Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2657–2666
Marseille, 11–16 May 2020

Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo)

Ludger Paschen[†], François Delafontaine[‡], Christoph Draxler[*], Susanne Fuchs[†], Matthew Stave[‡], Frank Seifart[†]
Leibniz-Zentrum Allgemeine Sprachwissenschaft[†], Laboratoire Dynamique Du Langage[‡], Bavarian Archive for Speech Signals[*]
Schützenstraße 18 10117 Berlin, 14 Avenue Berthelot 69007 Lyon, Schellingstraße 3 80799 München
{paschen, fuchs, seifart}@leibniz-zas.de, {matthew.stave, francois.delafontaine}@cnrs.fr, draxler@phonetik.uni-muenchen.de

Abstract

Natural speech data on many languages have been collected by language documentation projects aiming to preserve linguistic and cultural traditions in audiovisual records. These data hold great potential for large-scale cross-linguistic research into phonetics and language processing. Major obstacles to utilizing such data for typological studies include the non-homogenous nature of file formats and annotation conventions found both across and within archived collections. Moreover, time-aligned audio transcriptions are typically only available at the level of broad (multi-word) phrases but not at the word and segment levels. We report on solutions developed for these issues within the DoReCo (DOcumentation REference COrpus) project. DoReCo aims at providing time-aligned transcriptions for at least 50 collections of under-resourced languages. This paper gives a preliminary overview of the current state of the project and details our workflow, in particular standardization of formats and conventions, the addition of segmental alignments with WebMAUS, and DoReCo's applicability for subsequent research programs. By making the data accessible to the scientific community, DoReCo is designed to bridge the gap between language documentation and linguistic inquiry.

Keywords: corpus creation, endangered languages, phonetic databases

1. Introduction

Temporal patterns of speech are of central interest in the cognitive sciences as they provide key evidence for the architecture underlying the human language production system in terms of its cognitive-neural and physiological-articulatory bases (Jaeger and Buz, 2017). But most current approaches draw on phonetic measurements from only less than one percent of the world's languages, most of these (Indo-) European (Anand et al., 2015; Norcliffe et al., 2015). It has repeatedly been pointed out how problematic generalizations based on data from such a thin slice of the human population are (Henrich et al., 2010). Specifically, previous research has had few means to assess whether observed temporal patterns in speech are characteristic of human cognition and articulation in general or specific to just the few languages studied so far, predominantly English. For instance, recent claims on the neural basis for a preferred syllable rate (Assaneo and Poeppel, 2018) and on maximum achievable rates (Ghitza, 2014) still await systematic testing on a broad cross-linguistic sample. DoReCo addresses this problem by compiling a dataset for comparative studies of spoken language in a diverse sample of at least 50 languages.[1] We identify and extract suitable data from language-documentation collections (Himmelmann, 1998) archived at established repositories, and process these into a multilingual reference corpus consisting of one million words of transcribed, translated, and time-aligned corpus data with associated audio recordings. We carry out exemplary studies using this resource and make it available for future research. As the United Nations declared 2019 the International Year of Indigenous Languages, and 2022–2032 the International Decade of Indigenous Languages, DoReCo emphasizes the importance of global linguistic diversity and cultural heritage by enhancing the visibility of less resourced and often endangered languages and ensuring they are taken into account in the cognitive sciences.

2. The Pipeline

The workflow for turning language documentation data into a mineable time-aligned corpus within DoReCo is summarized in Figure 1 below.

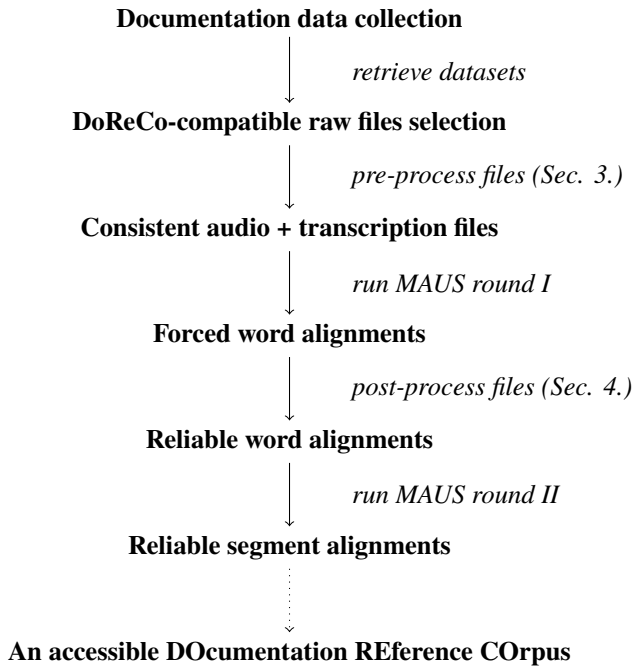

Figure 1: DoReCo pipeline.

[1] http://doreco.info

Each individual step will be discussed in more detail in the following sections, as indicated in the figure. Two research areas where DoReCo will be immediately useful – segmental length and information rate – will be presented in Section 5. Note that the main focus of this paper is on data selection, alignment and standardization; questions of curation and distribution will be briefly touched upon in Section 6.

3. Pre-processing

3.1. Sources

Most of the datasets in DoReCo stem from documentation initiatives of smaller and often endangered languages.[2] Table 5 lists the 40 languages that have been selected for inclusion in the corpus at the time of writing. While the selection is to some extent a convenience sample based on the availability of collections containing data that meet the specific requirements of DoReCo (Table 1), the corpus envisions a level of genealogical and areal diversity suitable for meaningful typological research. At the same time, we intend to take into account various registers and genres, ranging from personal and traditional narratives to procedural descriptions and multi-party conversations as long as they comply with the audio quality criteria listed below.

