Language Documentation & Conservation Special Publication 25 115
Doing corpus-based typology with spoken language data: State of the art
edited by Geoffrey Haig, Stefan Schnell, and Frank Seifart, pp. 115 – 139
https://nflrc.hawaii.edu/ldc/sp25
https://hdl.handle.net/10125/74659

Combining documentary linguistics and corpus phonetics to advance corpus-based typology

Frank Seifart

Leibniz-Centre General Linguistics (ZAS)

Abstract

This article argues that documentary linguistics and corpus phonetics can form a happy marriage in that corpora extracted from language documentation collections contain highly relevant data that can advance corpus phonetics by enabling broad comparative studies. To make this point, this article reviews previous research on phonetic lengthening at utterance boundaries and pause probabilities before nouns and verbs in ten languages. I then introduce the DoReCo initiative, which, based on experience gained from these studies, builds a database of time-aligned corpora from documentary collections of 50 languages for corpus phonetic research and other research purposes.

Keywords: documentary linguistics, corpus phonetics, typology, electronic corpora

1 Why study speech rate in fieldwork data?

This article argues that data on small and often endangered languages collected during language documentation fieldwork hold great potential for the cross-linguistic study of speech rate, specifically the slow versus fast pronunciation of words and the presence versus absence of pauses (Figure 1).[1] Such variation lies at the very heart of theories of human online language production, on the one hand, and theories of historical language change, on the other hand. Regarding the former, slow speech and pauses may be indicative of a high processing load, and thus provide us with a window into the cognitive-neural and physiological-articulatory bases of the human language production system (e.g. Jaeger & Buz 2017). For theories of language change, the emergence of phonetically reduced forms through fast pronunciation and their subsequent conventionalization plays a major role (e.g. Sóskuthy & Hay 2017). Also, the emergence of phonologically bound forms from function words may be facilitated by the absence of pauses between function words and lexical hosts (e.g. Himmelmann 2014).

Both theories of language production and theories of historical language change aim at species-wide applicability across any human population and any natural human language. But both are – so far – based on data from only a small fraction of attested human languages. This is at least in part due to difficulties in obtaining cross-linguistic data on speech rate. Measurements of articulation speed and pause occurrences have been less readily available for larger samples of languages than grammatical features such as structural information on basic word order, which is currently available for over 5000 languages (Hammarström 2015).[2] However, over the past couple of decades,

1 The research reported in Section 2 was supported by a grant from the Volkswagen Foundation's *Dokumentation Bedrohter Sprachen* (DoBeS) program (89 550), and the initiatve reported in Section 3 by a grant from the ANR and DFG programme *FRAL — Programme franco-allemand en Sciences humaines et sociales* (ANR-18-FRAL-0010/DFG-KR951/17-1). I am grateful for useful comments from anonymous reviewers and from Stefan Schnell.

2 It is worth noting here that the former is a feature of language use that has to be examined in annotated audio corpora whereas the latter is read off published descriptions, typically grammars. Availability of word order data from language use could be more informative for typological comparison but is much more restricted and limited to mostly written cor-

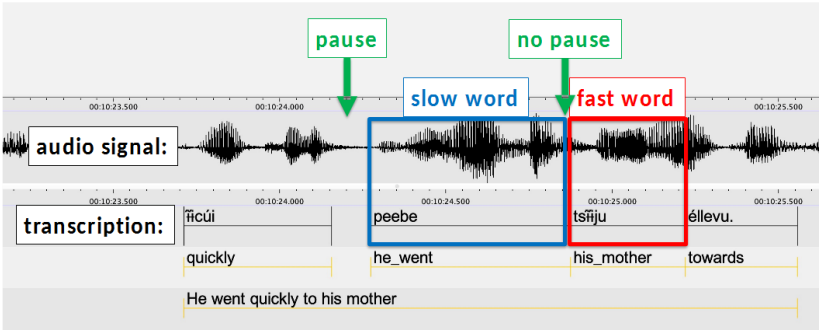

Figure 1 Utterance from a Bora (Northwest Amazon) myth, with transcription time-aligned with the audio, illustrating slow vs. fast pronunciation of words and the presence vs. absence of pauses. The four-segment word *peebe* is pronounced in approximately 600 milliseconds, the four-segment word *tsɨ́íju* in almost half that time, approximately 350 milliseconds (two orthographic vowel symbols, e.g. ⟨ee⟩, correspond to phonemically long vowels, e.g. /e:/, and ⟨ts⟩ to a single affricate phoneme). Source: llijchu_ine_I 099, `https://hdl.handle.net/1839/00-0000-0000-000C-DFBE-1`.

projects aiming at the documentation of small languages have produced audio materials with expert transcriptions, translations, and often further annotations, following documentary linguistics standards (Himmelmann 1998), for literally hundreds of languages from around the world (Seifart et al. 2018a). These materials are particularly well-suited to being processed for the cross-linguistic study of speech rate – and thus mitigate data scarcity in this area of study – for the following reasons:

pora and relatively small text samples. For instance, Futrell et al. (2020: 384) consider 54 languages with minimally 500 clauses represented in the collection of Universal Dependencies annotations (Nivre et al. 2016), Gerdes et al. (2021) consider 72 different language corpora from this collection, and Wälchli (2009: 81–85) studies word order in corpora of 100 languages from parallel Bible texts (gospel according to Mark) and selects those languages to match those represented in the core set of WALS languages (Dryer & Haspelmath 2013); see also Schnell et al. (this volume).

1. Speech rate phenomena can be studied in spoken text corpora that have been *collected for other purposes*, such as the traditional narratives and personal narratives ("original texts" in terms of Haig et al. 2011) that language documentation collections typically contain. Speech rate studies do not necessarily require targeted elicitation of complete verbal paradigms, negative evidence on ungrammatical word orders, etc.
2. Speech rate is *ubiquitous* and can therefore be studied in relatively *small corpora*, unlike certain grammatical structures that rarely occur, such as counterfactual constructions.
3. Speech rate can be *automatically measured relatively easily*, given phone-level time alignment, compared to more abstract structural properties, for instance morphological systems.
4. Relatively *shallow annotation* is sufficient to study speech rate – minimally a transcription that is close to a phonemic representation, although some studies might also require additionally at least a translation, and maybe morphological glossing and part of speech tagging, or annotation of prosodic units.
5. Annotations produced by language documentation projects are typically already *time-aligned* with audio at the level of utterances, or other multi-word units, often using the ELAN software (ELAN developers 2020), which greatly facilitates more fine-grained automatic time-alignment.

