss et al., 2022) swissnlp.org/datasets/ 🔒	GSW	200 hrs	🎙 ¶
Swiss Parliaments Corpus (Plüss et al., 2021a) www.cs.technik.fhnw.ch/i4ds-datasets	GSW	293 hrs	🎙 ¶
All Swiss German Dialects Test Set (Plüss et al., 2021b) cs.technik.fhnw.ch/i4ds-datasets	GSW 📍	13 hrs / 5.8k utterances	🎙 ¶
Gemeinderat Zürich Audio Corpus (Plüss et al., 2021b) www.cs.technik.fhnw.ch/i4ds-datasets	GSW	1208 hrs	🎙
Walliserdeutsch/RRO (Garner, 2014; Garner et al., 2014) zenodo.org/record/4580286 🔒	WAE	8.3 hrs	🎙 ✏
CABank Yiddish Corpus (Newman, 2015) ca.talkbank.org/access/Yiddish.html	YID (New York)	1 hr	🎙 ✎

Table 4: **Other audio corpora.** Abbreviations and symbols: *Rep.* = 'data representation,' 🔒 = access is not immediate, 📍 = fine-grained dialect distinctions, 🎙 = audio, ✎ = phonetic/phonemic transcription, ✏ = pronunciation spelling, **A** = LRL orthography, ¶ = normalized orthography. *It is unclear whether the RVG1 recordings are in regionally accented (Standard) German or whether they are in Low Saxon, Bavarian and other regional languages spoken in Germany, Switzerland, Austria and Northern Italy. **Transylvanian Saxon is a variety of Moselle Franconian that does not have its own ISO code. It is more closely related to Luxembourgish than to Standard German.

Corpus	Langs	Size
SwissCrawl (Linder et al., 2020) `icosys.ch/swisscrawl` 🔒	GSW	563k sents
SB-CH (Grubenmann et al., 2018) `github.com/spinningbytes/SB-CH` 🔒	GSW	ca. 116k sents
SwigSpot (Linder, 2018) `github.com/derlin/SwigSpot_Schwyzertuutsch-Spotting`	GSW	8k sents
Tatoeba (subset; with > 100 sents) `tatoeba.org`	NDS, YDD, GOS, FRR, SWG, LTZ, FRY, GSW, FAO, BAR	18.1k (NDS), 12.8k (YDD), 5.7k (GOS), 2.9k (FRR), 1.9k (SWG), 884 (LTZ), 641 (FRY), 474 (GSW), 417 (FAO), 227 (BAR) sents
Web to Corpus (W2C) (subset) (Majliš, 2011; Majliš and Žabokrtský, 2012) `hdl.handle.net/11858/00-097C-0000-0022-6133-9`	YID, FAO, LTZ, FRY, SCO, NDS, LIM	125 (YID), 102 (FAO), 81 (LTZ), 72 (FRY), 35 (SCO), 24 (NDS), 20 (LIM) MB
CC-100 (subset) (Wenzek et al., 2020) `data.statmt.org/cc-100/`	FRY, YID, LIM	174 MB (FRY), 51 MB (YID), 8.3 MB (LIM)
OSCAR (subset) (Abadji et al., 2022) `oscar-project.org/post/oscar-v22-01/` 🔒	YID, FRY, LTZ, NDS, GSW	15.8M (YID), 6.4M (FRY), 1.5M (LTZ), 1.0M (NDS), 37.4k (GSW) toks
Wikipedia (subset) `dumps.wikimedia.org`	see Appendix D	see Appendix D

Table 5: **Uncurated corpora.** 🔒 = Access is not immediate.

6.1 Recommendations

We make the following recommendations for researchers who work with LRL datasets:

- Investigate the quality of uncurated data, as it might be especially poor for LRLs.
- Check whether (pre-)training, development and test data are truly from independent datasets – the dearth of high-quality LRL data means that datasets may be likely to overlap.
- Consider quantitative work by dialectologists and sociolinguists who might not publish in typical NLP venues.