Media files	
1)	Good audio quality
2)	Little or no overlapping speech
Annotations	
3)	Minimum of 10k transcribed words
4)	Time alignment of annotation units (AUs)[3]
5)	Translation into majority language
6)	Interlinear morphological analysis (optional)
Metadata	
7)	Speakers: Identifiers, age, gender
8)	Recordings: Duration, sound quality
Commitment of corpus creators	
9)	Provide general background on language
10)	Provide grapheme-to-phoneme mapping table
11)	Clarify any issues during pre-processing
12)	Assist with identifying disfluencies
13)	Assist with identifying code-switching
14)	Data can be made available under CC BY-NC

Table 1: Requirements for inclusion in DoReCo.

Speech data are retrieved both from private collections and from deposits such as ELAR (https://elar.soas.ac.uk/), TLA (https://tla.mpi.nl/), PARADISEC (http://www.paradisec.org.au/), and Pangloss (https://lacito.vjf.cnrs.fr/pangloss/).

A fundamental principle of DoReCo is to utilize already existing materials rather than to create new recordings or annotations specifically for the purpose of the project. To that end, DoReCo relies on the cooperation of each of the institutions and individuals involved, and builds on existing scientific networks and personal contacts.

Since the working format during pre-processing is ELAN's .eaf format (ELAN developers, 2019), annotations that come in a different format have to be converted to .eaf before they can be further processed (see Section 3.3.). DoReCo will implement a versioning system to ensure interoperability between original files, files created in the course of DoReCo, and potential future user-generated annotations.

3.2. Consistency checks

During pre-processing, each file is inspected manually for more than 30 properties, including file format, naming conventions, sound quality, status of interlinearization, completeness of transcription and translation, alignment quality, and presence of non-alignable elements, e.g. in-line comments such as "[cough]". We perform conversions to the working file formats for pre-processing, which are .wav and .eaf, whenever necessary, and fix references to missing or alternatively named files in the MEDIA_DESCRIPTOR node in the .eaf file header.

In addition, a number of semi-automatic consistency checks are performed using custom-made scripts, in particular for validating ELAN's .eaf file format and finding undefined characters. The former is relevant because the aligner we use expects .eaf files to follow a specific xml template (The Language Archive, 2017), violations of which cannot be easily identified from within ELAN's GUI. The latter is important because the aligner maps every character (or combination of characters) on the transcription tier to a phoneme based on a language-specific correspondence table and cannot handle undefined characters. Missing character definitions must not only be avoided to prevent runtime errors but also to ensure a successful mapping to phonemes. During character checking, we regularly encounter inconsistencies such as occurrences of both a combined character and a sequence of base character plus diacritic for what appears as the same symbol to the human eye (e.g. U+0101 vs. U+0041 U+0304 for <ā>). Here, we consult with the respective corpus creator(s) whether these variants should be standardized in the source files themselves, or whether the variation should be captured by correspondence tables.[4] Character definition checks also have the potential to reveal other inconsistencies within collections, as transcription conventions may change over the time span of a documentation project; feedback on the exact units that contain potentially problematic content is provided as a service to the corpus creators.

While DoReCo will ultimately provide a transparent and unified system of file and tier names, it has proven useful to stick to the original naming conventions during preprocessing. The main reason for this is that it greatly facilitates communication with data donors, who can immediately react to any questions that may arise during preprocessing when they are formulated within the naming conventions that they are familiar with. We keep track of all

[2] A few majority languages such as Arabic, English, French, and German will likely be included in DoReCo as well to further diversify the sample.

[3] Annotation units are set during transcription, mostly for practical purposes, and roughly correspond to utterances. The average size of annotation units across corpora varies between about three and eight words.

[4] The former has obvious benefits for the original corpus, while the latter is easier to implement because mapping files need to be created independently. DoReCo offers assistance to corpus creators wishing to pursue the former path.

original file and tier names and their semantics, as this information is crucial for preparing the files for the first round of MAUS (Section 3.4.).

3.3. File formats and conversions

Data collection involves a plurality of file formats originating from different annotation tools like ELAN's .eaf (Sloetjes and Wittenburg, 2008), Praat's .TextGrid (Boersma and Weenink, 2019), or Pangloss' .xml (Michailovsky et al., 2014). These formats are converted into .eaf for DoReCo's internal procedures, i.e. for all pre-processing operations described in this section. For manual re-alignments, we use Praat's .TextGrid as a working format (Section 4.1.).

File conversions are carried out either with ELAN's import functions or with a custom-made conversion tool, which will be made publicly accessible both for reproducibility and general use at the end of the project. Our tool follows a "Swiss-army knife" logic similar to Pepper (Zipser and Romary, 2010), with an intermediate model from and to which all considered formats are converted. This intermediate model also allows for maximum control over any aspect of the data, including merging, replacing, renaming, etc. Our model is more semantic and thus more restrictive than Pepper's model, though, trading flexibility for more control over how translations are performed.

At the end of the DoReCo pipeline, annotation files are converted to a variety of formats and delivered to the linguistic community. This includes the above-mentioned formats, but also the TEI standard (Schmidt, 2011) and a tabular format. A major concern is to ensure that no original data is lost through the process, notably by injecting additional annotations provided by DoReCo back into the original files. Additional conversions are done for creators who have provided data in formats such as Toolbox (Buseman and Buseman, 2019) to ensure they have full access and use of the enhanced data.