This article presents two steps towards advancing corpus-based typology by bridging the gap between documentary linguistics and corpus-phonetic approaches to studying speech rate. The following section (Section 2) addresses the potentials of this approach by presenting two corpus-phonetic studies using a sample of ten languages. Section 3 presents the ongoing DoReCo initiative to transform language documentation materials into corpora for studying speech rate for a total of 50 languages. Section 4 concludes this article. It should be noted that this article represents to some extent my own personal perspective and trajectory, starting out as a language documentation 'practitioner' in the early 2000s, and then moving on to comparatively analyzing language documentation data in the projects presented in this article.

2 Two corpus-phonetic studies based on ten languages

2.1 Data sets

The two corpus-phonetic studies presented in the following (Sections 2.2 and 2.3) used data sets from ten languages (Table 1). These data sets includes previously published corpora on English and Dutch, in addition to seven corpora that were collected during language documentation projects that aimed at comprehensive documentation of language use in the respective communities (Baure, Bora, Chintang, Even, Hoocąk, N‖ng, and Texistepec) and one corpus of data collected during a large-scale study on contact-induced language change (Sakha). Data on these ten languages were processed and analyzed in collaborative projects between 2012 and 2018. The purpose of reporting on studies resulting from these projects here is, firstly, to highlight the potential of such data for corpus-phonetic research addressing theoretical research questions. Secondly, the limitations of these studies were taken into account when designing the subsequent DoReCo project.

2.2 Final lengthening

One type of local speech rate variation we studied is final lengthening. Final lengthening refers to the phonetic lengthening of segments preceding prosodic boundaries, which are often also marked by pauses, as in Figure 2. It is often linked to hypothetically species-wide cognitive processes like motor planning constraints (Byrd & Saltzman 2003), which suggests it should be observable across all natural human languages. But there are also indications that the extent and degree of final lengthening may serve as a listener-oriented strategy to signal different levels of constituency (e.g. Turk & Shattuck-Hufnagel 2000). If this is the case, cross-linguistic variation could be expected, and maybe also cross-cultural variation (Ordin et al. 2017). On which prosodic positions exactly final lengthening is realized is also known to depend on language-specific stress, mora, and vowel quantity characteristics, among others (Cho 2003: 125). In Bantu languages, for example, lengthening affects the penultimate, rather than ultimate syllables (Hyman 2013).

	typology				corpus	
language (family)	word order	phones/ word	stress vs. tone	vowel length	num. of words	reference
Baure (Arawakan)	VSO	5.73	stress	no	17563	Danielsen et al. 2009
Bora (Boran)	SOV	7.13	tone	yes	29795	Seifart 2009
Chintang (Sino-Tibetan)	SOV	5.14	stress	no	37731	Bickel et al. 2011
Dutch (Indo-European)	SOV	3.85	stress	yes	39448	CGN-consortium 2003
English (Indo-European)	SVO	3.70	stress	(tense-ness)	56136	Godfrey et al. 1992
Even (Tungusic)	SOV	5.79	stress	yes	37394	Pakendorf et al. 2010
Hoocąk (Siouan)	SOV	6.64	stress	yes	23176	Hartmann 2013
N‖ng (!Ui-Taa)	SVO	3.45	tone	no	25850	Güldemann et al. 2011
Sakha (Turkic)	SOV	5.77	stress	yes	31139	Pakendorf 2007
Texistepec (Mixe-Zoquean)	SOV	5.14	stress	yes	21315	Wichmann 1996

Table 1 Languages used in two pre-DoReCo corpus-phonetic studies.

Cross-linguistic studies of final lengthening are therefore crucial to tease apart such language-specific or culture-specific patterns from those that actually reflect properties of the presumably species-wide language production system. Speech rate variation has also received relatively little attention in approaches to 'prosodic typology' which have focused more on pitch and tone and to some extent rhythmic type (Jun 2005, 2014).

In Seifart et al. (2021) we studied final lengthening in the ten-language data set given in Table 1. These data were time-aligned at the word level, not at the phone level, which would have been beyond the scope of that re-

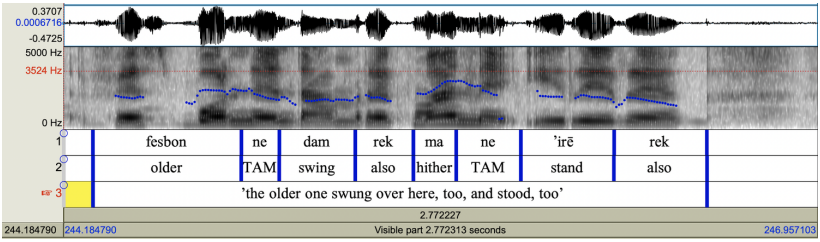

Figure 2 Vera'a example illustrating lengthening of *rek* 'also' in utterance-final vs. utterance-medial position (Schnell 2021: `veraa_anv_065`).

search project. We thus focused on lengthening of (orthographic) words[3] as a whole (following, e.g. Bell et al. 2003; Yuan et al. 2006), not syllables or segments. Specifically, we compared relative lengthening of utterance-final words (e.g. the final *rek* in Figure 2) and prefinal words (e.g. *'irē* in Figure 2) compared to the antepenultimate word and the word preceding that word (e.g. *ma* and *ne* in Figure 2), which are considered medial words (all utterance-initial words were excluded). Utterance boundaries were identified through the co-occurrence of automatically identified silent pauses and annotation-unit boundaries that were manually set by language experts according to semantic, syntactic, and prosodic criteria to segment texts mostly for practical purposes (for details, see Seifart et al. 2021).

