CLD²: Language Documentation Meets Natural Language Processing for Revitalising Endangered Languages

Roberto Zariquiey* **Arturo Oncevay†** **Javier Vera‡**
*Dep. of Humanities, Linguistics Unit, Pontificia Universidad Católica del Perú, Perú
†School of Informatics, University of Edinburgh, Scotland
‡Escuela de Ing. Informática, Pontificia Universidad Católica de Valparaíso, Chile
rzariquiey@pucp.edu.pe, a.oncevay@ed.ac.uk, javier.vera@pucv.cl

Abstract

Language revitalisation should not be understood as a direct outcome of language documentation, which is mainly focused on the creation of language repositories. Natural language processing (NLP) offers the potential to complement and exploit these repositories through the development of language technologies that may contribute to improving the vitality status of endangered languages. In this paper, we discuss the current state of the interaction between language documentation and computational linguistics, present a diagnosis of how the outputs of recent documentation projects for endangered languages are underutilised for the NLP community, and discuss how the situation could change from both the documentary linguistics and NLP perspectives. All this is introduced as a bridging paradigm dubbed as Computational Language Documentation and Development (CLD²). CLD² calls for (1) the inclusion of NLP-friendly annotated data as a deliverable of future language documentation projects; and (2) the exploitation of language documentation databases by the NLP community to promote the computerization of endangered languages, as one way to contribute to their revitalization.

1 Introduction

There are around 6,500 mutually unintelligible languages in the world (Hammarström et al., 2018). However, several thousand minority languages are in danger of being lost forever without leaving systematic records. In response to this, in the last decades *Documentary Linguistics* has become a major and vibrant field in Linguistics, which attempts to produce permanent records of the linguistic and cultural practices of the most threatened speech communities (Himmelmann (2012); Austin (2010); Woodbury (2011), among many others).

The outcomes of documenting a language in the frame of contemporary Documentary Linguistics often comprise large amounts of audio and video recordings, featuring collections of texts (often transcribed, translated and interlinearized), as well as lexical repertoires, framed as vocabularies or dictionaries, with different degrees of detail. These data are often deposited in international language archives, from which they can be accessed by scholars and members of speech communities. Transcription of texts is often conducted in the ELAN software (Max Planck Institute for Psycholinguistics, 2021), and interlinearization if often conducted using software tools, such as FLex (Summer Institute of Linguistics, 2021a) and Toolbox (Summer Institute of Linguistics, 2021b). The ideal outcome of this process are time-aligned parsed transcriptions with information about the morphological structure and the part-of-speech class of each lexical unit. Texts are often presented in .txt or .htlm formats.

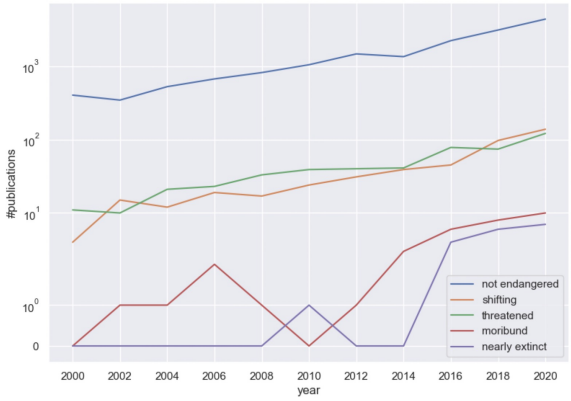

Figure 1: Number of publications in the ACL Anthology where languages are explicitly named in the title or abstract, and they are classified by their vitality from the Agglomerated Endangerment Status (Seifart et al., 2018). Vertical axis is in log-scale.

International language archives comprises documentation databases for several hundred languages. For instance, the Endangered Language Archive (ELAR) includes collections for 695 languages[1],

[1] https://www.elararchive.org/

Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 20 - 30

each of which may comprise several hours of transcribed and parsed speech, which represent several thousands of fully annotated sentences. These data has been produced in the frame of collaborative documentation projects with high ethical standards in terms of their methods, their outcomes and their dissemination. Thus, in principle, the data available through international language archives have been published with the permission of the linguistic communities involved, and therefore it is expected that they will be incorporated into new research, education and revitalisation projects, ideally with the participation of members of the communities culturally and linguistically linked to the data (Bird, 2020).