To researchers who create such datasets, we recommend to:

- Document the transcription principles (if the data were originally in an audio format) / if any standardized orthographies were used (if the language variety does not have an official orthography)
- The low number of available high-quality datasets per language variety means that the impact of losing such a resource is much greater. Therefore, please upload your corpus to an archive geared towards long-term data storage (like the CLARIN Language Resource Inventory,[18] the LRE Map or Zenodo).
- Provide easy-to-find documentation with details on the corpus size, data sources and the annotation procedure.

7 Conclusion

We have presented an analysis of over 60 corpora containing data in Germanic low-resource languages, with a focus on non-standardized or only recently standardized varieties. We additionally share the corpus overview on a public companion website (*[URL omitted in review version]*) that can easily be updated as more language resources are released.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Noëmi Aepli and Simon Clematide. 2018. Parsing approaches for Swiss German. In *Proceedings of the 3rd Swiss Text Analytics Conference (SwissText), Winterthur, Switzerland*.

[18] `clarin.eu/content/language-resource-inventory`

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Jean Anderson, Dave Beavan, and Christian Kay. 2007. SCOTS: Scottish corpus of texts and speech. *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, 1: Synchronic Databases:17–34.

Marco Angster, Marco Bellante, Raffaele Cioffi, and Livio Gaeta. 2017. I progetti DiWaC e ArchiWals. *Bollettino dell'Atlante Linguistico Italiano*, 41:83–94.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390. Association for Computational Linguistics.

Jeremy Barnes, Petter Mæhlum, and Samia Touileb. 2021. NorDial: A preliminary corpus of written Norwegian dialect use. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 445–451, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Martijn Bentum, Louis ten Bosch, Henk van den Heuvel, Simone Wills, Domenique van der Niet, Jelske Dijkstra, and Hans Van de Velde. 2022. A speech recognizer for Frisian/Dutch council meetings. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1009–1015, Marseille, France. European Language Resources Association.

Delphine Bernhard, Pascale Erhart, Dominique Huck, and Lucie Steiblé. 2019. Annotated corpus for the Alsatian dialects. Zenodo. Version 2.0.

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steible, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille, and Thomas Lavergne. 2018. Corpora with part-of-speech annotations for three regional languages of France: Alsatian, Occitan and Picard. In *11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Margaret Blevins. 2022. Texas German sample corpus. Texas Data Repository.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58, Kyiv, Ukraine. Association for Computational Linguistics.

Anne Breitbarth, Melissa Farasyn, Anne-Sophie Ghyselen, and Jacques Van Keymeulen. Het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten. *Handelingen – Koninklijke Zuid-Nederlandse maatschappij voor taal- en letterkunde en geschiedenis*, 72:23–38.

Libby Brooks and Alex Hern. 2020. Shock an aw: US teenager wrote huge slice of Scots Wikipedia. The Guardian.

Noah Bubenhofer, Martin Volk, Fabienne Leuenberger, and Daniel Wüest. 2015. Text+Berg–Korpus (release 151 v01). Digitale Edition des Jahrbuch des SAC 1864-1923, Echo des Alpes 1872-1924, Die Alpen, Les Alpes, Le Alpi 1925-2014, The Alpine Journal 1969-2008. Institut für Computerlinguistik, Universität Zürich.

Susanne Burger and Florian Schiel. 1998. RVG 1 – A database for regional variants of contemporary German. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pages 1083–1087, Granada, Spain.

Nicoletta Calzolari, Riccardo Del Gratta, Gil Francopoulo, Joseph Mariani, Francesco Rubino, Irene

Russo, and Claudia Soria. 2012. The LRE map. harmonising community descriptions of resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1084–1089, Istanbul, Turkey. European Language Resources Association (ELRA).

Cynthia G. Clopper and David B. Pisoni. 2006. The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*, 48.

Karen P. Corrigan, Isabelle Buchstaller, Adam Mearns, and Hermann Moisl. 2012. The diachronic electronic corpus of Tyneside English. Newcastle University.

Silvia Dal Negro and Simone Ciccolone. 2020. KONTATTO: A laboratory for the study of language contact in South Tyrol. *Sociolinguistica*, 34(1):241–247.