Our statistical analyses controlled for a number of factors that are known to also influence whether a word is phonetically lengthened, including the relative frequency of a word (rare words may be pronounced more slowly; Bell et al. 2009), a word's length (longer words tend to be phonetically contracted by 'polysyllabic shortening'; Lehiste 1972), as well as for idiosyncratic effects of individual speakers or individual texts. In a nutshell, the results show that utterance-final words are indeed lengthened across this areally, genealogically, and structurally diverse set of languages. This supports the hypothesis

3 The word boundaries used are those set by language experts based on language-specific criteria. Since relative durations of words are compared within, not across languages, potential differences regarding how word boundaries are defined are not problematic.

that final lengthening – in one form or another – is a general feature of human language production. The study also revealed differences between languages, depending primarily on average word length (see Table 1): For example, in English (which has an average of 3.7 segments per word), final words were lengthened by almost 60%, whereas in Bora (which has an average of 7.13 segments per word), final words were lengthened by only 10%. Additionally, in languages with short words, prefinal words are also lengthened, but not in languages with longer words. Both of these differences support the hypothesis that across these languages, final lengthening affects final segments within final words most strongly, and segments preceding these gradually less, with detectable lengthening extending beyond four segments up to seven segments back from the utterance boundary.

This study thus advanced the understanding of prosody and speech production through corpus-based typology: It provides comparative evidence for the universality of final lengthening, applying the same methods across various languages. At the same time, it showed that further progress hinges on two developments: First, it requires timing information at the more fine-grained level of phones to better understand the backward propagation of lengthening within and across words. Deriving syllable durations from phone durations will shed further light on the interaction of language-specific prosodic and metric structure with final lengthening. Secondly, a sample larger than ten will be necessary to follow up on the hypothesis that patterns of utterance-final lengthening might be prone to areal spread as pronunciation styles that traverse language boundaries in multilingual settings. Eventually, the study of such durational patterns could then also be enriched by information on intonational cues, boundary tones, pitch resets, etc., which will require, however, considerable amounts of manual expert annotation.

2.3 Pause probabilities

Another study using this data set, which focused on pause probabilities, serves here to illustrate the potential of speech-rate studies for addressing patterns of historical language change, resulting in typological preferences (Seifart et al. 2017; Seifart et al. 2018b). This research is based on the hypothesis that hesitation pauses may inhibit grammaticalization of function words into affixes, and that asymmetries in pause probabilities thus underlie typological affix

asymmetries. This process has been evoked to explain the preference for suffixes over prefixes in the world's languages by a higher likelihood of pauses before content words than after content words by Himmelmann (2014), as illustrated by the following, fictitious example (1):

(1) a. *It happened **in (...) France***
natural pause ➔ few case prefixes

b. *It happened **years (...)[?] ago***
pause not natural ➔ many case suffixes

Our ongoing research investigates pauses before nouns versus before verbs as a potential motivation for the fact that verbal prefixes are more common than nominal prefixes in the world's languages (preliminary results were reported in Seifart et al. 2017). Pause probabilities across the ten languages in Table 1 show a clear pattern of more pausing before nouns than before verbs in the majority of these languages, with some languages showing no difference in pausing, but few if any languages showing a reverse effect (no language showed differences in pausing after nouns vs. verbs). The results reported in Seifart et al. (2018b: 4) on pauses before nouns versus verbs in nine languages (Figure 3) found the apparently exceptional pattern of more pauses before verbs than before nouns only in one language, namely English, although this pattern was not statistically significant in the slightly different analytical setup reported in Seifart et al. (2017). The overwhelming cross-linguistic tendency for more pauses before nouns than before verbs can be explained by a higher processing cost afforded by noun use compared to verb use (Seifart et al. 2018b): The use of full lexical nouns is usually only appropriate for new or in other ways special referents and otherwise pronouns or gaps are used, while such replacement does not usually occur for verbs.

This research illustrates the importance of a cross-linguistic approach to studying speech rate patterns: A pattern instantiated by seven out of ten languages (or six out of nine in Seifart et al. 2018b), with no (or only one) counter-example, allows for a typologically valid generalization, but it could easily be missed if only one language, for example English, was studied. On the other hand, a sample of ten languages is clearly not nearly large enough for proper areal or typological control. For instance, one would want to properly control for word order in such a study, comparing, for example, verb-initial versus

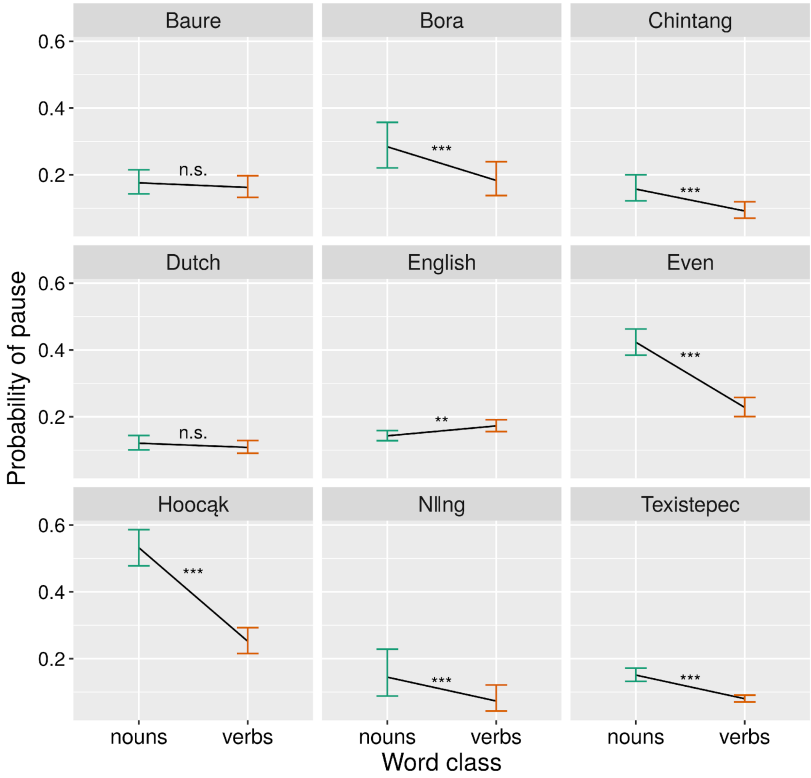

Figure 3 Pause probabilities before nouns vs. verbs in nine languages (reproduced from Seifart et al. 2018b).

verb-final languages. But even if a ten-language sample would be representative, it will naturally only contain one or two languages of the less common word order types (see Table 1), which is not enough to make generalizations.

