Final Report
The Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage

Harald Hammarström, Markus Forsberg and Shafqat Virk

Project name:	The Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage
PI Department:	Department of Linguistics and Philology Uppsala University Box 635 751 26 Uppsala Sweden
Participating Institutions:	Uppsala University Leiden University LLACAN, CNRS
Principal Investigator:	Dr. Harald Hammarström (1977-) harald.hammarstrom@lingfil.uu.se Department of Linguistics and Philology Uppsala University Box 635 751 26 Uppsala Sweden

1 Project Members

In the Swedish branch of the project, funded by RAÄ, the project members are:

Title	Name	Year of birth
Dr.	Harald Hammarström	1977
Dr.	Shafqat Virk	1979
Dr.	Markus Forsberg	1974

In the Dutch and French branches of the project, funded by the respective national agencies, the project members are:

Title	Name	Year of birth
Prof.	Marian Klamer	1965
Dr.	Søren Wichmann	1964
Dr.	Stéphane Robert	1958
Dr.	Guillaume Segerer	1965

2 Summary Results

The diversity of the world's 7 500 languages embodies a wealth of information on human cognition and the history of populations. As languages go extinct, the linguistic heritage of human kind increasingly resides in printed grammars and dictionaries, which are rapidly piling up. The present project has:

- Provided access to the world's linguistic heritage by making an existing collection of more than 9 000 PDF documents no longer protected by to copyright in a stable archive enriched by added metadata and computational tools developed to search information within the texts. The documents can be browsed at `https://spraakbanken.gu.se/korp/?mode=dream` and downloaded wholesale at `https://spraakbanken.gu.se/en/resources/dream`.

- Developed automatic Information Extraction tools specifically targeted to the domain of linguistic diversity. These tools enable human researchers in typological and historical linguistics to obtain items of information on a much broader scale in a less time-consuming way. Application to the copyright-free as well as copyrighted texts has resulted in the largest databases available for some linguistic features.

- A collection of dictionaries have been revamped into apps for mobile devices to be distributed to speakers of minority languages, handing back to these speakers some of their linguistic heritage. This was done by the French branch of the project.

3 Background and Aims

The diversity of the world's 7 500 languages embodies a wealth of information on human cognition and the history of populations. As languages go extinct, the linguistic heritage of human kind increasingly resides in printed grammars and dictionaries, which are rapidly piling up. The words, text and grammar of the world's languages constitute one of the most important elements of our cultural heritage and carries unique value for the respective speakers, the general public and linguistic researchers.

At the same time, technological advances have opened up new ways to access these legacy materials. Books can be digitized, accessed over the internet and searched/analyzed using natural language processing tools.

The DReaM project sought to transform the heritage materials on the world's minority languages from being scattered on bookshelves to an accessible, usable and enhanced digital form. Their curated digital availability immediately allows for regeneration of dictionaries in smartphone apps as well as automatic extraction of pieces of linguistic information using state-of-the-art tools in Natural Language Processing and Machine Learning.

4 Theory and Method

To locate and collect the relevant documents which are scattered across a wide range of languages, publication types and outlets is not a small task. Thanks to existing efforts in the Glottolog project (`glottolog.org`, Hammarström et al. 2021) we were able to jumpstart and continue the methodology of going through handbooks, overviews and expert literature to find the most extensive descriptive materials for each of the languages of the world. Major and minor online repositories were searched for relevant documents and those that could not be found electronically were sought out in libraries and digitized manually. Each document was provided with metadata concerning the language it is written in (the meta-language, usually English, French, German,

Spanish or Mandarin Chinese), the language(s) described in it (the target language, typically one of the thousands of minority languages throughout the world), and the type of description (comparative study, description of a specific features, phonological description, grammar sketch, full grammar etc).

Once digitized, text files for all non-copyrighted texts with linguistic descriptions were made openly available in a central, continuously curated archive accessible through the Internet. This will be hosted by Språkbanken (the Swedish Language Bank, `http://spraakbanken.gu.se`) of Gothenburg University, which has already hosted linguistic corpora for four decades and has a stable infrastructure for long-term future archiving and distribution of digital materials. The collection was be enriched with extensive metadata and exposed to relevant research tools at Språkbanken, such as Korp (Borin et al. 2012, `https://spraakbanken.gu.se/korp`), a wordbased research tool which can improve the searchability of texts.