Language databases, however, are often underexploited for further developments. Although field linguists very often incorporate revitalisation components in their documentation projects, language *documentation* and language *revitalisation* are not equivalent in terms of their frames, methods and outcomes. Language revitalisation will surely take advantage of the data produced in language documentation projects, by actively using such records in community-based revitalisation programs, which may take various shapes according to the needs of the community and/or the scope of the project. Although it is true that creating a language repository alone cannot revert language endangerment or decay, there are several ways in which documentation data can be integrated into revitalisation projects. Here, we focus on one, associated with the perspective of language technologies. Language technologies offer a promising perspective for language revitalisation, not only because technological gadgets such smart phones are becoming more popular even in rural areas, but also because they are inexpensive. The concern about language endangerment is a fundamental issue in contemporary approaches to Computational Linguistics, and in the last years, the "computerisation" of minority languages has become a growing field in NLP research (Berment, 2002). NLP developments' potential contribution to revitalising endangered languages is high, but there is still moderate interaction between Documentary Linguistics and NLP research for language revitalisation.

In this paper, we reflect on the necessity of increasing the interactions between Documentary Linguistics and NLP. This is not a novel point in the literature (see particularly (Levow et al., 2017)), but to our knowledge this is the first attempt to put some ideas on this topic together in a position paper. We hope that the proposals we dubbed here as Computational language Documentation and Development (CLD²) will stimulate debate and more vibrant interactions between documentary linguists and NLP developers.

2 Language documentation and language revitalisation

Language documentation (or documentary linguistics) emerged at the end of the last century as a research program whose primary motivation lies in the concern about the accelerating loss of language diversity in the world. As a response, language documentation aims to create permanent records of the linguistic and cultural practices of the most threatened speech communities (Himmelmann, 1998; Austin, 2010; Woodbury, 2011). These records are framed as databases, ideally including several hours of audio and video recordings of monologue and dialogue texts belonging to various genres and topics (e.g. traditional tales and myths, verbal art, jokes, historical facts, life stories, cultural knowledge, among others). A good portion of these recordings is transcribed, translated and parsed. Each transcribed sentence is expected to be time-aligned and to include an orthographic or IPA representation, a morphemic parse, glossing, information about parts of speech and a free translation.

Producing such linguistic databases is a long-term and time-consuming task that may take several years and requires considerable funding. The expectation is that these linguistic databases, conceptualised as multipurpose repositories deposited and curated in international archives, will be preserved for posterity and thus will support community-based revitalisation projects in the future. Although it is true that language documentation projects very often incorporate revitalisation components, they are inevitably marginal since the documentation itself is the main focus of documentary linguistics. Therefore, the contribution of language documentation to language revitalisation is potentially significant but mainly indirect: the linguistic repositories produced in the frame of language documentation projects can indeed contribute to future revitalisation projects, but crafting and archiving a repository is not expected to have an inherent positive impact on the vitality status of an endangered language.

`collections/`, consulted on February, 28th, 2022

3 Language documentation and computational linguistics

Most interactions between computational linguistics and documentary linguistics relate to the release of software tools for language documentation, processing and archiving (van Esch et al., 2019; Anastasopoulos et al., 2020). Computational linguists and computer scientists have developed advanced software tools to assist field linguists in the various processes of contemporary language documentation, making them less time-consuming, more efficient and more systematic. These tools have been crucial for the exponential growth of language documentation on a global scale.