3 Building the DoReCo database of 50 languages

3.1 The DoReCo project and corpus

What does it take to build a broader cross-linguistic database to further advance comparative corpus phonetics, or – more generally – corpus-based typology? This section discusses central aspects of DoReCo, an initiative to create a Language *Do*cumentation *Re*ference *Co*rpus of a diverse sample of 50 languages (Seifart et al. 2022). Since 2019, we have been building DoReCo from contributions of fieldwork corpora. Our research aims in DoReCo are to study cross-linguistic processes of phonetic contraction and lengthening, on the one hand, and the temporal distribution and rate of information-bearing units such as morphemes, on the other hand. For this purpose, we provide time-aligninment of annotations with audio at the phone level as a result of data processing within the DoReCo project. We expect that the carefully manually checked, consistent, and time-aligned phonemic transcriptions furnished by DoReCo will also be useful for subsequent research projects once DoReCo is finalized and published (expected in 2022). The following sections describe data selection (Section 3.2) and data processing and archiving (Section 3.3) in DoReCo.

3.2 Extracting linguistic corpora from eclectic collections

The first task in building DoReCo was to identify data sets that met a set of criteria in terms of audio and annotation quality and quantity. The great majority of DoReCo 'data sets' (or 'corpora') are subsets of larger language documentation 'collections'. These collections, as a whole, are extremely diverse in terms of the types of data they contain and how they are annotated, as well as in terms of how much data they contain. This diversity results, on the one hand, from the framework of documentary linguistics (Himmelmann 1998) that underlies the creation of these collections. According to this framework, language documentations should serve multiple purposes, that is the data should be accessible, interesting, and useful for many different potential uses and users. Potential uses may include various scientific disciplines (e.g.

anthropology, [ethno]botany, different fields of linguistics), but also the general public, and native speakers and their descendants. On the other hand, field workers carrying out language documentation also have research aims of their own which shape data collection and data processing choices and depth. This results in collections that might contain not only various text genres, but also elicited word lists, grammaticality judgments, photographs, recordings of music, etc.

For inclusion in DoReCo, we identified collections that contained data sets that, taken together, fulfilled the following set of criteria:

1. Within DoReCo, the *minimum corpus size* is 35000 phones, that is phonetic realizations of phonemes. This threshold was set in terms of phones instead of words (the standard measure for corpus size) to achieve comparable corpus sizes across synthetic versus analytic languages. On average, this threshold corresponds to about 10000 words.
2. The data set consists mostly of *narrative genres* (personal and/or traditional), among which a minority may have been collected from stimuli such as the *Pear* story (Chafe 1980), but a limited amount of conversational data are also included. Not included are elicitations of isolated sentences or words.
3. In terms of *annotation*, data need to be transcribed and translated into a major language. The transcription must be close to a phonemic representation. For the small languages DoReCo targets, this is usually the case, as they are generally transcribed using recently developed orthographies for primarily oral languages. For a subset of languages, DoReCo further requires morphological segmentation, glossing, and part-of-speech tags.
4. In terms of *audio quality*, there should be no, or very little, overlapping speech, which means texts selected for DoReCo are largely monological. Background noise should be absent or minimal, and the frequency range of the recording should be wide.
5. Transcriptions must be *time-aligned* at the level of multi-word units, for example utterance or prosodic phrase, typically through the ELAN annotation software (for examples, see Figure 1 and Figure 2). This criterion is included because time-alignment at the level of such

units makes the automatic phone-level time-alignment within them much more reliable.

6. The annotation files must, eventually, be *made publicly available* under a Creative Commons Attribution license (CC-BY), which allows re-users to distribute, remix, adapt, and build upon the material in any medium or format so long as attribution is given to the creator. Optionally, re-use can be restricted to non-commercial use (CC-BY-NC), to sharing potentially modified material under identical terms (CC-BY-SA), or to copying and distributing the material in unchanged form only (CC-BY-ND).[4] Corresponding audio files should ideally be made available under one of these licenses, too, but must at least be accessible for registered users in a recognized data repository.
7. There needs to be a *responsive corpus creator* for answering questions that arise during data processing.

Among the several hundred collections held in the major language documentation archives, like TLA, ELAR, AILLA, Paradisec, and Pangloss, only a fraction of data sets meet all seven criteria. Criterion two, consistent annotation, is the one that is most frequently not met for a minimum of 35000 phones (criterion one). In particular, there are not many collections that contain at least 35000 phones (or about 10000 words) that are not only transcribed and translated but also morphologically annotated.[5]

In a number of cases, there are also hurdles to make language documentation data publicly available (criterion five). Often, fieldworkers have obtained consent from speakers to make data available under very specific conditions, for instance involving registration and acceptance of the code of conduct by the DoBeS program,[6] or other schemes that prohibit commercial uses. In

4 https://creativecommons.org/about/cclicenses/

5 A clear step towards producing and archiving more extensively annotated data sets seems to have occurred when the ELDP (Endangered Languages Documentation Program) funding body implemented a scheme in which the release of subsequent grant installments is contingent upon meeting previously agreed-upon archiving plans (Holton & Seyfeddinipur 2018).

6 https://dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf

some cases, corpus creators prefer to add restrictions, like NC or ND, to enforce that commercial or scientific re-use involves prior contact of potential users with corpus creators to obtain consent. For these reasons, it would not have been possible to include 50 languages in the DoReCo set if the requirement had been that all data must be made accessible under a CC-BY license without further restrictions, even though that is what DoReCo strives for. Languages currently being processed in DoReCo are listed in Table 2.

3.3 DoReCo data processing and archiving

The main feature of DoReCo data processing is a close-to-phonemic transcription of the data that is time-aligned with the audio signal at the level of phones for the purpose of conducting corpus-phonetic research. Three steps are involved in furnishing these time-aligned transcriptions from multipurpose language-documentation data (Figure 4; for details, see Paschen et al. 2020a). Firstly, orthographic representations are converted to phonemic ones (so-called g2p, i.e. grapheme-to-phoneme, mapping). Secondly, inconsistencies between transcription and audio must be manually resolved. In materials that also serve the purpose of preserving cultural heritage and that are meant for distribution among the speech communities, repetitions are typically omitted, speech errors are corrected, and full forms are written instead of reduced ones (e.g. *will not* instead of *won't*). For optimization for corpus phonetic studies, these inconsistencies are resolved by DoReCo project members in consultation with corpus contributors with the aim of providing transcriptions that closely match what is actually present in the audio signal. Finally, this transcription is automatically time-aligned with audio at the phone level using the MAUS alignment software (Kisler et al. 2012). This step involves manual corrections of word start- and endtimes, the most labor-intensive aspect of DoReCo data processing. The time-aligned phonemic transcriptions are then added as an additional layer to the annotations that had been provided by corpus contributors, including translation, and potentially morpheme segmentation and glosses (Figure 4).