Iben Nyholm Debess. 2019. FADAC Hamburg 1.0. guide to the Faroese Danish corpus Hamburg. *Kieler Arbeiten zur skandinavistischen Linguistik*, 6.

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. 2020. Open-source multi-speaker corpora of the English accents in the British Isles. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6532–6541, Marseille, France. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nina Dobrushina and Elena Sokur. 2022. Spoken corpora of Slavic languages. *Russian Linguistics*, 46:77–93.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. SwissDial: Parallel multidialectal corpus of spoken Swiss German. *Computing Research Repository*, arXiv:2103.11401.

Michael Elmentaler, Joachim Gessinger, Jens Lanwer, Peter Rosenberg, Ingrid Schröder, and Jan Wirrer. 2015. Sprachvariation in Norddeutschland (SiN). In Roland Kehrein, Alfred Lameli, and Stefan Rabanus, editors, *Regionale Variation des Deutschen*, pages 397–424. De Gruyter.

Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W. Stemle. 2019. DIDI - the DiDi corpus of South Tyrolean CMC 1.0.0. Eurac Research CLARIN Centre.

Frisian Academy. Corpus spoken Frisian. Department of Social Sciences (Frisian Academy) and Department of Linguistics (Frisian Academy).

Livia Gaeta. 2020. The Observer's Paradox meets corpus linguistics: Written and oral sources for the Walser linguistic islands in Italy. In *Endangered linguistic varieties and minorities in Italy and the Balkans*, Vienna. VLACH.

Livio Gaeta, Marco Angster, Raffaele Cioffi, and Marco Bellante. 2022. Corpus linguistics for low-density varieties. minority languages and corpus-based morphological investigations. *Corpus*, 23.

Philip Garner. 2014. Walliserdeutsch. Zenodo.

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

Marta Ghilardi. 2019. Eliciting comparable spoken data in minor languages: first observations from the corpus Kontatti. *Suvremena lingvistika*, 45(88):231–246.

Giellatekno. Korp.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

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).

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University – Computer and Information Sciences*, 33(5):497–507.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2022. Glottolog 4.7. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://glottolog.org, accessed on 2023-02-03.

Wilbert Heeringa, Gosse Bouma, Martha Hofman, Eduard Drenth, Jan Wijffels, and Hans Van de Velde. 2021. POS tagging, lemmatization and dependency parsing of West Frisian.

Wilbert Heeringa, Keith Johnson, and Charlotte Gooskens. 2009. Measuring Norwegian dialect distances using acoustic features. *Speech Communication*, 51(2):167–183.

Nora Hollenstein and Noëmi Aepli. 2015. A resource for natural language processing of Swiss German dialects. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2015*, pages 108–109.

Anton Karl Ingason, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Joel C. Wallenberg. 2012. Faroese parsed historical corpus (FarPaHC) 0.1. CLARIN-IS.

Jógvan í Lon Jacobsen. 2022. Flertalsformer af ari-ord i den færøske talesprogsbank. *Nordlyd*, 46(1):103–113.

Janne Bondi Johannessen. 2015. The corpus of American Norwegian speech (CANS). In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 297–300, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Janne Bondi Johannessen, Kristin Hagen, André Lynum, and Anders Nøklestad. 2012. OBT+ stat. a combined rule-based and statistical tagger. In Gisle Andersen, editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, volume 49 of *Studies in Corpus Linguistics*, page 51. John Benjamins Publishing.

Janne Bondi Johannessen, Joel James Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus–an advanced research tool. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 73–80, Odense, Denmark. Northern European Association for Language Technology (NEALT).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug. 2022. The Norwegian Dialect Corpus treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4827–4832, Marseille, France. European Language Resources Association.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Thomas Lavergne, Gilles Adda, Martine Adda-Decker, and Lori Lamel. 2014. Automatic language identity tagging on word and sentence-level in multilingual text sources: A case-study on Luxembourgish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3300–3304, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jérémy Leixa, Valérie Mapelli, and Khalid Choukri. 2014. Inventaire des ressources linguistiques des langues de France. Version 1.1. Evaluations and Language resources Distribution Agency (ELDA).