3.4. MAUS round I

WebMAUS is a public web service to automatically time-align orthographic text with a given signal (Kisler et al., 2012). It is based on MAUS (Munich AUtomatic Segmentation) (Schiel, 2004). The general principle of MAUS is that a pronunciation hypothesis graph is generated from a sequence of phonemes created from an orthographic transcript using a grapheme to phoneme converter. The signal file is then force-aligned with the hypothesis graph and the path with the highest overall probability is chosen. MAUS has been shown to achieve 95% agreement with human annotators (Kipp et al., 1997).

In the first round of forced alignment within DoReCo, .eaf files that have gone through pre-processing are used. They contain at least one tier with an orthographic transcript within so-called annotation unit (AU) intervals. These annotation units in general consist of several words. For each of these annotation units, the word boundaries are computed automatically using WebMAUS.

MAUS supports more than 20 languages, and it features a language independent mode, which is the mode we use in DoReCo. Furthermore, languages not (yet) supported directly by MAUS can be processed by providing a mapping file which maps orthographic strings to phonemic representations in SAMPA. For DoReCo, a specific web service has been implemented by extending the existing WebMAUS service. These extensions allow

- the use of .eaf files as input,
- specifying which tiers from the .eaf file to process,
- custom mappings for the grapheme to phoneme conversion (g2p), and
- the option to exclude certain tokens from processing, e.g. in-line comments.

The custom mappings and exceptions are language-specific, and they are provided as .csv or .txt files. The relevant tier name(s) of the .eaf source files are given as a .csv file. Files are submitted to the DoReCo web service either via the graphical user interface or via a command line call using `curl`. The latter is most commonly used for batch processing.

For processing, the .wav audio and .eaf annotation files are uploaded to the server. The server then does some minor preprocessing (e.g. removing blanks from filenames), checks whether all required files have been provided, and then extracts annotation unit segments from the specified tiers in the .eaf file. The service can handle simple parent AUs with discrete timestamps as well as dependent AUs or AUs with symbolic subdivision. The extracted segments are used to perform a chunk-wise grapheme to phoneme conversion and the actual MAUS segmentation. The result is a .TextGrid file containing consistent segmentations on three levels: orthography, canonical form, and phonemes. Only the word intervals on the orthography tier will be edited in Praat, while the other tiers are given for reference to facilitate corrections.

4. Post-processing

4.1. Manual word boundary correction

By *post-processing* we understand a range of manual corrections and refinements that are carried out following the first round of forced alignment. The central component of post-processing is manually checking the word start- and endpoints produced by WebMAUS and readjusting them where necessary.

The WebMAUS service infers word boundaries from forced segment alignments given phrase-size annotation units as input. Several factors may impact the accuracy of these alignments, and hence the amount of manual corrections (for more details on inter-annotator agreement, see Section 4.3.). As a general rule, alignment quality is best for (i) short AUs, (ii) AUs with no internal pauses, and (iii) good sound quality, which particularly refers to the absence of irregular background noise[5]. Figure 2 shows an extreme case of gross misalignment when all of these factors are combined. Additionally, certain segment types and combinations are intrinsically challenging for forced alignment tools (Gonzalez et al., Ms). Within DoReCo, the language-independent

[5] MAUS has proven surprisingly robust against the effect of continuous background noise such as constant bird singing.

model of MAUS was found to frequently require manual refinements in the case of voiceless stops, sibilants with high-frequency noise (see Figure 3), and vowel-nasal sequences.

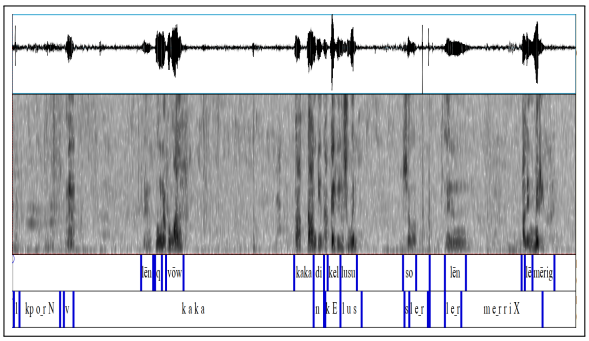

Figure 2: Praat editor window showing word boundaries resulting from forced alignment (bottom tier) and manual correction (top tier) for a recording from the Vera'a corpus. Severe readjustments were necessary due to the length of the AU (14 sec), several pauses, and irregular background noise.

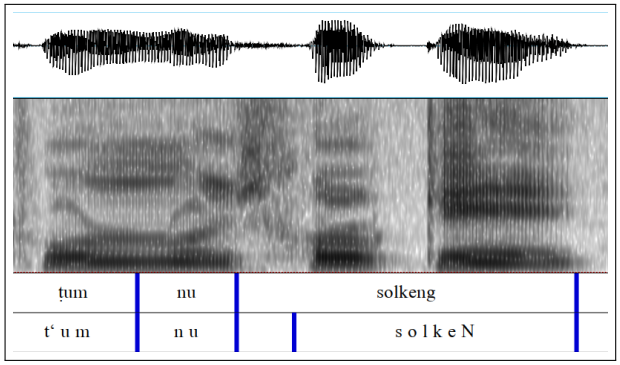

Figure 3: Praat editor window showing results of forced alignment (bottom tier) and manual correction (top tier) for a recording from the Anal corpus. The forced aligner misinterpreted the initial part of the turbulent noise in the sibilant /s/ as a pause. This was corrected by merging the empty interval with the one containing the fricative.

All team members involved in post-processing received rigorous phonetic training. Inter-annotator agreement was monitored and yielded satisfying results (Section 4.3.). Overall, the manual correction of word intervals is the most time-consuming and labor-intensive part of the whole DoReCo enterprise. However, the huge effort can be justified in light of the tremendous potential and impact that we believe a cross-linguistic corpus of spoken language equipped with reliable word and segment alignments has to offer.