Various tools from the fields of Natural Language Processing and Machine Learning are available to extract information automatically from raw text. Of particular relevance are Optical Character Recognition (OCR) stemming, vector space semantics, TF-IDF, clustering, and various forms of supervised ML techniques including Deep Learning. Some techniques developed specifically for the domain of extracting linguistic information have also been recently developed, principally around the DReaM project itself (e.g., Virk et al. 2017, Macklin-Cordes et al. 2017 and below). Existing manually curated databases of linguistic features are used for supervised learning as well as evalution. Extracting information can be done on all digitized descriptions, not only the ones out of copyright.

Dictionaries digitized according to the methodology of the RefLex project (Segerer 2016) preserve all the relevant information to be rendered in other formats. Web-accessible dictionaries serve the research community well, but the preferred form of engagement with speaker communities is through smartphones which is by now practical enough. This form of digital distribution is especially beneficial for minority groups who are otherwise forced to conduct all digital activities using another larger language other than their own minority language.

5 Main Outcomes

At present, over 50 000 linguistic descriptive works in electronic form have been gathered in the DReaM corpus. Of these, around 12 000 consist of grammars and grammar sketches and the remainder of lexical resources and miscellaneous materials such as phonological descriptions, comparative studies, sociolinguistic studies and descriptions of specific aspects of language structure. About one quarter of the documents were amassed within the auspices of the present project, building on an existing initiative and is subject to further expansion in the future. The collection consists of (1) out-of-copyright texts digitized by national libraries, archives, scientific societies and other similar entities, (2) texts posted online with a license to use for research usually by university libraries and non-profit organizations (notably the Summer Institute of Linguistics, `http://www.sil.org`), (3) documents scanned directly from hard copies under publisher copyright where quotations of short extracts are legal.

For each document, there are features manually curated annotations of

(i) the language it is written in (the meta-language, usually English, French, German, Spanish, Russian or Mandarin Chinese, see Table 1),

(ii) the language(s) described in it (the vernacular, typically one of the thousands of minority languages throughout the world), and

(iii) the type of description (comparative study, description of a specific feature, phonological description, grammar, bibliography, sociolinguistic study, overview etc, see Hammarström and Nordhoff (2011).

The collection has been OCRed using ABBYY Finereader 14 using the meta-language as recognition language. The original digital documents are of quality varying from barely legible typescripts to high-quality scans and even born-digital documents. The OCR correctly recognizes most tokens of the set meta-language, but, particular to our collection, most documents contain a fraction of tokens which do not belong to the set meta-language but to the minority language(s) being discussed. These tokens are typically recognized poorly, as expected from the dictionary/training-heavy, contemporary techniques for OCR. We cannot easily improve on the OCR on a scale relevant for the present collection but some post-correction of OCR output very relevant for the genre of linguistics is possible and advisable (see Hammarström et al.

Meta-language		**# docs**	**# pages**
English	eng	24 081	3 493 543
French	fra	3 632	666 187
German	deu	2 798	434 647
Spanish	spa	2 628	382 859
Portuguese	por	1 343	177 615
Russian	rus	993	267 263
Dutch	nld	662	92 447
Mandarin Chinese	cmn	537	131 687
Indonesian	ind	477	72 148
Italian	ita	402	71 552
…	…	…	…

Table 1: Meta-languages of documents in DReaM corpus.

2017) and if there is training data, more powerful supervised techniques can be applied (Dannélls and Virk 2021).

The collection contains descriptive information of varying amounts for no less than 7 521 languages (very close to the total number of languages known) and grammars/grammar sketches for 4 527 languages (very close to the total number of languages for which such a description exists, Hammarström et al. 2018). A listing of the collection can be enumerated via the open-access bibliography Glottolog (`glottolog.org`, Hammarström et al. 2021). The copyright-free subset of documents can be browsed at `https://spraakbanken.gu.se/korp/?mode=dream` and downloaded wholesale from `https://spraakbanken.gu.se/en/resources/dream`. The DReaM corpus is introduced and described to a wider audience for the first time in Virk et al. (2020a). Figure 1 shows a screenshot of Korp interface to the DReaM corpus hosted at Språkbanken, Gothenburg University.