Contemporary language documentation implies a large amount of technical sophistication for managing, annotating, processing and archiving lasting and large repositories (Himmelmann, 2006; Austin, 2006; Woodbury, 2003, among many others). This could not be achieved without the contribution of computer scientists (particularly software developers). In the last decades, we have witnessed the release of specialised software tools nowadays customary for language documentation, speech analysis and linguistic fieldwork. Field linguist's Toolbox (before "Shoebox") (Summer Institute of Linguistics, 2021a) and more recently Fieldworks (FLex) (Summer Institute of Linguistics, 2021b) are data management and analysis tools for field linguists developed by the Summer Institute of Linguistics, which are used in language documentation and taught in linguistics schools worldwide. Toolbox and Flex allow to create dictionaries, which can be used for morphosyntactic parsing and annotation of transcribed texts. Transcription is often conducted in a different and nowadays very popular software called ELAN (Max Planck Institute for Psycholinguistics, 2021), developed by the Max Planck Institute for Psycholinguistics. ELAN allows to visualise and play audio and video files in order to create time-aligned transcriptions and translations. ELAN can also be used for morphological parsing, but most linguists prefer to conduct such tasks in Toolbox or FLex since ELAN transcriptions can be easily exported into these programs. In Toolbox or Flex, each sentence in an ELAN file (containing a transcription and a free translation) can receive morphemic parsing, morpheme-by-morpheme glossing and parts of speech tags, among any other relevant information in the frame of a specific project. The resulting Toolbox/Flex files are text files that can be opened back in ELAN, in PRAAT (a phonetics analyser) (Boersma and Weenink, 2001), or to be processed in Python or any other programming language as plain texts. This is shown in Figure 2.

In sum, there have been several attempts from the computational side trying to create or incorporate intelligent components in language documentation tools and procedures (Good et al., 2014; Arppe et al., 2017, 2019; van Esch et al., 2019; Anastasopoulos et al., 2020). We find a one-direction application (computation into language documentation), but there are still few developments in the other direction (language documentation into computation). One of our takes in this paper is that language documentation can significantly contribute to computational linguistics by providing data and insights to develop NLP tools for endangered languages.

4 NLP has not really met endangered language documentation

As mentioned before, NLP has mainly focused on aiding the language documentation pipeline. However, has NLP taken advantage of the outputs of the documentation projects, especially for endangered languages?

4.1 Data

To address that question, we looked into the central repository of NLP publications: the ACL Anthology[2], the language inventory of massive multilingual datasets in NLP research (UniMorph (McCarthy et al., 2020), Universal Dependencies (Nivre et al., 2020), Tatoeba (Tiedemann, 2020))[3], and the central database of language documentation projects for endangered languages: The Endangered Languages Archive, or ELAR, which is supported by the Endangered Languages Documentation Programme or ELDP[4].

Besides, we work with the list of languages from Glottolog 4.4 (Hammarström et al., 2021), which is an extended inventory of living and extinct languages, including metadata such as geographical location and other properties. Moreover, we use the Agglomerated Endangerment Status (AES) classification proposed by Seifart et al. (2018) to distinguish the vitality status of the language inventory.

[2] https://aclanthology.org/

[3] We chose these datasets as they are the most diverse collections according to their language inventory.