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Lucy Linder. 2018. SwigSpot – creation of a Swiss German dataset. Master's thesis, University of Applied Sciences and Arts Western Switzerland.

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 12th Language Resources and Evaluation Conference*, pages 2706–2711, Marseille, France. European Language Resources Association.

Chloé Lybaert, Bernard De Clerck, Jorien Saelens, and Ludovic De Cuypere. 2019. A corpus-based analysis of V2 variation in West Flemish and French Flemish dialects. *Journal of Germanic Linguistics*, 31(1):43–100.

Petter Mæhlum, Andre Kåsen, Samia Touileb, and Jeremy Barnes. 2022. Annotating Norwegian language varieties on Twitter for part-of-speech. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 64–69, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Martin Majliš. 2011. W2C – Web to Corpus – corpora. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Martin Majliš and Zdeněk Žabokrtský. 2012. Language richness of the web. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2927–2934, Istanbul, Turkey. European Language Resources Association (ELRA).

Alice Millour and Karën Fort. 2018. À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. *Revue TAL*.

Alice Millour and Karën Fort. 2020. Text corpora and the challenge of newly written languages. In *1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*, Proceedings of the 1st Joint SLTU and CCURL Workshop, Marseille, France.

John Nerbonne, Wilbert Heeringa, Jelena Prokić, and Martijn Wieling. 2021. Dialectology for computational linguists. In Marcos Zampieri and Preslav Nakov, editors, *Similar Languages, Varieties, and Dialects: A Computational Perspective*, Studies in Natural Language Processing, pages 96–118. Cambridge University Press.

Zelda Kahan Newman. 2015. Discourse markers in the narratives of New York Hasidim. more V2 attrition. In Janne Bondi Johannessen and Joseph C. Salmons, editors, *Germanic heritage languages in North America. Acquisition, attrition and change*, pages 178–197. John Benjamins.

Francis Nolan and Brechtje Post. 2014. The IViE Corpus. In *The Oxford Handbook of Corpus Phonology*. Oxford University Press.

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The LIA treebank of spoken Norwegian dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jean-Marie Pierrel, Christophe Parisse, Jérôme Blanchard, Etienne Petitjean, and Frédéric Pierre. 2017. ORTOLANG: a French infrastructure for Open Resources and TOols for LANGuage. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, volume 136, pages 102–112. Linköping University Electronic Press.

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*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021a. Swiss Parliaments Corpus, an automatically aligned Swiss German speech to Standard German text corpus. In *Proceedings of the Swiss Text Analytics Conference 2021*.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2021b. SwissText 2021 task 3: Swiss German speech to Standard German text. In *Proceedings of the Swiss Text Analytics Conference 2021*.

Alan Ramponi. 2022. NLP for language varieties of Italy: Challenges and the path forward. *Computing Research Repository*, arXiv:2209.09757.

Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajič, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez-Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. 2020. European Language Grid: An overview. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3366–3380, Marseille, France. European Language Resources Association.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic parsed historical corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984, Istanbul, Turkey. European Language Resources Association (ELRA).

Beatrice Santorini. 2021. Penn parsed corpus of historical Yiddish. Version 1.0.

Odette Scharenborg, Laurent Besacier, Alan Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen,

Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux. 2020. Speech technology for unwritten languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:964–975.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019a. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53:837–863.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019b. ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache. *Linguistik Online*, 98(5):425–454.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset).

Ingrid Schröder. 2011. Sprachvariation in Norddeutschland (SiN). Archived in Hamburger Zentrum für Sprachkorpora. Version 0.1.

Sarah Schulz and Nora Ketschik. 2019. From 0 to 10 million annotated words: part-of-speech tagging for Middle High German. *Language Resources and Evaluation*, 53:837–863.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China. Association for Computational Linguistics.

Janine Siewert, Yves Scherrer, and Jörg Tiedemann. 2021. Towards a balanced annotated Low Saxon dataset for diachronic investigation of dialectal variation. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 242–246, Düsseldorf, Germany. KONVENS 2021 Organizers.