There several avenues to enhance access, usability and information extraction from the document collection. The key property characterizing the DReaM corpus is that its documents describe languages. Thanks to the high-quality bibliographical metadata associated with all the documents, it is possible to make targeted searches of all kinds — something not possible in, e.g., google books (Hammarström 2021).

There are regularities in document structure that can be recognized automatically with good accuracy. Many descriptive publications contain separate sections for morphosyntactic description, text and lexicon and the each

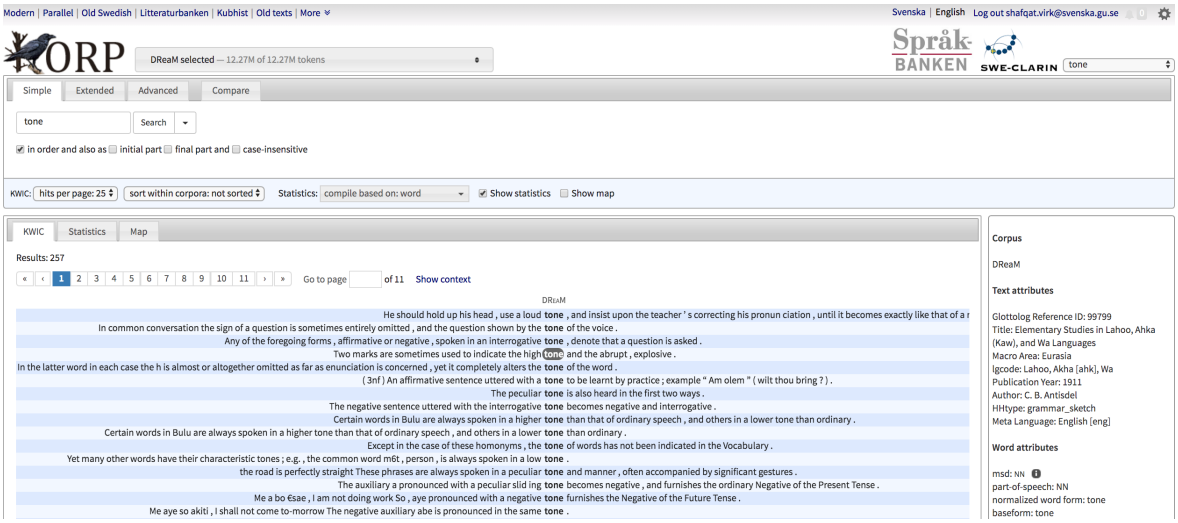


Figure 1: Screenshot of the 'Basic' search mode in the Korp interface to the DReaM corpus.

page of each type tends to have a characteristic word frequency distribution. This coupled with the fact that the different types come in coherent sections makes it possible to slice descriptive publications into the respective sections (Hammarström 2021a) which is useful for many downstream search tasks (e.g., search only the grammar section, flag a document as containing a large dictionary even if this is not highlighted in its title, count/normalize the length by the appropriate section, etc.).

The most conspicuous formal, or semi-formal, elements of all scientific publications are the cited bibliographical items. State-of-the-art methodology from other fields of science needed some extension for the DReaM corpus because of its extraordinary heterogeneity in times, styles and languages. Initial work on this task has been submitted (Hammarström 2021) and needs to be continued further in order to produce a web-of-science specific to descriptive linguistics. Another valuable (semi-)formal element of recent grammars is so-called Interlinear Glossed Text (IGT, see Round et al. 2020 and references therein). Because of crucial OCR issues, extraction of IGT has not been attempted in the present project, but could in principle be essayed on the born-digital subset of DReaM.