[4] https://www.eldp.net/

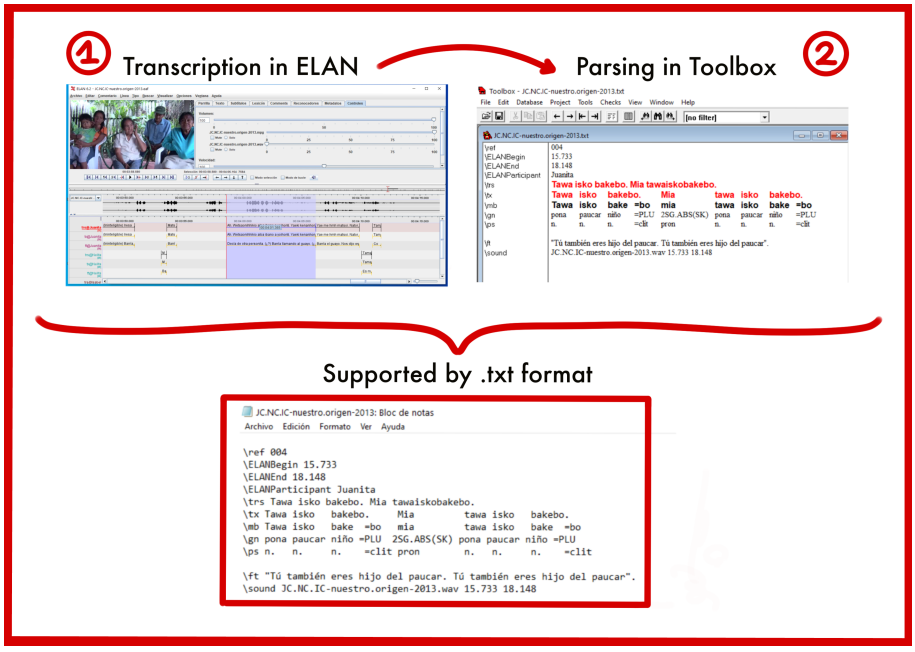

Figure 2: Graphic representation of the standard computational frame of language documentation: transcription is conducted in ELAN; ELAN files are imported into Toolbox or FLex where they are fully parsed and glossed. Crucially, we are dealing with .txt files throughout the process, which enormously facilitates their manipulation in any programming language

The classes are, from more to less vital: not endangered, shifting, threatened, moribund, nearly extinct and extinct[5].

4.2 Processing

With the language inventory and their vitality status, we first identified all the publications in the ACL Anthology (both conference and workshop proceedings) whose title or abstract explicitly includes the name of a language[6]. We manually clean false positives, such as concise language names (less than five characters) that can be confused with English words or acronyms.

A similar procedure is done with the ELAR database: all the projects are extracted, the language names are matched with the Glottolog inventory, and we manually curated potential false positives. From all the 570 projects published in the ELAR database, we identified 307 language names matching with the Glottolog database. With this, we obtained geographical information for 286 languages.

The procedure is similar for the massively multilingual (MM) datasets (Unimorph, Universal Dependencies and Tatoeba), and the language identifiers (ISO code or name) are matched with the Glottolog inventory. Details of the considered languages are shown in Table 1[7].

4.3 Results

First, we look into how the NLP literature has considered endangered languages across time. Figure 1 shows that, in the current century, there is a considerable growth of publications for languages across different revitalisation status. For instance, articles about languages with shifting or threatened status have increased from ten to a hundred papers annually, but there is a very shy increase of the moribund or nearly extinct languages (from zero to ten annually), which are the most endangered ones. This is highly contrasted by the continuous increment of NLP publications for not endangered languages (from hundreds to thousands annually).

Then, we observe the overlap of the language coverage between the ELAR database, the ACL Anthology and the language inventory of massive multilingual datasets above-mentioned. Figure 3 shows the cross-over in a map. The very low overlapping was expected: from the ELAR inventory

[5]We do not consider the extinct languages in our analysis

[6]We are aware that this was not an extended practice previously, but the Bender's Rule (Bender, 2011) has remarked it recently. Moreover, if a work does not specify which language is working on, we can expect the target to be English or very well-known established multilingual datasets.