4.2. Marking non-alignable units

The second main objective of post-processing is identifying portions in a recording that are not suitable for segmental alignment, either due to internal factors (disfluencies, code switching) or external factors (overlaps, heavy background noise). An overview of all the labels used for marking non-alignable portions is given in Table 2. The final column illustrates (using fictitious examples) wrapping pre-existing transcriptions in the labels; if transcriptions are not available, the labels are inserted as they are, replacing any previous content in the respective interval. Our labels make use of an inbuilt feature of WebMAUS that ignores characters within angle brackets. By labelling, we ensure every speech event in the audio signal is mapped to some entry on the transcription tier, be it parsed or not.

Type	Label	Example
Filled pause	<<fp>>	<<fp> *uhm* >
Prolongation	<<pr>>	<<pr> *looonger* >
False start	<<fs>>	<<fs> *fal–* >
Ideophone	<<id>>	<<id> *tick* >
Onomatopoeic	<<on>>	<<on> *moo* >
Foreign material	<<fm>>	<<fm> *Weberei* >
Unidentifiable	<<ui>>	<<ui>?? >

Table 2: Labels for disfluent or non-alignable speech.

It should be stressed that the present classification system is merely a practical solution developed for the purpose of marking units that are problematic for alignment. The authors are aware of the difficulties surrounding the taxonomy of disfluencies and the issue of at times unclear or conflicting definitions; see Shriberg (1993), Eklund (2004), Belz (2017), and Brugos et al. (2019) for discussion. We are confident that our system is sufficient to serve the goal of marking segments for exclusion from the second round of forced alignment. As a byproduct, general pointers to disfluent speech will be available alongside alignments to invite further cross-linguistic research into these elements. The <<ui>> label is reserved for unidentifiable stretches of speech, caused by overlaps, external noise such as rainfall or thunder, or other factors that made certain words or phrases unintelligible even to the language expert delivering the transcriptions. This label is also used for songs and stylistically marked speech such as poetic recitation, for the simple reason that transcriptions for these parts are often not included in the transcriptions to begin with.

4.3. Inter-annotator agreement

Several measurements of inter-annotator agreement were performed in order to evaluate the degree to which human annotators agree among each other and the amount of adjustments that were made to the forced alignments produced by WebMAUS. Corrections performed by human annotators comprise (i) the displacement of units, i.e. shifting word boundaries, (ii) the addition or removal of units, and (iii) editing of a unit's content. We will focus here on displacement, addition, and removal of intervals.

Table 3 shows the extent to which human annotators agree on whether or not and where to add or remove intervals. The numbers reflect a test sample consisting of 6 annotation files from 3 different languages (2 files per language) with a total of 4125 orthographic words, each processed by four human annotators (H1–H4). We can observe an average of 93–96% matching units across annotators and languages, indicating a high reliability for the labelling task.

When looking only at matching units, i.e. word intervals

IAA	Anal	Resígaro	Vera'a	Total
H1-H2	93.46%	95.48%	98.33%	95.76%
H1-H3	92.43%	95.93%	98.03%	95.46%
H1-H4	90.23%	94.88%	94.46%	93.19%
H2-H3	95.25%	95.32%	98.30%	96.29%
H2-H4	92.08%	94.35%	94.54%	93.66%
H3-H4	90.30%	94.42%	94.20%	92.97%

Table 3: Agreement on unit addition/removal.

that annotators do agree on, we can determine the amount of overlap between those units. The results, aggregated over all languages, are given in Table 4. The table shows that human annotators (columns 2–5) agree for 91-95% of unit size and position, while agreement between humans' and MAUS' forced alignments is at 82–86%.

IAA	H1	H2	H3	H4	MAUS
H1	—	93.68%	94.56%	92.34%	84.24%
H2		—	93.90%	91.32%	82.04%
H3			—	92.37%	83.61%
H4				—	85.58%

Table 4: Agreement on unit overlap.

The following histograms display the distribution of distances between moved boundaries for matching units. Figure 4 shows a quasi-normal distribution for humans vs. MAUS with the majority of items in the range of 10–100 ms and a substantial amount of items up to the 1000 ms mark. This distribution is expected: When a human annotator shifts a boundary, it is either due to minor errors at the segment level (displacement usually less than 100 ms) or due to serious misalignments (displacement up to several seconds). Figure 5 shows a positive skew for inter-human agreement, with more items in the region between 1–5 ms and substantially fewer items beyond the 100 ms mark compared to Figure 4. This shows that humans agree most of the time about the region to which a boundary should be moved, with the obvious caveat that agreement in the range of 10 ms or less is nearly impossible to achieve.

4.4. MAUS round II

Having manually corrected word start and end times, and having identified filled pauses, code-switching etc., the recordings are once more processed by MAUS. In this second round of forced alignment, MAUS computes segment alignments based on the manually corrected word intervals. The constraints for the segmentation differ from the first round. Now, the input is the manually corrected .TextGrid file. Furthermore, it is assumed that word segments at this stage do not contain any pauses, so MAUS is configured to not insert new pause segments.[6] Effectively, in this round MAUS simply realigns the phoneme boundaries within the boundaries of every word in the selected tiers.

[6] Occasionally, silent pauses are also found within words, in particular in highly synthetic languages like Arapaho. In such cases, a silent interval is added and the transcribed word is split accordingly during post-processing.