To optimally enable further re-use, DoReCo provides links to Glottolog (Hammarström et al. 2021) for genealogical, areal, and typological information on each language (see Table 2), assuring also interoperability with other

	language	glottocode	family	area	contributor(s)
1	Anal	anal1239	Sino-Tibetan	EURAS	Pavel Ozerov
2	Arapaho	arap1274	Algic	N AMER	Andrew Cowell
3	Asimjeeg Datooga	tsim1256	Nilotic	AFRICA	Richard Griscom
4	Baïnounk Gubëeher	bain1259	Atlantic-Congo	AFRICA	Alexander Yao Cobbinah
5	Beja	beja1238	Afro-Asiatic	AFRICA	Martine Vanhove
6	Bora	bora1263	Boran	S AMER	Frank Seifart
7	Cabécar	cabe1245	Chibchan	N AMER	Juan Diego Quesada, Stavros Skopeteas, Carolina Pasamonik, Carolin Brokmann, Florian Fischer
8	Cashinahua	cash1254	Panoan	S AMER	Sabine Reiter
9	Daakie	port1286	Austronesian	PAPUN	Manfred Krifka
10	Dalabon	ngal1292	Gunwinyguan	AUSTR	Maïa Ponsonnet
11	Dolgan	dolg1241	Turkic	EURAS	Alexandre Arkhipov
12	English	sout3282	Indo-European	EURAS	Nils Norman Schiborr
13	Evenki	even1259	Tungusic	EURAS	Olga Kazakevich, Elena Klyachko
14	Fanbyak	orko1234	Austronesian	PAPUN	Mike Franjieh
15	French (Switzerland)	stan1290	Indo-European	EURAS	Mathieu Avanzi, Marie-José Béguelin, Gilles Corminboeuf, Federica Diémoz, Laure Anne Johnsen
16	Goemai	goem1240	Afro-Asiatic	AFRICA	Birgit Hellwig
17	Gorwaa	goro1270	Afro-Asiatic	AFRICA	Andrew Harvey
18	Hoocąk	hoch1243	Siouan	N AMER	Iren Hartmann
19	Jahai	jeha1242	Austroasiatic	EURAS	Niclas Burenhult
20	Jejuan	jeju1234	Koreanic	EURAS	Soung-U Kim
21	Kakabe	kaka1265	Mande	AFRICA	Alexandra Vydrina
22	Kamas	kama1378	Uralic	EURAS	Valentin Gusev, Tiina Klooster, Beáta Wagner-Nagy, Alexandre Arkhipov

	language	glottocode	family	area	contributor(s)
23	Komnzo	komn1238	Yam	PAPUN	Christian Döhler
24	Light Warlpiri	ligh1234	(mixed)	AUSTR	Carmel O'Shannessy
25	Lower Sorbian	lowe1385	Indo-European	EURAS	Hauke Bartels, Marcin Szczepański, Kamil Thorquint-Stumpf, Serbski institut
26	Mojeño Trinitario	trin1278	Arawakan	S AMER	Françoise Rose
27	Movima	movi1243	(isolate)	S AMER	Katharina Haude
28	Nafsan (South Efate)	sout2856	Austronesian	PAPUN	Nick Thieberger
29	Nisvai	nisv1234	Austronesian	PAPUN	Jocelyn Aznar
30	Northern Alta	nort2875	Austronesian	PAPUN	Alexandro Garcia Laguia
31	Northern Kurdish (Kurmanji)	khor1267	Indo-European	EURAS	Geoffrey Haig, Maria Vollmer, Hanna Thiele
32	N‖ng	nngg1234	Tuu	AFRICA	Tom Güldemann, Martina Ernszt, Sven Siegmund, Alena Witzlack-Makarevich
33	Pnar	pnar1238	Austroasiatic	EURAS	Hiram Ring
34	Resígaro	resi1247	Arawakan	S AMER	Frank Seifart
35	Ruuli	ruul1235	Atlantic-Congo	AFRICA	Alena Witzlack-Makarevich, Saudah Namyalo, Anatol Kiriggwajjo, Zarina Molochieva, Amos Atuhairwe
36	Sadu	sadu1234	Sino-Tibetan	EURAS	Xianming Xu, Bibo Bai, Yan Yang
37	Sanzhi Dargwa	sanz1248	Nakh-Daghestanian	EURAS	Diana Forker, Nils Norman Schiborr
38	Savosavo	savo1255	Austronesian	PAPUN	Claudia Wegener

	language	glottocode	family	area	contributor(s)
39	Sümi	sumi1235	Sino-Tibetan	EURAS	Amos Teo, H Salome Kinny
40	Svan	svan1243	Kartvelian	EURAS	Jost Gippert
41	Tabaq (Karko)	kark1256	Nubian	AFRICA	Birgit Hellwig, Gertrud Schneider-Blum, Ismail Khaleel Bakheet Khaleel
42	Teop	teop1238	Austronesian	PAPUN	Ulrike Mosel
43	Texistepec Popoluca	texi1237	Zoque	N AMER	Søren Wichmann
44	Urum	urum1249	Turkic	EURAS	Stavros Skopeteas, Violeta Moisidi, Nutsa Tsetereli, Johanna Lorenz, Stefanie Schröter
45	Vera'a	vera1241	Austronesian	PAPUN	Stefan Schnell
46	Warlpiri	warl1254	Pama-Nyungan	AUSTR	Carmel O'Shannessy
47	Yali (Apahapsili)	apah1238	Nuclear Trans-New-Guinea	PAPUN	Sonja Riesberg
48	Yongning Na	yong1270	Sino-Tibetan	EURAS	Alexis Michaud
49	Yucatec Maya	yuca1254	Mayan	N AMER	Stavros Skopeteas
50	Yurakaré	yura1255	(isolate)	S AMER	Jeremías Ballivián Torrico, Sonja Gipper

Table 2 Languages for which data sets have been contributed to the DoReCo project and that are being processed at the time of writing. Note that the final set of data sets to be made available in 2022 may deviate from this list. Glottocode refers to language identification provided by Glottolog (Hammarström et al. 2021), from which family classification is also taken. Geographic macro-areas are assigned following WALS (Dryer & Haspelmath 2013): AFRICA – Africa; AUSTR – Australia; EURAS – Eurasia; N AMER – North America; PAPUN – Papunesia; S AMER – South America.