Because of its representativeness the DReaM corpus can used for some important meta-level questions in linguistics. The Dutch branch of the DReaM

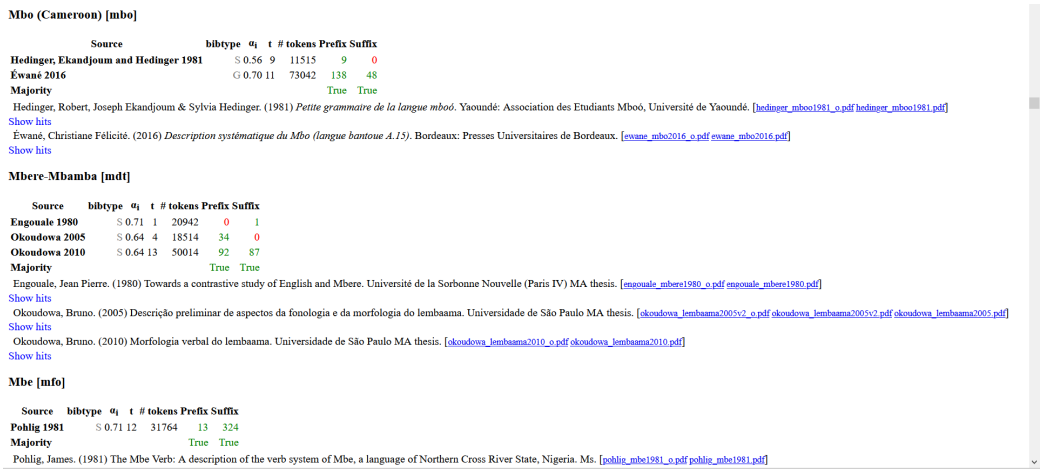
project developed a schema for obtaining all technical terms in linguistics from the corpus (Wichmann 2021). A concrete application is the submission by Hammarström (2022) which attempts to measure how well described a language is using such lists of technical terms.

A central achievement of the DReaM project is the emergence of (semi-)automatically constructed databases of various features of the languages of the world. Two parallel approaches have been developed: one "deep" which extracts information using frame-semantics and Deep Learning and one "shallow" which exploits occurrence statistics of key terms in a convenient way.

The deep approach requires resources/annotation following the classic theory of frame semantics (Malm et al. 2018, Virk et al. 2019a). Once annotations exist for a given domain and targeted feature, the sought-after information can be obtained even as it is expressed in a myriad of different ways in different grammars. Some research has utilized classic supervised Machine Learning techniques (Aslam 2019, Foster 2019, Virk et al. 2019b) and, more recently, state-of-the-art Deep Learning techniques (Virk et al. 2021b). A general strategy for reducing the manual labour for a given domain has also been devised (Virk et al. 2021a).

The shallow approach requires no annotation/resources and no human tuning or other intervention, but hinges on the existence of some key term(s) that are associated with the feature of interest (Hammarström et al. 2021). The simplicity and broad applicability of this technique has resulted in by far the biggest cross-linguistic databases to date on features such as prefixation/suffixation (Hammarström 2021b), gender/noun classes (Allassonnière-Tang et al. 2021), classifiers (Tang et al. 2022), language endangerment (Zariquiey et al. ress), ejectives (Urban and Moran 2021) and others to follow in the future with a public interface (Hammarström 2021). Figure 2 shows a screenshot with a snippet of the search/extraction output for the search terms `prefix` and `suffix`. For each language and corresponding grammatical descriptions, the number of hits is shown, alongside the automatically calculated threshold t that determines whether the number of hits is 'significant' (see Hammarström et al. 2021 for details).

Both the deep and shallow approach target specific features identified beforehand. A more ambitious goal to obtain a general profile with all features for each language has been articulated (Virk et al. 2020b) and serves as the long-term objective beyond the present project.