[7]Data is published in https://github.com/aoncevay/cld2

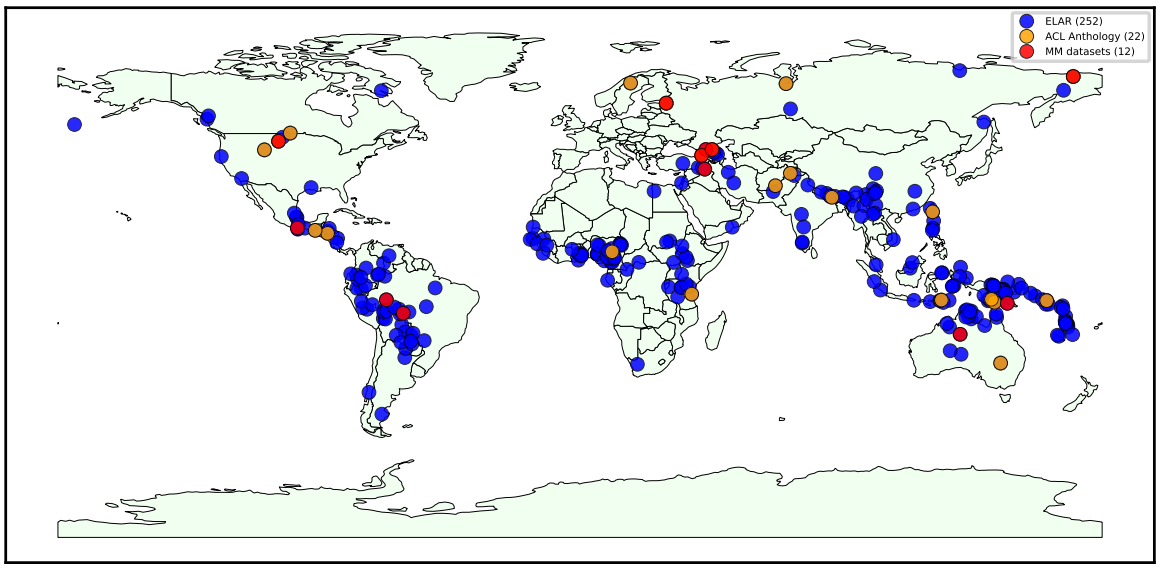

Figure 3: World map with languages in ELAR database and ACL Anthology. For the the present study, we only consider the languages of the ELAR database (570), whose names appear in *Glottolog* (version 4.4). This selection consists in 286 languages with geographical information. With this, 252 languages only belong to ELAR database (in blue); 22 languages belong to both ELAR database and ACL Anthology (in orange); and 12 languages belong to both ELAR database and massively multilingual (MM) datasets (Unimorph, Universal Dependencies and Tatoeba) (in red).

(286)[8], there are only 22 languages with at least one entry in the ACL Anthology (7.7%), and also 12 languages from this inventory included in at least one massive multilingual NLP dataset (4.2%). This two lists of languages overlaps only in 5 languages (Lakota, Laz, Chechen, Chukchi and Ingrian). Moreover, the geo-localisation allows us to observe the potential of these under-utilised resources in terms of representation for NLP research. Geographical areas such as the Americas, Africa, South-East Asia or Australia are better covered by language documentation projects than NLP resources and studies. Regional initiatives, such as Masakhane for Africa (Nekoto et al., 2020), or AmericasNLP (Mager et al., 2021), must look towards these still unexplored resources for extending their language coverage.

4.4 Discussion

The NLP community is recently more aware of the importance of language diversity in their research (Bender, 2009, 2011). Typologically-diverse language data allows to discuss results more broadly and to identify potential flaws of the proposed methods in languages with typologically uncommon grammatical properties and categories (O'Horan et al., 2016; Ponti et al., 2019). Furthermore, it has been pointed out that minority languages are indeed expected to exhibit unusual typological trends and non-prototypical degrees of complexity (Trudgill, 2011, 2010). Therefore, accessing and processing databases of a wide sample of endangered languages data would be beneficial for the NLP agenda.

However, as we observed, this has not been a priority. Why? We argue that this is mainly because of the visibility, accessibility, and readability of the data (from the NLP perspective):

Visibility Language documentation archives are mostly known in the linguistic community. The NLP community should look for data beyond the usual repositories. Besides ELAR, other famous repositories are the Archive of the Indigenous Languages of Latin America (AILLA)[9] from the University of Texas, The Language Archive (TLA)[10] from the Max Planck Institute for Psycholinguistics, and the Pacific and Regional Archive for Digi-

[8] We do not consider all languages in ELAR inventory (570) because languages in ELAR database are identified in most cases only by their names (and not by ISO codes), which match with the Glottolog database for 307 languages.

[9] https://ailla.utexas.org/

[10] https://archive.mpi.nl/tla/

tal Sources in Endangered Cultures (PARADISEC). [11]

Accessibility Most of the language documentation databases are open-source, but one often needs to become a registered user in order to access the materials deposited in the language archives. Furthermore, some linguists block fully public access to their records as a way to protect speech community's rights.

Readability Although most language documentation outputs video, audio and text files (plain texts or interlineal glossed texts, known as IGT), they are not labelled or processed for immediate use for NLP developments. If we observe the example in Figure 2, we can quickly identify potential resources for morphological segmentation and analysis, part-of-speech tagging, and machine translation. However, IGT is partially standardised, as not all the annotations follow the same label schema.