Janine Siewert, Yves Scherrer, Martijn Wieling, and Jörg Tiedemann. 2020. LSDC – a comprehensive dataset for Low Saxon dialect classification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 25–35, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Gary Simons and Steven Bird. 2003. The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. *Literary and Linguistic Computing*, 18(2):117–128.

Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. 2022. Creating a basic language resource kit for Faroese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4637–4643, Marseille, France. European Language Resources Association.

Marjoleine Sloos, Eduard Drenth, and Wilbert Heeringa. 2018. The Boarnsterhim corpus: A bilingual Frisian-Dutch panel and trend study. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Per Erik Solberg and Pablo Ortiz. 2022. The Norwegian parliamentary speech corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1003–1008, Marseille, France. European Language Resources Association.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian dependency treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 789–795, Reykjavik, Iceland. European Language Resources Association (ELRA).

Språkbanken. NB Tale – speech database for Norwegian. National Library of Norway. Last updated 2015-12-22.

Språkbanken and Fróðskaparsetur Føroya. FTS – Faroese text collection. Språkbanken Text (Department of Swedish at the University of Gothenburg) and the University of Faroe Islands. Last modified: 2015-05-27.

Elisabeth Stark, Simone Ueberwasser, and Anne Göhring. Corpus "What's up, Switzerland?". University of Zurich.

STIH. 2020. Bisame_gsw (alsacien) : corpus brut et annoté. ORTOLANG (Open Resources and TOols for LANGuage).

Sali A. Tagliamonte. 2007. Representing real language: Consistency, trade-offs and thinking ahead! In Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, editors, *Creating and Digitizing Language Corpora*, volume 1: Synchronic Databases, pages 205–240. Palgrave Macmillan UK, London.

Tekstlab. 2020. TAUS – the spoken language investigation in Oslo. Version 3.

Francis Tyers, Mariya Sheyanova, Aleksandra Martynova, Pavel Stepachev, and Konstantin Vinogorodskiy. 2018. Multi-source synthetic treebank creation

for improved cross-lingual dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150, Brussels, Belgium. Association for Computational Linguistics.

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

University of Helsinki. 2006. The Helsinki corpus of British English dialects. Department of Modern Languages, University of Helsinki.

University of Munich. Audioatlas siebenbürgisch-sächsischer Dialekte (ASD). Institut für deutsche Kultur und Geschichte Südosteuropas, Institut für romanische Philologie, IT-Gruppe Geisteswissenschaften. LMU Munich.

Jacques Van Keymeulen, Veronique De Tier, Anne Breitbarth, Anne-Sophie Ghyselen, and Melissa Farasyn. 2019. Het dialectologische corpus 'Stemmen uit het verleden' van de Universiteit Gent. *Volkskunde*, 120(2):193–204.

Dieter Van Uytvanck, Claus Zinn, Daan Broeder, Peter Wittenburg, and Mariano Gardellini. 2010. Virtual Language Observatory: The portal to the language resources and technology universe. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Elaine Vaughan and Brian Clancy. 2016. Sociolinguistic information and Irish English corpora. In Raymond Hickey, editor, *Sociolinguistics in Ireland*, pages 365–388. Palgrave Macmillan, London.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Manuela Weibel and Muriel Peter. 2020. Compiling a large Swiss German dialect corpus. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Linda Wiechetek, Katri Hiovain-Asikainen, Inga Lill Sigga Mikkelsen, Sjur Moshagen, Flammie Pirinen, Trond Trosterud, and Børre Gaup. 2022. Unmasking the myth of effortless big data - making an open source multi-lingual infrastructure and building language resources from scratch. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Martijn Wieling and John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics*, 1(1):243–264.

Emre Yilmaz, Maaike Andringa, Sigrid Kingma, Jelske Dijkstra, Frits van der Kuip, Hans Van de Velde, Frederik Kampstra, Jouke Algra, Henk van den Heuvel, and David van Leeuwen. 2016. A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4666–4669, Portorož, Slovenia. European Language Resources Association (ELRA).