Mbo (Cameroon) [mbo]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Hedinger, Ekandjoum and Hedinger 1981	S	0.56	9	11515	9	0
Éwané 2016	G	0.70	11	73042	138	48
Majority					True	True

Hedinger, Robert, Joseph Ekandjoum & Sylvia Hedinger. (1981) *Petite grammaire de la langue mboó*. Yaoundé: Association des Etudiants Mboó, Université de Yaoundé. [hedinger_mboo1981_o.pdf hedinger_mboo1981.pdf]
Show hits
Éwané, Christiane Félicité. (2016) *Description systématique du Mbo (langue bantoue A.15)*. Bordeaux: Presses Universitaires de Bordeaux. [ewane_mbo2016_o.pdf ewane_mbo2016.pdf]
Show hits

Mbere-Mbamba [mdt]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Engouale 1980	S	0.71	1	20942	0	1
Okoudowa 2005	S	0.64	4	18514	34	0
Okoudowa 2010	S	0.64	13	50014	92	87
Majority					True	True

Engouale, Jean Pierre. (1980) Towards a contrastive study of English and Mbere. Université de la Sorbonne Nouvelle (Paris IV) MA thesis. [engouale_mbere1980_o.pdf engouale_mbere1980.pdf]
Show hits
Okoudowa, Bruno. (2005) Descrição preliminar de aspectos da fonologia e da morfologia do lembaama. Universidade de São Paulo MA thesis. [okoudowa_lembaama2005v2_o.pdf okoudowa_lembaama2005v2.pdf okoudowa_lembaama2005.pdf]
Show hits
Okoudowa, Bruno. (2010) Morfologia verbal do lembaama. Universidade de São Paulo MA thesis. [okoudowa_lembaama2010_o.pdf okoudowa_lembaama2010.pdf]
Show hits

Mbe [mfo]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Pohlig 1981	S	0.71	12	31764	13	324
Majority					True	True

Pohlig, James. (1981) The Mbe Verb: A description of the verb system of Mbe, a language of Northern Cross River State, Nigeria. Ms. [pohlig_mbe1981_o.pdf pohlig_mbe1981.pdf]

Figure 2: Sample search/extraction output for the search terms `prefix` and `suffix`. The sources are spelled out with links to full-text and displayable hit snippets.

6 Positioning of the Outcomes in a National and International Research Context

Språkbanken is the leading national archive for language-related corpora and is active in the development of software tools, formats, archiving practices and interfaces at the highest international level.

The DReaM corpus is the broadest cross-linguistic grammar collection ever collected and is fully up-to-date with international language catalogues and annotation standards in the field of diversity linguistics.

The work on automatic information extraction on raw text descriptions is nearly the only work so far on this very topic. The possibilites are far from exhausted and, in particular, standard tools from NLP and Information Retrieval, e.g., (multilingual) stemming, (multilingual) vector space semantics, multi-word expressions, etc. remain to be amply utilized.

7 Relevance of the Outcomes for the Cultural Environment, Culture Heritage and Cultural Environment Work

Thousands of out-of-copyright linguistic descriptions are now gathered in a central archive for long-term free browsing and download. Moreover, searches and extraction results from the full set of publications in the DReaM corpus are available for anyone interested in the world's linguistic heritage. Since the DReaM corpus covers nearly every language, there is something to be found for all appetites.

For researchers in diversity linguistics, the possibility to do large extractions and build databases relatively rapidly is an enormously powerful tool. In turn, greater availability of large databases of linguistic facts allow for broader analyses and novel analyses bearing on all aspects of our cultural heritage.

Aside from academic researchers, the most important audience for linguistic heritage data are the speaker communities themselves. In the context of language endangerment, heritage data have a crucial role to play in stimulating the younger generation's interest and capabilities in the language (Tsunoda 2005). Most minority languages throughout the world are spoken by socio-economically challenged groups who, paradoxically but typically, have little access to materials on their own language. In the digital age, this gap can be shrunk considerably.

Other fields of humanities, such as history, ethnography and archaeology (especially of the comparative kinds), also rely heavily on textual sources and similarly face challenges due to the accumulating amounts of data which cannot be immediately accessed or comprehensively searched. The DReaM project provides an example on how to utilize technological developments on the vast backlog of legacy data also for these fields.

8 Dissemination of Outcomes: Present and Future

Dissemination to the academic world is done through a series of seminars, conference presentations and academic publications in national and international linguistics, computational linguistics and information retrieval venues.