In sum, NLP is not taking advantage of all the resources potentially available for different applications. Moreover, from the three previously explained factors, readability is the hardest to overcome. One of our takes in this paper is to push the NLP community to focus more on the parsing and processing of the already published data, which is unlikely to be modified, unfortunately[12]. For instance, there should be paid more attention to IGT parsing research (Lewis and Xia, 2010; Round et al., 2020) or to the establishment of a more universally-readable IGT schema (Palmer and Erk, 2007). All this is complementary to the last point of Section 3, as we expect that, ideally, future deliverables of documentation projects could consider the annotation schema and resources that are more easily readable for NLP research.

5 CLD²: Computational Language Documentation and Development

Computational linguistics and language documentation share not only the assumption that technology plays an important role in the design and development of language-related projects, but also a crucial concern about language endangerment and loss. This concern is obvious from the perspective of language documentation, in the sense that it assumes itself as a response to language endangerment Himmelmann (2006, 5). A similar shift towards minority languages can be found in contemporary approaches to computational linguistics. Berment (2002) regrets that less than 1% of the world's languages have been correctly "computerised". That is, for Berment (2002), the fact that 99% of the world's languages lack computational tools (NLP tools as spell-checking or machine translation) requires immediate attention. Since the seminal article by Krauss (1992), language endangerment and language dormancy is a major concern for both current language documentation and computational linguistics.

This paper takes the shared interest in linguistic diversity found in language documentation and computational linguistics further by proposing a paradigm that assumes an intense and multifaceted interaction between the two: Computational Language Documentation and Development (CLD²). CLD² assumes, following (Berment, 2002), that "computerisation" should be understood as one main task in language documentation and, at the same time, proposes a basic protocol to carry out this task. This basic protocol is based on a straightforward idea according to which any documentation project, in addition to its customary outcomes (audio and video recordings, transcriptions, morphological parsing and glossing, and free translations), should include NLP-friendly annotated data as its deliverables:

1. Monolingual and parallel corpora[13] in a digital format, ideally taken from a specific domain or discourse that is relevant for the language speaker community;
2. A public representative set of sentences annotated in universal frameworks for morphology and syntax, such as Universal Morphology (McCarthy et al., 2020) and Universal Dependencies (Nivre et al., 2020)[14], which are well-known in the NLP field; and
3. A communication describing the main characteristics of the released Universal Depen-

[11] https://www.paradisec.org.au

[12] Most of the language documentation projects that are published might do not have extra funding allocated for any update, or new funding will be required for the job.

[13] Translations paired with English or another relevant language spoken in the specific region, such as Spanish in Latin America.

[14] The identification of syntax dependencies and their annotation is not common in language documentation projects. However Croft et al. (2017) have argued that the UD scheme shares crucial principles with typological research. Indeed, research on linguistic typology may benefit from the development of an annotation scheme like UD and vice-versa.

dencies (Nivre et al., 2020) treebank and Universal Morphology (McCarthy et al., 2020) dataset, so that NLPers can understand the particularities and challenges of the data.

We attempt then to draw documentary and computational linguists' attention towards the potentialities of a more integral and systematic collaboration between them. On the one hand, field linguists may get involved in creating relevant products from the NLP perspective (e.g. preparing representative treebanks taking as a starting point their own data). On the other hand, NLPers can get involved in the development of processes and protocols that may contribute to the transformation of linguistic data of the traditional sort into formats that may support NLP developments.

According to Forcada (2006, 1), one feature for a language to be considered as a minor one is the few to zero availability of machine-readable resources. There are features such as the number of speakers or literacy speakers that may support the definition of a minor language in a general overview, but we want to emphasise the computational perspective in Forcada's statement. Dictionaries, translated text or annotated corpora, that are currently part of a standard language documentation process, are instances of machine-readable data. We consider that linguistic corpora are insufficient to disentangle the relationship between a language and its characterisation as a minor language. We claim the need to develop more multiple resources to support a consistent revitalisation of the language. However, we do not mean that all language documentation processes should include a massive technology development by itself. The magnitude of such a project would be cost-prohibitive. Nevertheless, we have identified some elements that might be included in a documentation process that could drive a "computerisation" effect in the studied language.