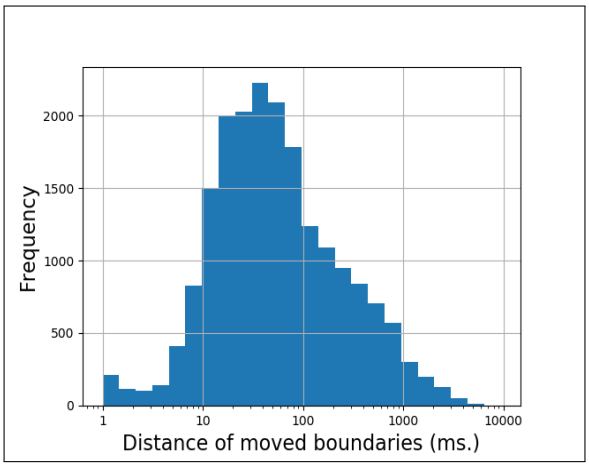

Figure 4: Frequency distribution of distances between MAUS-aligned and manually adjusted word boundaries. Unchanged boundaries (distance = 0) make up for 15.1k or 45.70% of all boundaries (Σ = 33k).

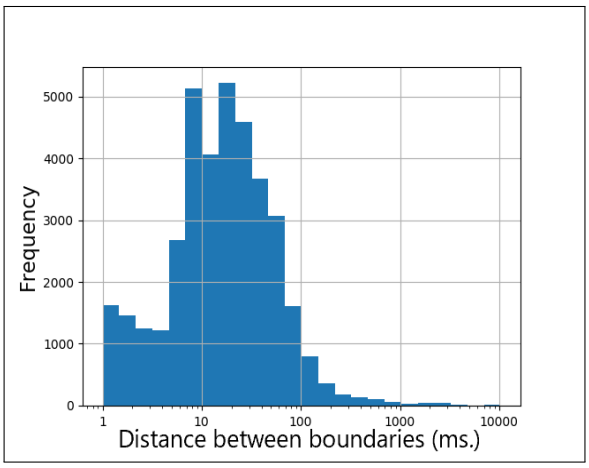

Figure 5: Frequency distribution of distances between word boundaries among four human annotators. Unchanged boundaries (distance = 0) make up for 15.4k or 31.05% of all boundaries (Σ = 49.5k).

At the end of the second round, the resulting .TextGrid file contains the original annotation unit segments, the word segments within each annotation unit, and phoneme segments within each word. It is then ready to be reintegrated into the .eaf working format (see Section 3.3.).

As a final note, ELAN allows the definition of dependencies between tiers. For example, a dependent tier can be defined to fill the interval specified by the parent tier without gaps. Technically, this is achieved by sharing timepoints between annotation levels in the .eaf file. In DoReCo, the boundaries on additional tiers generated through the process are independent of each other and independent of other tiers. This grants maximum flexibility with respect to any subsequent processing steps.

5. DoReCo research projects

In this section, we outline two broad research areas for which DoReCo's time-aligned transcriptions and annotations will be particularly useful.

5.1. Lengthening and compression

Speech production is a process that unfolds in time. Research into this process has investigated various time scales ranging from the subphonemic level to phones, syllables, words, phrases, utterances etc. Languages are known to differ substantially with respect to all aspects of speech production (rhythm, prosody, accent, stress, syllable structure, phonotactics, phoneme inventory); a comprehensive overview of timing and rhythm is given by Fletcher (2010). Some temporal aspects, however, might also be universal, given that speech production is tightly linked to the biological, physical and cognitive capacities of humans (Ohala, 1983; Lindblom, 1986; Maddieson, 2006). For example, MacNeilage (1998) suggests that the modulation between mouth opening and closing that constitutes syllable rhythm may have its origin in mastication, chewing and sucking. In addition, modulations of the jaw correlated with the amplitude envelope of the acoustic signal have been discussed in line with neural oscillations in the auditory cortex within a range of 2 to 7 Hz (Chandrasekaran et al., 2009).

The unified MAUS segmentations and the diverse set of languages in DoReCo will allow us to (i) provide solid cross-linguistic evidence regarding the assumed universality of final lengthening, and (ii) examine cross-linguistic differences in the degree of compressibility of segments. Final lengthening, i.e. increased segment duration before prosodic boundaries, has been reported for a variety of languages (Fletcher, 2010). It is unclear, however, whether it is uniformly observed in a broad sample of languages, and how it is related to speech planning and other factors. We will compare the average durations of segments preceding pauses with durations of the same segments in non-prepausal positions. We will additionally consider the extent of lengthening depending on segment type and other phonological properties.

With respect to compressibility, we will identify common consonants and vowels which occur across many DoReCo languages. We will then extract the duration of each of these segments together with the number of segments and morphemes of the word (to account for polysyllabic shortening), the position in the word (to account for positional effects), the phonetic context (to account for coarticulation), and the number of consonants and vowels in the respective phoneme inventory (to account for phoneme density). On the basis of these data we will consider the elasticity of various types of segments across languages. We expect voiceless fricatives to be less flexible than vowels and sonorants, for instance. Results will be discussed in light of theoretical accounts such as incompressibility limits (Klatt, 1979), as well as with respect to the effects of universal aerodynamics and motor constraints.

5.2. Speech rate and information density

It is widely believed that much of the complexity and efficiency of linguistic structure is motivated by two competing pressures: the pressure to maximize output, by expressing a lot of information quickly, and the pressure to minimize effort, by reducing processing costs in cognition, perception, and articulation. Analysis of these pressures has led to the formulation of universal claims about language production, such as the principle of Uniform Information Density (Aylett and Turk, 2004; Frank and Jaeger, 2008), which predicts that speakers will optimize their speech production to be always near the speech channel's maximum information capacity, and claims about language structure, such as the observation that the frequency and length of linguistic units follow a Zipfian distribution (Zipf, 1949; Haspelmath, 2004). Testing such universal claims of information transmission and structure is ideally done on a diverse set of well-annotated corpora, such as those found in DoReCo.