Jihene Younes, Emna Souissi, Hadhemi Achour, and Ahmed Ferchichi. 2020. Language resources for Maghrebi Arabic dialects' NLP: A survey. *Language Resources and Evaluation*, 54(4):1079–1142.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Chiara Alzetta, Erik Andersen, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Þórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Juan Belieni, Kepa Bengoetxea, Yifat Ben Moshe, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Maria Clara Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit,

Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Boris Katz, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Mitsuya Sasaki, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine

Siewert, Einar Freyr Sigurðsson, João Ricardo Silva, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Barbara Sonnenhauser, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2022. Universal Dependencies 2.11. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Caleb Ziems, William Held, Jingfeng Yang, and Diyi Yang. 2022. Multi-VALUE: A framework for cross-dialectal English NLP. *Computing Research Repository*, arXiv:2212.08011.

A Written representations

Table 6 provides examples of different types of written representations and showcases how diverse each category can be.

Examples 1a, 2a, 3a, 4a/b, 5a and 6a are written in **standardized orthographies** (or in lowercased versions of standard orthographies with no pronunciation). Of these, sentences 1a, 4a and 5a are written in orthographies developed for their respective low-resource languages **A**, while 2a, 3a and 6a are normalized and written in the orthographies of closely related standard languages ¶ (e.g. Elfdalian written in Swedish or Swiss German written in Standard German).

Sentences 4b and 6a present two examples of ad-hoc **pronunciation spellings** ✎. These kinds of spellings vary from speaker to speaker, and one and the same speaker might also choose different spellings of the same word at different times.

Phonetic or phonemic transcriptions ✎ have different styles depending on each corpus's transcription guidelines. Examples 1b and 1c are written in modified versions of SAMPA and X-SAMPA, and the corpora come with sufficient documentation to automatically convert these transcriptions into IPA (1c, 2c). (The superscript symbols ^{1} and ^{2} in example 2c are commonly used to indicate the Norwegian pitch accent.) The transcription styles presented in examples 3b and 5b are based on Norwegian and Standard German orthography, respectively. What sets them apart from pronunciation spellings is that they are consistent across the entire corpus and that they follow linguistic rationales that often are outlined in the corpus documentation.

B Resources by language

Table 7 provides an overview of resource types by language variety.

C Overlapping data sources

Several of the corpora mentioned in this article overlap with each other:

- *UD Faroese OFT* and the *Korp* subcorpus of the background corpus of the Faroese *BLARK 1.0* contain material from the Faroese Wikipedia.
- The *NDC Treebank* uses data from the *Nordic Dialect Corpus.*
- *UD Low Saxon LSDC* is an annotated subset of *LSDC.*
- *UD Frisian/Dutch Fame* is an annotated subset if *FAME.*
- Many of the sentences in *UD Swiss German UZH* are also in *NOAH's corpus.* Both of these corpora contain material from the Alemannic Wikipedia.
- *SB-CH* contains *NOAH's corpus.*