We want to emphasise the development of multipurpose linguistic databases, specifically aiming at language technologies, whose implementation will not radically increment the amount of expected work for the linguist. Language technologies are purpose-specific programmes that try to address language-related tasks from spell- or grammar-checking to automatic machine translation. Based on such databases, NLPers and field linguists may work together to develop NLP toolkits for minority languages. An NLP Toolkit is a set of different tools made to computerise a language fully. We then take inspiration from the Basic Language Resource Kit (Krauwer, 2003) and also consider established annotation frameworks, such as UD or UniMorph, and current state-of-the-art methods in NLP, such as transfer learning. With transfer learning protocols, especially multilingual pretraining (Lauscher et al., 2020; Ebrahimi and Kann, 2021), CLD² projects might automatise learning tasks by taking advantage of larger amounts of multilingual data and tools. A learning task in this context may refer to a specific NLP or functionality, such as a dependency parser, which has been trained to learn how to parse the syntax in a textual sentence. Finally, we list the main tools that such basic toolkits could have:

1. Morphological tools: such as morphological analysis, to determine the base form or lemma of an inflected word and its morphological features; morphological segmentation, to identify the canonical or surface morphemes (Mager et al., 2020); and morphological reinflection (Pimentel et al., 2021), which exploits UniMorph data. Morphological knowledge is usually crafted in language documentation projects (see Figure 2), so these deliverables could be the most manageable.

2. Spell-checker: to detect and automatic correct of spelling errors. Dictionary-based spell-checkers can be easily retrieved from a documentation project with a lexicon as an output, whereas rule-based ones can be adapted from a finite-state morphological analyser. Data-driven spell-checking is also possible to develop from monolingual data only.

3. Syntactic parser: to analyse the relationships between the words and phrases that compose a text. A dependency syntax parser can be developed using UD annotated data, and is also benefited for transfer learning and pretraining approaches (Lauscher et al., 2020). Current language documentation projects do not usually focus on this kind of annotation, but we emphasise that it might be relevant for research not only on NLP but also in linguistic typology (Croft et al., 2017).

4. Part-of-Speech tagger and Named Entity Recognition: both tasks are sequence taggers, and are two of the tasks that have been benefited the most from multilingual pretraining, and few- or zero-shot learning (Lauscher et al.,

2020; Ebrahimi and Kann, 2021). POS tagging could be easily adapted from the current glossing annotation, whereas NER annotation can be quickly extended or marked in the glosses.

Besides these tools, further developments that can be achieved for endangered languages, such as machine translation, are very appealing. However, we also need to point out that, despite the progress of the pretraining approaches and the use of few labelled examples, a translation system (or other kinds of NLP tools) should not be deployed with low-quality outputs, as it can mislead the user. Limitations of their usage should be assessed according to the annotated data used and the purpose of the systems.

6 Conclusion

CLD² calls for an enrichment of language documentation projects by means of incorporating components, outcomes and methods from NLP research, as a strategy to promote the computarisation and revitalisation of minority languages. This paper shows that most of the interactions between computational linguistics and language documentation are framed as software developments that facilitate the various processes involved in documenting a language. The potential contributions of language documentation and language repositories to NLP research are under-exploited and deserve urgent attention from the NLP community. At the same time field linguists may also incorporate into the outcomes of their projects, data crafted into paradigms that can be automatically used for NLP developments (Universal Dependencies and/or Universal Morphology, for instance).

This will benefit not only language documentation and computational linguistics scholars but also typologists and speech communities, as research in NLP has recently paid some attention to linguistic typology as a substantial source of linguistics knowledge to improve performance in different algorithms and technologies (O'Horan et al., 2016; Ponti et al., 2019). Indigenous communities, in turn, are highly enthusiastic about the computerisation of their languages as a political strategy that vindicates their languages and demonstrates that they are as valuable as major European languages. CLD² can significantly contribute to this aim by promoting productive exchanges among field linguists, NLP researchers and members of indigenous communities as part of multi-component projects that put language revitalisation at their core.

7 Acknowledgements

The first author acknowledges the support of CONCYTEC-ProCiencia, Peru, under the contract 183-2018-FONDECYT-BM-IADT-MU from the funding call E041-2018-01-BM.

References

Antonios Anastasopoulos, Christopher Cox, Graham Neubig, and Hilaria Cruz. 2020. Endangered languages meet Modern NLP. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 39–45, Barcelona, Spain (Online). International Committee for Computational Linguistics.

Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer, and Lane Schwartz, editors. 2017. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, Honolulu.

Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer, Lane Schwartz, and Miikka Silfverberg, editors. 2019. *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*. Association for Computational Linguistics, Honolulu.

Peter K Austin. 2006. Data and language documentation. *Essentials of language documentation*, 178:87.

Peter K. Austin. 2010. Communities, ethics and rights in language documentation. In Peter K. Austin, editor, *Language documentation and description*, volume 7, pages 34–54. London: School of Oriental and African Studies.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Vincent Berment. 2002. Several directions for minority languages computerization. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.

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.

Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets universal dependencies. In *TLT*, pages 63–75.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Mikel Forcada. 2006. Open source machine translation: an opportunity for minor languages. In *Proceedings of the Workshop "Strategies for developing machine translation for minority languages", LREC*, volume 6, pages 1–6.

Jeff Good, Julia Hirschberg, and Owen Rambow, editors. 2014. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Harald Hammarström, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A. Westenberg, and Bettina Speckmann. 2018. Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation*, 12:359–392.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. Glottolog 4.4. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at `http://glottolog.org`. Accessed on 2021-05-20.

Nikolaus Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Nikolaus Himmelmann. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation & Conservation*, 6:187–207.

Nikolaus P Himmelmann. 2006. Language documentation: What is it and what is it good for. *Essentials of language documentation*, 178(1).

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):1–10.

Steven Krauwer. 2003. The basic language resource kit (blark) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM*, volume 2003, pages 8–15.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Gina-Anne Levow, Emily M. Bender, Patrick Littell, Kristen Howell, Shobhana Chelliah, Joshua Crowgey, Dan Garrette, Jeff Good, Sharon Hargus, David Inman, Michael Maxwell, Michael Tjalve, and Fei Xia. 2017. STREAMLInED challenges: Aligning research interests with shared tasks. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–47, Honolulu. Association for Computational Linguistics.

William D Lewis and Fei Xia. 2010. Developing odin: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*, 25(3):303–319.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250, Online. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Max Planck Institute for Psycholinguistics. 2021. *ELAN (Version 6.2)*. The Language Archive, Nijmegen. Https://archive.mpi.nl/tla/elan.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe,

Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexis Palmer and Katrin Erk. 2007. IGT-XML: An XML format for interlinearized glossed text. In *Proceedings of the Linguistic Annotation Workshop*, pages 176–183, Prague, Czech Republic. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Erich Round, Mark Ellison, Jayden Macklin-Cordes, and Sacha Beniamine. 2020. Automated parsing of interlinear glossed text from page images of grammatical descriptions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2878–2883, Marseille, France. European Language Resources Association.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4e):324–345.

Summer Institute of Linguistics. 2021a. *Field linguist's Toolbox (Version 1.6.4)*. Http://www.fieldlinguiststoolbox.org/?i=1.

Summer Institute of Linguistics. 2021b. *Fieldworks (Version 9.0)*. Https://software.sil.org/fieldworks/.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Peter Trudgill. 2010. Contact and sociolinguistic typology. In Raymond Hickey, editor, *The Handbook of Language Contact*, pages 299–319. Oxford: Wiley-Blackwell.

Peter Trudgill. 2011. *Sociolinguistic Typology: social determinants of linguistic complexity*. Oxford: Oxford University Press.

Daan van Esch, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 14–22, Honolulu. Association for Computational Linguistics.

Anthony Woodbury. 2011. Language documentation. In Peter Austin and Julia Sallabank, editors, *Handbook of Endangered Languages*, The Cambridge Handbook of Endangered Languages, pages 159–186. Cambridge: Cambridge University Press.

Anthony C Woodbury. 2003. Defining documentary linguistics. *Language documentation and description*, 1(1):35–51.

A AES status for massively multilingual datasets

AES status	Tatoeba	Unimorph	UD
not endangered	164	60	52
threatened	71	25	16
shifting	44	17	16
moribund	11	4	2
nearly extinct	7	4	1
extinct	24	17	11

Table 1: Agglomerated Endangerment Status (AES) (Seifart et al., 2018) statistics for MM databases (Tatoeba, Unimorph and Universal Dependencies).