The morphological annotation in 30 of the DoReCo corpora provide the means to test some of these universal claims. Two main branches of research will involve (i) establishing whether there is an optimal global "attractor state" for information rate across languages and (ii) whether languages tend to package comparable amounts of information in interpausal units. Previous work has described similar information rates (information per syllable) for 17 languages (Coupé et al., 2019), using parallel corpora to establish a fixed quantity of information across the languages. We will expand on these results by comparing a different measure of information rate (morphemes per unit time) across 30 corpora and testing whether information rates cluster within an optimal attractor state that is fast enough to convey useful information and slow enough to limit communication costs. We expect information rate to vary according to language-specific morphological characteristics, with lower information rates for languages with highly fusional morphology and also for languages with a high degree of "hidden complexity". With regards to information packaging, we will test the hypothesis that languages tend to package equivalent amounts of information (morphemes) between pauses. We speculate that this will be true regardless of whether the language is more synthetic or analytic (that is, regardless of the number of words), perhaps as a result of a universal human cognitive preference for informational length of planning units.

In a related branch of research, we have utilized the morpheme annotations in corpora of nine languages in DoReCo to test the effects of two universal laws that make predictions about the length of linguistic units (Stave et al., 2020): Zipf's Law (Zipf, 1935), which predicts shorter units when the corpus frequency is higher, and Menzerath's Law (Menzerath, 1928), which predicts shorter units when the carrier unit is longer. These laws make similar predictions, but are based on quite different aspects of linguistic structure. We find that both laws contribute to predicting morpheme length, with Zipf's Law having around twice as much predictive power, and that the negative correlation between morpheme length and frequency is more pronounced in longer carrier words.

6. Outlook

With 40 languages already at advanced stages of processing eleven months into the project, DoReCo will soon be able to publish a fully processed corpus containing a diverse set of 50 languages with annotated time-aligned speech data of at least 10,000 words each. The data will be offered as downloads to registered users, who will need to accept agreements on fair use before being granted access. Data will be made available both as individual file downloads

and as customized packages using sophisticated filters via a web interface. We will also provide links to the deposits where the original files are archived and publish transparent guidelines on how contributors should be credited when using data collected by them. Making this resource available widens the cross-linguistic scope for research in corpus phonetics (Liberman, 2019), linguistic complexity, and other fields, beyond well-studied Western languages (Henrich et al., 2010). Since DoReCo's tools will be made publicly available, its workflow can be used for other low-resourced languages in the future if corpus creators are willing to invest the necessary time and effort. In the long run, DoReCo intends to set new standards for cross-linguistic corpora and encourage future language documentation activities to apply those standards when collecting new data.

7. Acknowledgements

This research is funded by DFG and ANR in the *Programme franco-allemand en Sciences humaines et sociales* (grant numbers KR 951/17-1; ANR-18-FRAL-0010-01). We are also grateful to Thomas Kisler, Florian Schiel, and Raphael Winkelmann (Bavarian Archive for Speech Signals) for their support with adapting WebMAUS to the particular workflow of DoReCo.

Language	Family	Macro-area	Glottocode	Donator/Reference
Anal	Sino-Tibetan	Eurasia	anal1239	(Ozerov, 2018)
Arapaho	Algic	N America	arap1274	(Cowell, 2019)
Asimjeeg Datooga	Nilotic	Africa	isim1234	(Griscom, 2018)
Beja	Afro-Asiatic	Africa	beja1238	(Vanhove, 2017)
Bora	Boran	S America	bora1263	(Seifart, 2009)
Daakaka	Austronesian	Papunesia	daka1243	(von Prince, 2013)
Daakie	Austronesian	Papunesia	port1286	(Krifka, 2016)
Fanbyak	Austronesian	Papunesia	orko1234	(Franjieh, 2018)
Goemai	Afro-Asiatic	Africa	goem1240	(Hellwig, 2003)
Gorwaa	Afro-Asiatic	Africa	goro1270	(Harvey, 2017)
Gubëëher	Atlantic-Congo	Africa	bain1259	(Bèye, 2012)
Gurindji	Pama-Nyungan	Australia	guri1247	(Meakins, 2016a)
Gurindji Kriol	(Mixed)	Australia	guri1249	(Meakins, 2016b)
Jahai	Austroasiatic	Eurasia	jeha1242	(Burenhult, 2016)
Kagate (Syuba)	Sino-Tibetan	Eurasia	kaga1252	(Gawne, 2019)
Kakabe	Atlantic-Congo	Africa	kaka1265	(Vydrina, 2013)
Kamas	Uralic	Eurasia	kama1378	(Gusev and Klooster, 2018)
Katla	Atlantic-Congo	Africa	katl1237	(Hellwig, 2007)
Komnzo	Yam	Papunesia	wara1294	(Döhler, 2019)
Lower Sorbian	Indo-European	Eurasia	lowe1385	(Bartels et al., 2016)
Mavea	Austronesian	Papunesia	mafe1237	(Guérin, 2006)
Mojeño Trinitario	Arawakan	S America	trin1274	(Rose, 2018)
Movima	(isolate)	S America	movi1243	(Haude and Beuse, 2016)
Mwotlap	Austronesian	Papunesia	motl1237	(François, 2017)
Nafsan	Austronesian	Papunesia	sout2856	(Thieberger and Brickell, 2019)
Northern Alta	Austronesian	Papunesia	nort2875	(Garcia-Laguia, 2017)
Pnar	Austroasiatic	Eurasia	pnar1238	(Kruspe, 2019)
Resígaro	Arawakan	S America	resi1247	(Seifart, 2019)
Ruuli	Atlantic-Congo	Africa	ruul1235	(Witzlack-Makarevich et al., 2019)
Sadu	Sino-Tibetan	Eurasia	sadu1234	(Xu et al., 2012)
Savosavo	Austronesian	Papunesia	savo1255	(Wegener, 2016)
Tabaq (Karko)	Nubian	Africa	kark1256	(Hellwig, 2007)
Teop	Austronesian	Papunesia	teop1238	(Mosel and Schnell, 2015)
Totoli	Austronesian	Papunesia	toto1304	(Leto et al., 2010)
Urum	Turkic	Eurasia	urum1249	(Skopeteas, 2018)
Vera'a	Austronesian	Papunesia	vera1241	(Schnell, 2015)
Yali	Trans-New-Guinea	Papunesia	angg1239	(Riesberg et al., 2016)
Yanomama	Yanomamic	S America	yano1262	(Ferreira, 2020)
Yongning Na	Sino-Tibetan	Eurasia	yong1270	(Michaud, 2017)
Yurakaré	(isolate)	S America	yura1255	(Van Gijn et al., 2012)