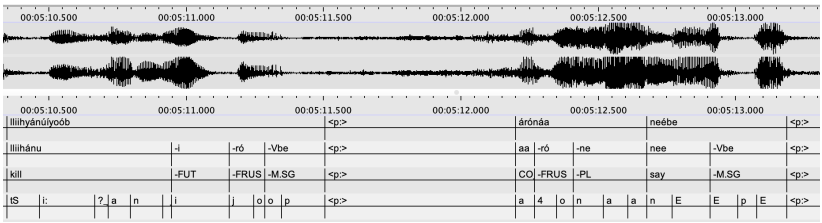

Figure 4 Screenshot of an EAF (ELAN) file from the Bora DoReCo data set (Seifart 2021: `meenujkatsi` 5:10) illustrating (i) orthographic transcription (first line) with hand-corrected word start- and endtimes and with representation of reduced forms (first word ending in *-b* vs. last word *-be*, same underlying suffix), (ii) roots and affixes with interlinear morpheme glosses, time-aligned based on (iii) automatically time-aligned phones (last line, SAMPA transcription, based on g2p mapping). Abbreviations used in glossing: FUT – future; FRUS – frustrative; M – masculine; SG – singular; CO – coordination; PL – plural. Additional abbreviations: `p:` – pause; `V` (as in *-Vbe*) – lengthening of preceding vowel. The example translates as '...he wanted to kill (him). But he said...'

resources in the CLLD ('cross-linguistic linked data') framework, like WALS.[7] Furthermore, DoReCo data sets include information such as the g2p mapping, and 'documentation' of tier names and glossing conventions for each data set (von Prince & Nordhoff 2020). Morphological glossing, which is also checked for consistency, is used within DoReCo for research on morphological complexity, also independently of time-alignment (Stave et al. 2021).

Once finalized, DoReCo will be sustainably archived within the French national research data infrastructure Huma-Num,[8] which provides handle-PIDs and guarantees longevity through the commitment of the French government. DoReCo data sets are conceived as citable resources, with each data set treated as a contribution, authored by the corpus creator, in an edited volume (e.g. Schnell 2021). Similar to peer-reviewed publications, DoReCo

7 `https://clld.org/datasets.html`

8 `https://huma-num.fr/`

data sets are guaranteed to fulfill a set of quality standards as a result of having undergone extensive quality checks and data processing by members of the DoReCo team. In this sense DoReCo is comparable to research databases like WALS (Dryer & Haspelmath 2013) or corpora like Switchboard (Godfrey & Holliman 1993) and Multi-CAST (Haig & Schnell 2021). And in this sense, DoReCo is different from archives or repositories, like TLA or ELAR, which accept a wide variety of data.

In summary, DoReCo is an initiative to advance corpus-based typology through processing existing data rather than by collecting new data, along with a number of related initiatives, each targeting different analytical goals (Schnell 2018): Among these is the Multi-CAST project and corpus, which annotated and analyzed language documentation data for grammatical relations, anaphoric reference, and animacy (Haig & Schnell 2021; Haig et al., this volume). Another is the three-participant project (Margetts et al., in press), which annotated and analyzed language documentation data for the expression of three-participant events. Both of these overlap with DoReCo in their language samples, which means that for some language documentation corpora, multiple layers of newly added annotations are available by now.

4 Conclusion and outlook

Bridging the gap between documentary linguistics and corpus phonetics involves both challenges and benefits on both ends. Regarding challenges, the framework of documentary linguistics still struggles with reconciling multipurpose data collection for the preservation of cultural heritage with producing and processing carefully curated data sets for specific scientific research questions. Corpus phonetics, on the other hand, as conceived by, for example, Liberman (2019), faces the challenge of developing methods that are applicable to comparative analyses of corpora that are relatively small compared to those on well-resourced languages, for instance regarding word frequency counts as control factors, or balanced sampling for, for instance speaker age and sex (for further discussion, see Strunk et al. 2020).

Regarding benefits, this approach provides a broad cross-linguistic basis for corpus-phonetic evidence for the human language production system, and for principles of language change. For instance, preliminary analyses

of 15 DoReCo languages show that final lengthening interacts more strongly than previously assumed with language-specific features such as phonemic vowel length distinctions (Paschen et al. 2020b) and thus also reflects learned, language-specific aspects of human language production, rather than purely responding to cognitive or articulatory constraints. On the other hand, the approach advocated in the current chapter addresses the fact that very little use in cross-linguistic studies has been made of the materials created by the huge, collective effort that went into creating the language documentation materials currently held at archives such as TLA and ELAR. The initiatives presented here are thus also an attempt to 'mobilize' data contained in these repositories, using speech rate as one example of an area of study that has been identified as "low hanging fruit" (Seifart 2012) for such mobilization efforts. These efforts also aim at enhancing visibility and representation of small and often endangered languages in the language sciences: Amazonian Resígaro, Australian Dalabon, South African N‖ng, and Siberian Dolgan are just as important as English and Dutch in offering insights into what human language is and can do.

References

Bell, Alan & Brenier, Jason M. & Gregory, Michelle & Girand, Cynthia & Jurafsky, Daniel. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1). 91–111. (https://doi.org/10.1016/j.jml.2008.06.003).

Bell, Alan & Jurafsky, Daniel & Fosler-Lussier, Eric & Girand, Cynthia & Gregory, Michelle & Gildea, Daniel. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America* 113(2). 1001–1024. (https://doi.org/10.1121/1.1534836).

Bickel, Balthasar & Stoll, Sabine & Gaenszle, Martin & Rai, Novel Kishore & Lieven, Elena & Banjade, Goma & Bhatta, Toya Nath & alii. 2011. *Audiovisual Chintang corpus (ca. 160000 words, plus paradigm sets and grammar sketches, ethnographic descriptions, photographs)*. Nijmegen: The Language Archive. (https://hdl.handle.net/1839/00-0000-0000-0005-6F41-C).