The dissemination has also benefitted from collaborations with four associated partners that span the Western, Russian and Chinese scientific communities, although the COVID-pandemic hindered a physical meeting with associated partners. These activities will continue in the future and we also envisage an extension into other fields such as genetics and ethnography where the data and/or search paradigm are of considerable interest.

The project's free and open web presence also caters to the general public and for this reason long-term support, uncomplicated interfaces and well-documented background are key criteria. The design of the project is such that metadata and back links are preserved at every stage to ensure credit is traceable to the efforts of the work in the legacy sources. Advertisements have been posted on the Språkbanken blog[1] and the resources are linked, in a series of steps, to the major entry points Glottolog, Wikipedia and the Open Language Archives Community (OLAC).

Information from the largest (semi-)automatically generated cross-linguistic databases can be expected to trickle down to wider audiences through publications — especially, high-profile publications such as Allassonnière-Tang et al. 2021.

Avant-garde dissemination to the speaker communities is achieved through the smartphone dictionaries subproject by the French leg of DReaM. Here legacy dictionaries are brought back to the speech communities via fieldwork and local capacity building.

References

Allassonnière-Tang, M., Lundgren, O., Robbers, M., Cronhamn, S., Larsson, F., Her, O.-S., Hammarström, H., and Carling, G. (2021). Expansion by migration and diffusion by contact is a source to the global diversity of linguistic nominal categorization systems. *Nature: Humanities and Social Sciences Communications*, 8(331):1–6, 1–50.

Aslam, M. I. (2019). Semantic frame based automatic extraction of typological information from descriptive grammars. Master's thesis, University of Skövde.

[1] `https://spraakbanken.gu.se/blogg/index.php/2020/04/07/a-multilingual-annotated-corpus-of-worlds-natural-language-descriptions/`

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp — the corpus infrastructure of språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 474–478, Istanbul, Turkey. European Language Resources Association (ELRA).

Dannélls, D. and Virk, S. (2021). A supervised machine learning approach for post-ocr error detection for historical text. In Dobnik, S., Johansson, R., and Ljunglöf, P., editors, *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020), 25-27 November 2020*, pages 13–20. Linköping: Linköping Electronic Press.

Foster, D. (2019). Automatic frame-semantic parsing for linguistic descriptions: Extracting typological linguistic information from unstructured text. Master's thesis, University of Gothenburg.

Hammarström, H., Castermans, T., Forkel, R., Verbeek, K., Westenberg, M. A., and Speckmann, B. (2018). Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation*, 12:359–392.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2021). Glottolog 4.5. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at `http://glottolog.org`. Accessed on 2021-12-10.

Hammarström, H., Virk, S. M., and Forsberg, M. (2017). Poor man's OCR post-correction: Unsupervised recognition of variant spelling applied to a multilingual document collection. In *Proceedings of the Digital Access to Textual Cultural Heritage (DATeCH) conference*, pages 71–75. Göttingen: ACM.

Hammarström, H. (2021). Bibliographical parsing of descriptive linguistic literature. Submitted.

Hammarström, H. (2021). Gramfinder: Human and machine reading of grammatical descriptions of the languages of the world. In Dragut, E. C., Li, Y., Popa, L., and Vucetic, S., editors, *3rd Workshop on Data Science with Human in the Loop, DaSH@KDD, Virtual Conference, August 15, 2021*, pages 1–6.

Hammarström, H. (2021a). Inventory and content separation in grammatical descriptions of languages of the world. In Berget, G., Hall, M. M., Brenn,

D., and Kumpulainen, S., editors, *Linking Theory and Practice of Digital Libraries: 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13-17, 2021, Proceedings*, pages 129–140. Berlin: Springer.

Hammarström, H. (2021b). Measuring prefixation and suffixation in the languages of the world. In *Proceedings of The 3rd Workshop on Research in Computational Typology and Multilingual NLP*, pages 81–89. Stroudsburg, PA: Association for Computational Linguistics (ACL).

Hammarström, H. (2022). How good is this grammar? term-counting techniques for measuring the comprehensiveness of grammatical description. Submitted.