Table 5: DoReCo language sample as of February 2020. Classification based on Hammarström et al. (2019).

8. Bibliographical References

Anand, P., Chung, S., and Wagers, M. (2015). Widening the Net: Challenges for Gathering Linguistic Data in the Digital Age. Paper submitted to the National Science Foundation as part of its SBE 2020 planning activity.

Assaneo, M. F. and Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science Advances*, 4(2):eaao3842, February.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Belz, M. (2017). Glottal filled pauses in German. In Proceedings of DiSS 2017, Royal Institute of Technology, Stokholm, Sweden, pages 5–8.

Boersma, P. and Weenink, D. (2019). Praat: doing phonetics by computer. Computer program. Version 6.1.05, retrieved 16 October 2019 from http://www.praat.org/.

Brugos, A., Langston, A., Shattuck-Hufnagel, S., and Veilleux, N. (2019). A cue-based approach to prosodic disfluency annotation. In Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia, pages 3413–3417.

Buseman, A. and Buseman, K. (2019). Field Linguists Toolbox.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology*, 5(7):e1000436.

Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9).

Eklund, R. (2004). Disfluency in Swedish human-human and human-machine travel booking dialogues. Ph.D. thesis, Linköpings Universitet.

ELAN developers. (2019). ELAN (Version 5.7) [Computer software]. (June 14, 2019).

Fletcher, J. (2010). The prosody of speech: timing and rhythm. In William J. Hardcastle, John Laver and Fiona Gibbon (Eds.), *The Handbook of Phonetic Sciences*. Malden, MA:Wiley Online Library, chapter 15pp. 521–602.

Frank, A. and Jaeger, F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In Proceedings of the Annual Meeting of the Cognitive Science Society.

Ghitza, O. (2014). Behavioral evidence for the role of cortical oscillations in determining auditory channel capacity for speech. *Frontiers in Psychology*, 5(652), July.

Gonzalez, S., Grama, J., and Travis, C. E. (Ms). Comparing the performance of major forced aligners used in sociophonetic research. Manuscript, August 2019.

Harald Hammarström, et al., editors. (2019). Glottolog 4.0. Max Planck Institute for the Science of Human History, Jena.

Haspelmath, M. (2004). Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. `https://ling.auf.net/lingbuzz/004531`.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3):61–83.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36(1):161–195.

Jaeger, T. F. and Buz, E. (2017). Signal Reduction and Linguistic Encoding. In Eva M. Fernández and Helen Smith Cairns (Eds.), *The Handbook of Psycholinguistics*. Hoboken, NJ:John Wiley & Sons, pp. 38–81.

Kipp, A., Wesenick, B., and Schiel, F. (1997). Pronunciation modeling applied to automatic segmentation of spontaneous speech. In Proc. Eurospeech, pages 1023–1026, Rhodes.

Kisler, T., Schiel, F., and Sloetjes, H. (2012). Signal processing via web services: the use case WebMAUS. In Proceedings Digital Humanities, pages 30–34, Hamburg.

Klatt, D. H. (1979). Synthesis by rule of segmental durations in English sentences. *Frontiers of Speech Comm. Res.*, pages 287–299.

Liberman, M. Y. (2019). Corpus Phonetics. *Annual Review of Linguistics*, 5(1):91–107.

Lindblom, B. (1986). Phonetic universals in vowel systems. *Experimental phonology*, 1344.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and brain sciences*, 21(4):499–511.

Maddieson, I. (2006). In search of universals. In Ricardo Mairal and Juana Gil (Eds.), *Linguistic universals*. Cambridge:Cambridge University Press, pp. 80–100.

Menzerath, P. (1928). Über einige phonetische Probleme. In Actes du premier congrès international de linguistes.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., and Adamou, E. (2014). Documenting and researching endangered languages. *Language Documentation & Conservation*, 8:119–135.

Norcliffe, E., Harris, A. C., and Jaeger, T. F. (2015). Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances. *Language, Cognition and Neuroscience*, 30(9):1009–1032, October.

Ohala, J. J. (1983). The origin of sound patterns in vocal tract constraints. In Peter F. MacNeilage (Ed.), *The production of speech*. New York:Springer, pp. 189–216.

Schiel, F. (2004). MAUS goes iterative. In Proc. LREC, pages 1015–1018, Lisbon, Portugal.

Schmidt, T. (2011). A TEI-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative*, 1.

Shriberg, E. E. (1993). Preliminaries to a theory of speech disfluencies. Ph.D. thesis, University of California at Berkeley.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In 6th international Conference on Language Resources and Evaluation (LREC 2008).

Stave, M., Paschen, L., Pellegrino, F., and Seifart, F. (2020).

Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzeraths laws. To appear in: Linguistic Vanguard.

The Language Archive. (2017). ELAN Annotation Format, EAF, Schema version: 3.0. Max Planck Institute for Psycholinguistics, Nijmegen.