Byrd, Dani & Saltzman, Elliot. 2003. The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics* 31(2). 149–180. (https://doi.org/10.1016/S0095-4470(02)00085-2).

CGN-consortium, ELIS Gent, Language and Speech Nijmegen. 2003. *Corpus Gesproken Nederlands*. Nijmegen: Nederlandsle Taalunie.

Chafe, Wallace (ed.). 1980. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

Cho, Taehong. 2003. Prosodic boundary strengthening in the phonetics–prosody interface. *Language and Linguistics Compass* 10(3). 120–141. (https://doi.org/doi.org/10.1111/lnc3.12178).

Danielsen, Swintha & Riedel, Franziska & Admiraal, Femmy & Terhart, Lena. 2009. *Baure Documentation*. Nijmegen: The Language Archive. (https://hdl.handle.net/1839/00-0000-0000-000D-8382-B).

Dryer, Matthew S. & Haspelmath, Martin (eds.). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://wals.info/).

ELAN developers. 2020. *ELAN (Version 5.9)*. Nijmegen: Max Planck Institute for Psycholinguistics. (https://archive.mpi.nl/tla/elan).

Futrell, Richard & Levy, Roger P. & Gibson, Edward. 2020. Dependency locality as an explanation principle for word order. *Language* 76(2). 371–412.

Gerdes, Kim & Kahane, Sylvain & Chen, Xinying. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa* 6(1). 1–31. (https://doi.org/{10.5334/gjgl.764}).

Godfrey, John & Holliman, Edward. 1993. *Switchboard-1 Release 2 LDC97S62*. Philadelphia, PA: Linguistic Data Consortium.

Godfrey, John & Holliman, Edward & McDaniel, Jane. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In IEEE (ed.), *Proceedings of ICASSP-92*: *1992 IEEE International Conference on Acoustics, Speech and Signal Processing, 23–26 March 1992, The San Francisco Marriot, San Francisco, California*, 517–520. New York: Institute of Electrical and Electronics Engineers. (https://doi.org/10.1109/ICASSP.1992.225858).

Güldemann, Tom & Ernszt, Martina & Siegmund, Sven & Witzlack-Makarevich, Alena. 2011. *A Text documentation of N|uu*. London: ELAR. (https://hdl.handle.net/2196/00-0000-0000-0002-F81F-F).

Haig, Geoffrey & Schnell, Stefan (eds.). 2021. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. Version 2108. (https://multicast.aspra.uni-bamberg.de/).

Haig, Geoffrey & Schnell, Stefan & Schiborr, Nils N. This volume. Universals of reference in discourse and grammar: Evidence from the Multi-CAST collection of

spoken corpora. In Haig, Geoffrey & Schnell, Stefan & Seifart, Frank (eds.), *Doing corpus-based typology with spoken language corpora*: *State of the art* (*Language Documentation & Conservation* special publication 25), 141–177. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/74660).

Haig, Geoffrey & Schnell, Stefan & Wegener, Claudia. 2011. Comparing corpora from endangered languages: Explorations in language typology based on original texts. In Haig, Geoffrey & Nau, Nicole & Schnell, Stefan & Wegener, Claudia (eds.), *Documenting endangered languages*: *Achievements and perspectives*, 55–86. Berlin: Mouton de Gruyter. (https://doi.org/10.1515/9783110260021.55).

Hammarström, Harald. 2015. *The basic word order typology: An exhaustive study*. Paper presented at the Closing Conference of the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 1–3 May 2015.

Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2021. *Glottolog 4.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://doi.org/10.5281/zenodo.4761960).

Hartmann, Iren. 2013. *Hoocąk Corpus*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(2). 161–195.

Himmelmann, Nikolaus P. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language* 90(4). 927–960.

Holton, Gary & Seyfeddinipur, Mandana. 2018. Reflections on funding to support documentary linguistics. In McDonnell, Bradley & Berez-Kroeker, Andrea L. & Holton, Gary (eds.), *Reflections on Language Documentation 20 Years After Himmelmann 1998*, 100–109. Honolulu: University of Hawai'i Press. (https://hdl.handle.net/10125/24812).

Hyman, Larry M. 2013. Penultimate lengthening in Bantu. In Bickel, Balthasar & Grenoble, Lenore A. & Peterson, David A. & Timberlake, Alan (eds.), *Language Typology and Historical Contingency*: *In honor of Johanna Nichols*, 309–330. Amsterdam: John Benjamins.

Jaeger, T. Florian & Buz, Esteban. 2017. Signal Reduction and Linguistic Encoding. In Fernández, Eva M. & Smith Cairns, Helen (eds.), *The Handbook of Psycholinguistics*, 38–81. Hoboken, NJ: Wiley. (https://doi.org/10.1002/9781118829516.ch3).

Jun, Sun-Ah (ed.). 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, Sun-Ah (ed.). 2014. *Prosodic Typology II: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Kisler, Thomas & Schiel, Florian & Sloetjes, Han. 2012. *Signal processing via web services: The use case WebMAUS*. Paper presented at the Digital Humanities Conference 2012, Hamburg, Germany, 16–22 July 2012.

Lehiste, Ilse. 1972. The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America* 51(6B). 2018–2024. (https://doi.org/10.1121/1.1913062).

Liberman, Mark. 2019. Corpus phonetics. *Annual Review of Linguistics* 5(1). 91–107. (https://doi.org/10.1146/annurev-linguistics-011516-033830).

Margetts, Anna & Hellwig, Birgit & Riesberg, Sonja (eds.). In press. *The expression of caused accompanied motion events*. Amsterdam: John Benjamins.

Nivre, Joakim & de Marneffe, Marie-Catherine & Ginter, Filip & Goldberg, Yoav & Hajič, Jan & Manning, Christopher D. & McDonald, Ryan & alii. 2016. Universal Dependencies 1.0: A multilingual treebank collection. *Proceedings of the 16th International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016*. 1659–1666.

Ordin, Mikhail & Polyanskaya, Leona & Laka, Itziar & Nespor, Marina. 2017. Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Memory & Cognition* 45(5). 863–876. (https://doi.org/10.3758/s13421-017-0700-9).