From the Faroese BLARK recordings (Simonsen et al., 2022):		
1a	**A**	vit høvdu matpakka við og eg hugnaði mær óført
1b	✎	vId h9dI m%EApaHga v%i: o e h%u:najI mar %OW:f9zd
1c	✎	vɪd hœdɪ ˈmɛa̯pʰaʰga ˈviː o e ˈhuːnajɪ maɹ ˈɔu̯ːfœs̬d
"We had lunchboxes with us and I enjoyed myself greatly."		
From the Norwegian NB Tale corpus (Språkbanken):		
2a	¶	Etter litt godsnakk kom tre av kyrne mot han mens den fjerde glei og fall
2b	✎	""{t@4 l"it g""u:snAkk k"Om t4"e: "A:v C"y:n'@ m"u:t "An m"ens d_= fj""{:d'@ gl"eI "O: f"Al
2c	✎	²ɛtəɹ ¹lɪt ²guːsnɑkk ¹kɔm ¹tɪeː ¹ɑːv ¹çyːnə ¹muːt ¹ɑn ¹mɛns dn̩ ²fjæːɖə ¹glɛɪ̯ ¹oː ¹fɑl
"After some coaxing, three of the cows came towards him while the fourth one slipped and fell."		
From the Norwegian part of the Nordic Dialect Corpus (Johannessen et al., 2009):		
3a	¶	det er slik at de fleste kommer jo att når de får # unger
3b	✎	de e sjlik att dæi fLeste kjemme jo att nårr dæi fær # onnga
"The thing is that most people return when they have *[brief pause]* kids."		
From the Elfdalian part of the Nordic Dialect Corpus (Johannessen et al., 2009):		
4a	**A**	wen wa wen war eð før ien månað ? juni ?
4b	¶	vad va- vad var det för en månad ? juni ?
"What, wha-, what month was it? June?"		
From UD Low Saxon LSDC (Siewert et al., 2021):		
5a	**A**	Nu leyt em de böyse vynd disse nacht kyn ouge an enander doon.
5b	✎	Nu leit em de baise Find düse Nacht kinn Auge an enander dohn.
"Now the wicked wind didn't let them get a wink of sleep that night."		
From the Swiss German ArchiMob corpus (Scherrer et al., 2019b):		
6a	¶	können sie ihre jugendzeit beschreiben
6b	✎	chönd sii iri jugendziit beschriibe
"Can you describe your youth?"		
From the BISAME corpus (STIH, 2020):		
7a	✎	Niema hat salamols gweßt as die Werter vum franzeescha kumma.
"Nobody knew then that the words came from French."		

Table 6: **Examples of written representations.** Symbols: ✎ = phonetic/phonemic transcription, ✎ = pronunciation spelling, **A** = LRL orthography, ¶ = normalized orthography.

Language		Dialect/ Location	Morpho-syntax	Semantic	Parallel	Un-curated text	Curated data
North Germanic							
FAO	Faroese	📍	✔			✔	🎙 ✎ A
NOR	(non-std.) Norwegian	📍	✔				🎙 ✎ ✎ ¶
OVD	Elfdalian	📍					A ¶
SWE	(non-std.) Swedish	📍					✎ ¶
Anglo-Frisian							
SCO	Scots					✔	✎ ¶
ENG	(non-std.) English	📍					🎙 ¶
FRY	West Frisian	📍	✔			✔	🎙 A
FRR	North Frisian					✔	
STQ	Saterland Frisian					✔	
Low German							
NDS	Low Saxon	📍	✔	✔		✔	🎙 ✎ A
FRS	East Frisian Low Saxon						🎙
GOS	Gronings				✔	✔	
Macro-Dutch							
VLS	West Flemish	📍	✔			✔	✎ ¶
ZEA	Zeelandic					✔	
Middle German							
LTZ	Luxembourgish					✔	A
KSH	Colognian					✔	
LIM	Limburgish					✔	
PFL	Palatine German					✔	
PDC	Pennsylvania Dutch					✔	
YID*	Yiddish					✔	🎙 ✎
Upper German							
DEU	(non-std.) German						🎙 ✎ ¶
BAR	Bavarian		✔	✔	✔	✔	🎙 ✎ ✎ ¶
CIM	Cimbrian						🎙 ¶
SWG	Swabian					✔	
GSW	Swiss German	📍	✔	✔	✔	✔	🎙 ✎ ✎ ¶
WAE	Walser	📍					🎙 ✎

Table 7: **Corpora by language variety.** For ease of reference, the language are sorted by Germanic sub-branches (based on Glottolog (Hammarström et al., 2022)). *Glottolog discerns between Eastern Yiddish (Middle German) and Western Yiddish (Upper German). Symbols: 🎙 = audio, ✎ = phonetic/phonemic transcription, ✎ = pronunciation spelling, A = LRL orthography, ¶ = normalized orthography.

- The *Annotated Corpus for the Alsatian Dialects* contains articles from the Alemannic Wikipedia that were explicitly tagged as Alsatian.
- *TaPaCo* is a subset of *Tatoeba.*
- Any corpus that includes data from the internet might overlap with the uncurated datasets in Section 5.3.