Hammarström, H., Her, O.-S., and Tang, M. (2021). Term-spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. In Dobnik, S., Johansson, R., and Ljunglöf, P., editors, *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020), 25-27 November 2020*, pages 27–34. Linköping: Linköping Electronic Press.

Hammarström, H. and Nordhoff, S. (2011). Langdoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language*, 3(2):31–43.

Macklin-Cordes, J. L., Blackbourne, N. L., Bott, T. J., Cook, J., Ellison, T. M., Hollis, J., Kirlew, E. E., Richards, G. C., Zhao, S., and Round, E. R. (2017). Robots who read grammars. Poster presented at CoEDL Fest 2017, Alexandra Park Conference Centre, Alexandra Headlands, QLD.

Malm, P., Virk, S. M., Borin, L., and Saxena, A. (2018). Lingfn: Towards a domain-specific linguistic framenet. In Torrent, T. T., Borin, L., and Baker, C. F., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) Workshop 5 — International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons*, pages 37–43. European Language Resources Association (ELRA), Paris, France.

Round, E., Ellison, M., Macklin-Cordes, J., and Beniamine, S. (2020). Automated parsing of interlinear glossed text from page images of grammatical

descriptions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2871–2876, Marseille, France. European Language Resources Association.

Segerer, G. (2016). Reflex: la reconstruction sans peine. *Faits de Langues*, 47:201–214.

Tang, M., Her, O.-S., and Hammarström, H. (2022). Defining numeral classifiers and identifying classifier languages of the world. Submitted.

Tsunoda, T. (2005). *Language Endangerment and Language Revitalization*, volume 148 of *Trends in Linguistics: Studies and Monographs*. Berlin: Mouton de Gruyter.

Urban, M. and Moran, S. (2021). Altitude and the distributional typology of language structure: Ejectives and beyond. *PLoS ONE*, 16(2)(e0245522):1–36.

Virk, S., Forsberg, M., and Hammarström, H. (2017). Textcat for language profiling. Submitted.

Virk, S., Malm, P., Borin, L., and Saxena, A. (2019a). Lingfn: A framenet for the linguistics domain. In *CICLing 2019: Short Oral Presentations and Poster Session*, pages 1–12.

Virk, S. M., Dannélls, D., Borin, L., and Forsberg, M. (2021a). A data-driven semi-automatic framenet development methodology. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1471–1479, Held Online. INCOMA Ltd.

Virk, S. M., Foster, D., Sheikh Muhammad, A., and Saleem, R. (2021b). A deep learning system for automatic extraction of typological linguistic information from descriptive grammars. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1480–1489, Held Online. INCOMA Ltd.

Virk, S. M., Hammarström, H., Forsberg, M., and Wichmann, S. (2020a). The DReaM corpus: A multilingual annotated corpus of grammars for the world's languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 871–877. Marseille, France: European Language Resources Association, Marseille, France.

Virk, S. M., Hammarström, H., Borin, L., Forsberg, M., and Wichmann, S. (2020b). From linguistic descriptions to language profiles. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 23–27. Marseille: European Language Resources Association (ELRA).

Virk, S. M., Muhammad, A. S., Borin, L., Aslam, M. I., Iqbal, S., and Khurram, N. (2019b). Exploiting frame-semantics and frame-semantic parsing for automatic extraction of typological information from descriptive grammars of natural languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, page 1247–1256. Varna, Bulgaria: NCOMA Ltd.

Wichmann, S. (2021). Pipeline for a data-driven network of linguistic terms. In Dobnik, S., Johansson, R., and Ljunglöf, P., editors, *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020), 25-27 November 2020*, pages 66–71. Linköping: Linköping Electronic Press.

Zariquiey, R., Arakaki, M., Vera, J., Torres, G., Cuba, C., Barrientos, C., García, A., Ingunza, A., and Hammarström, H. (in press). Linking endangerment databases and descriptive linguistics: an assessment of the use of terms relating to language endangerment in grammars. *Language Documentation & Conservation*, page 27pp.