Pakendorf, Brigitte (ed.). 2007. *Documentation of Sakha (Yakut)*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Pakendorf, Brigitte & Matić, Dejan & Aralova, Natalia & Lavrillier, Alexandra. 2010. *Documentation of the dialectal and cultural diversity among Ėvens in Siberia*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Paschen, Ludger & Delafontaine, François & Draxler, Christoph & Fuchs, Susanne & Stave, Matthew & Seifart, Frank. 2020a. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC'20), Marseille, France, 13–16 May 2020*. 2657–2666.

Paschen, Ludger & Fuchs, Susanne & Seifart, Frank. 2020b. *Phonological vowel length interacts with final lengthening*. Paper presented at the 12th International Seminar on Speech Production (ISSP2020), Providence, United States of America, 14–18 December 2020. (https://issp2020.yale.edu/S06/paschen_06_15_205_poster.pdf).

Schnell, Stefan. 2018. Reflections on the role of language documentations in linguistic research. In McDonnell, Bradley & Berez-Kroeker, Andrea L. & Holton, Gary (eds.), *Reflections on Language Documentation 20 Years After Himmelmann 1998*, 173–182. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/24818).

Schnell, Stefan. 2021. Vera'a DoReCo data set. In Seifart, Frank & Paschen, Ludger & Stave, Matthew (eds.), *Language Documentation Reference Corpus (DoReCo) 0.1*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft and Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Schnell, Stefan & Haig, Geoffrey & Seifart, Frank. This volume. The role of language documentation in corpus-based typology. In Haig, Geoffrey & Schnell, Stefan & Seifart, Frank (eds.), *Doing corpus-based typology with spoken language corpora*: *State of the art* (*Language Documentation & Conservation* special publication 25), 1–28. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/74656).

Seifart, Frank. 2009. Bora documentation. In Seifart, Frank & Fagua, Doris & Gasché, Jürg & Alvaro Echeverri, Juan (eds.), *A multimedia documentation of the languages of the People of the Center*: *Online publication of transcribed and translated Bora, Ocaina, Nonuya, Resígaro, and Witoto audio and video recordings with linguistic and ethnographic annotations and descriptions*. Nijmegen: The Language Archive. (https://hdl.handle.net/1839/00-0000-0000-0008-38E5-2).

Seifart, Frank. 2012. The threefold potential of language documentation. In Seifart, Frank & Haig, Geoffrey & Himmelmann, Nikolaus P. & Jung, Dagmar & Margetts, Anne & Trilsbeek, Paul (eds.), *Potentials of language documentation*: *Methods, analyses, and utilization*, 64–72. Honolulu, HI: University of Hawai'i Press.

Seifart, Frank. 2021. Bora DoReCo data set. In Seifart, Frank & Paschen, Ludger & Stave, Matthew (eds.), *Language Documentation Reference Corpus (DoReCo) 0.1*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft and Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Seifart, Frank & Evans, Nicholas & Hammarström, Harald & Levinson, Stephen C. 2018a. Language documentation 25 years on. *Language* 94(4). e324–e345. (https://doi.org/10.1353/lan.2018.0070).

Seifart, Frank & Paschen, Ludger & Stave, Matthew (eds.). 2022. *Language Documentation Reference Corpus (DoReCo) 1.0*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft and Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). (https://doreco.huma-num.fr/).

Seifart, Frank & Strunk, Jan & Bickel, Balthasar. 2017. *Recurrent patterns in the distribution of speech pauses cause languages to develop more prefixes in verbs than in nouns: An exhaustive study*. Paper presented at the 12th Meeting of the Association for Linguistic Typology, Canberra, Australia, 12–14 August 2017.

Seifart, Frank & Strunk, Jan & Danielsen, Swintha & Hartmann, Iren & Pakendorf, Brigitte & Wichmann, Søren & Witzlack-Makarevich, Alena & de Jong, Nivja H. & Bickel, Balthasar. 2018b. Nouns slow down speech across structurally and culturally diverse languages. *Proceedings of the National Academy of Sciences of the*

United States of America 115(22). 5720–5725. (https://doi.org/10.1073/pnas.1800708115).

Seifart, Frank & Strunk, Jan & Danielsen, Swintha & Hartmann, Iren & Pakendorf, Brigitte & Wichmann, Søren & Witzlack-Makarevich, Alena & Himmelmann, Nikolaus P. & Bickel, Balthasar. 2021. The extent and degree of utterance-final word lengthening in spontaneous speech from 10 languages. *Linguistics Vanguard* 7(1). (https://doi.org/10.1515/lingvan-2019-0063).

Sóskuthy, Márton & Hay, Jennifer. 2017. Changing word usage predicts changing word durations in New Zealand English. *Cognition* 166. 298–313. (https://doi.org/10.1016/j.cognition.2017.05.032).

Stave, Matthew & Paschen, Ludger & Pellegrino, François & Seifart, Frank. 2021. Optimization of morpheme length: A cross-linguistic assessment of Zipf's and Menzerath's laws. *Linguistics Vanguard* 7(s3). (https://doi.org/10.1515/lingvan-2019-0076).

Strunk, Jan & Seifart, Frank & Danielsen, Swintha & Hartmann, Iren & Pakendorf, Brigitte & Wichmann, Søren & Witzlack-Makarevich, Alena & Bickel, Balthasar. 2020. Determinants of phonetic word duration in ten language documentation corpora: Word frequency, complexity, position, and part of speech. *Language Documentation & Conservation* 14. 423–461. (https://hdl.handle.net/10125/24926).

Turk, Alice E. & Shattuck-Hufnagel, Stefanie. 2000. Word-boundary-related duration patterns in English. *Journal of Phonetics* 24(4). 397–440. (https://doi.org/doi.org/10.1006/jpho.2000.0123).

von Prince, Kilu & Nordhoff, Sebastian. 2020. An empirical evaluation of annotation practices in corpora from language documentation. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC'20), Marseille, France, 13–16 May 2020*. 2778–2787.

Wälchli, Bernard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13(1). 77–94. (https://doi.org/doi.org/10.1515/LITY.2009.004).

Wichmann, Søren. 1996. *Cuentos y colorados en popoluca de Texistepec*. Copenhagen: C.A. Reitzel.

Yuan, Jiahong & Liberman, Mark & Cieri, Christopher. 2006. Towards an integrated understanding of speaking rate in conversation. *Interspeech 2006*. 541–644.