D Wikipedia statistics

Table 8 provides a comparison of Wikipedia sizes and user (vs. bot) activity.[19]

The sizes of the small Germanic Wikipedias vary considerably from wiki to wiki (there are just under 2k Pennsylvania Dutch articles, while the (German) Low Saxon Wikipedia has over 84k articles), as does the number of recently active contributors (from 6 active non-bot users per month for Ripurarian/Colognian, Palatine German and Pennsylvania Dutch to 70 for Scots).

While bots can be used for automating many tasks that are unrelated to the textual diversity of a wiki (e.g. cleaning up article redirection pages), they can also be used to automatically create short template-based articles.[20] The share of manual edits (i.e. edits not by bots) is very varied across wikis – only about a quarter of all edits in the Pennsylvania Dutch Wikipedia have been made manually, compared to 79 % in the North Frisian Wikipedia. However, there is a clear trend towards a much larger proportion of manual edits: the vast majority of edits made only in the past year were manual edits.

Some of the wikis are written according to one or more orthographies, while others either do not include any spelling recommendations at all or encourage editors to use whatever pronunciation spelling they prefer.[21] Several of these wikis include (some) articles with metadata specifying which variety the document is written in.[22]

[19] The data sources are the automatically updated list of Wikipedia sizes at `meta.wikimedia.org/wiki/List_of_Wikipedias_by_language_group#Germanic` (last accessed 2023-01-31) and Wikimedia's metadata via `wikimedia.org/api/rest_v1/`. Scripts will be available via *[URL omitted in review version]*.

[20] For an example for the latter, see `nds.wikipedia.org/wiki/Bruker:ArtikelBot`

[21] The Dutch Low Saxon Wikipedia, for instance, recommends using the *Nysassiske Skryvwyse* (one of many proposed orthographies for Low Saxon): `nds-nl.wikipedia.org/wiki/Wikipedia:Spelling`. The Ripurarian/Colognian Wikipedia, on the other hand, encourages idiosyncratic spellings: `ksh.wikipedia.org/wiki/Wikipedia:Schrievwies`.

[22] E.g. `bar.wikipedia.org/wiki/Kategorie:Artikel_nach_Dialekt`

Wikipedia & Language		Articles (01/2023)	Manual edits (2001–2022)	Manual edits (2022)	Monthly editors (2022)
nds	NDS (Germany)	84 k	44 %	99 %	30
lb	LTZ	61 k	43 %	85 %	56
fy	FRY	50 k	60 %	99 %	54
sco	SCO	39 k	53 %	63 %	70
als	GSW + SWG	30 k	69 %	100 %	58
bar	BAR	27 k	68 %	63 %	39
frr	FRR	17 k	79 %	85 %	16
yi	YID	15 k	49 %	97 %	35
li	LIM	14 k	42 %	75 %	21
fo	FAO	14 k	41 %	99 %	29
vls	VLS	8 k	45 %	79 %	16
nds-nl	NDS (Netherlands)	8 k	40 %	68 %	14
zea	ZEA	6 k	47 %	98 %	10
stq	STQ	4 k	38 %	81 %	8
ksh	KSH + other Ripuarian	3 k	32 %	99 %	6
pfl	PFL	3 k	65 %	72 %	6
pdc	PDC	2 k	27 %	92 %	6
en	ENG	6608 k	90 %	92 %	102 574
de	DEU	2765 k	91 %	93 %	16 141
nl	NLD	2114 k	68 %	66 %	3521
da	DAN	289 k	63 %	64 %	711
is	ISL	56 k	54 %	79 %	118

Table 8: **Wikipedia statistics.** 'Manual edits' include the proportion of edits (of content pages) performed by registered non-bot users or anonymous editors (out of the total number of content page edits performed by anyone, including bots). The number of monthly editors is the mean number of registered non-bot users who edited at least one content page, per month. English, German, Dutch (NLD), Danish (DAN) and Icelandic (ISL) are included for comparison.