Zipf, G. K. (1935). The psycho-biology of language. The M.I.T. Press, Cambridge, MA. Reprinted 1968.

Zipf, G. K. (1949). Human behavior and the principle of least effort. Addison-Wesley Press, Cambridge, MA.

Zipser, F. and Romary, L. (2010). A model oriented approach to the mapping of annotation formats using standards. In Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010, Malta.

9. Language Resource References

Bartels, Hauke and Thorquindt-Stumpf, Kamil and Meschkank, Jan and Neumann, Colett and Szczepańsk, Marcin. (2016). Text and audio corpus of native Lower Sorbian. Nijmegen: TLA.

Niclas Burenhult. (2016). Jahai. Nijmegen: TLA.

Bèye, Amadou. (2012). The language of material culture in Baïnounk Gubëëher. London, SOAS: Endangered Languages Archive.

Cowell, James Andrew. (2019). Arapaho. University of Colorado.

Döhler, Christian. (2019). Komnzo text corpus. Zenodo.

Ferreira, Helder Perri. (2020). Documentation and description of the Yanomama of Papiu, an endangered Yanomami language of Brazil. London, SOAS: Endangered Languages Archive.

François, Alexandre. (2017). Corpus mwotlap. Pangloss collection, LACITO-CNRS.

Franjieh, Michael. (2018). The languages of northern Ambrym, Vanuatu: an archive of linguistic and cultural material from the North Ambrym and Fanbyak languages. London, SOAS: Endangered Languages Archive.

Garcia-Laguia, Alexandro. (2017). Documentation of Northern Alta, a Philippine Negrito language. London, SOAS: Endangered Languages Archive.

Gawne, Lauren. (2019). Kagate (Syuba), an endangered Tibeto-Burman language of Nepal. London, SOAS: Endangered Languages Archive.

Griscom, Richard. (2018). Documenting Isimjeeg Datooga. London, SOAS: Endangered Languages Archive.

Guérin, Valérie. (2006). Documentation of Mavea. London, SOAS: Endangered Languages Archive.

Gusev, Valentin and Klooster, Tiina. (2018). INEL Kamas Corpus. Hamburger Zentrum für Sprachkorpora.

Harvey, Andrew. (2017). Gorwaa: an archive of language and cultural material from the Gorwaa people of Babati (Manyara region, Tanzania). London, SOAS: Endangered Languages Archive.

Haude, Katharina and Beuse, Silke Angelika. (2016). Movima. Nijmegen: TLA.

Hellwig, Birgit. (2003). Goemai texts. London, SOAS: Endangered Languages Archive.

Hellwig, Birgit. (2007). A documentation of Tabaq, a Hill Nubina language of the Sudan, in its sociolinguistic context. London, SOAS: Endangered Languages Archive.

Krifka, Manfred. (2016). Daakie. Nijmegen: TLA.

Kruspe, Nicole. (2019). Pnar.

Leto, Claudia and Alamudi, Winarno S. and Himmelmann, Nikolaus P. and Kunht-Saptodewo and Riesberg, Sonja and Basri, Hasan. (2010). DoBeS Totoli Documentation. Nijmegen: TLA.

Meakins, Felicity. (2016a). Gurindji. Nijmegen: TLA.

Meakins, Felicity. (2016b). Gurindji Kriol. Nijmegen: TLA.

Michaud, Alexis. (2017). Na Corpus. Pangloss collection, LACITO-CNRS.

Mosel, Ulrike and Schnell, Stefan. (2015). Multi-CAST Vera'a. Haig, Geoffrey and Schnell, Stefan (eds.), Multi-CAST: Multilingual corpus of annotated spoken texts.

Ozerov, Pavel. (2018). A community-driven documentation of natural discourse in Anal, an endangered Tibeto-Burman language. London, SOAS: Endangered Languages Archive.

Riesberg, Sonja and Walianggen, Kristian and Zöllner, Siegfried. (2016). DoBeS Documentation Summits in the Central Mountains of Papua. Nijmegen: TLA.

Rose, Françoise. (2018). Corpus mojeño trinitario. Online database.

Schnell, Stefan. (2015). Multi-CAST Vera'a. Haig, Geoffrey and Schnell, Stefan (eds.), Multi-CAST: Multilingual corpus of annotated spoken texts.

Seifart, Frank. (2009). Bora documentation. Nijmegen: TLA.

Seifart, Frank. (2019). Resígaro. Nijmegen: TLA.

Skopeteas, Stavros. (2018). Urum. Universität Göttingen.

Thieberger, Nick and Brickell, Timothy. (2019). Multi-CAST Vera'a. Haig, Geoffrey and Schnell, Stefan (eds.), Multi-CAST: Multilingual corpus of annotated spoken texts.

Van Gijn, Rik and Hirtzel, Vincent and Gipper, Sonja. (2012). The Yurakaré archive. Nijmegen: TLA.

Vanhove, Martine. (2017). Corpus bedja, projet CORLI. LLACAN.

von Prince, Kilu. (2013). Daakaka. Nijmegen: TLA.

Vydrina, Alexandra. (2013). Description and documentation of the Kakabe language. London, SOAS: Endangered Languages Archive.

Wegener, Claudia. (2016). Savosavo. Nijmegen: TLA.

Witzlack-Makarevich, Alena and Namyalo, Saudah and Kiriggwajjo, Anatol and Molochieva, Zarina and Atuhairwe, Amos. (2019). A corpus of spoken Ruuli. Makerere University & Hebrew University of Jerusalem.

Xu, Xianming and Bai, Bibo and Yang, Yan. (2012). Linguistic and cultural documentation of Sadu. London, SOAS: Endangered Languages Archive.