RESEARCH REPORT

PROSODY AND CORPORA

Heliana MELLO
Universidade Federal de Minas Gerais (UFMG)

Amina METTOUCHI
Ecole Pratique des Hautes Etudes (EPHE)

Marianne MITHUN
University of California at Santa Barbara (UCSB)

Alessandro PANUNZI
Università degli Studi di Firenze (UNIFI)

Tommaso RASO
Universidade Federal de Minas Gerais (UFMG)

OPEN ACCESS

EDITORS
– Miguel Oliveira, Jr. (UFAL)
– René Almeida (UFS)

REVIEWERS
– Oliver Niebuhr (SDU)
– Philippe Mareüil (LIMSI–CNRS)

ABOUT THE AUTHORS
– Heliana Mello
Original Draft, Review & Editing.
– Amina Mettouchi
Original Draft & Review.
– Marianne Mithun
Original Draft & Review.
– Alessandro Panunzi
Original Draft & Review.
– Tommaso Raso
Original Draft & Review.

DATES
–Received: 05/19/2021
–Accepted: 07/10/2021
–Published: 08/01/2021

HOW TO CITE
MELLO, H.; METTOUCHI, A.; MITHUN, M.; PANUNZI, A.; RASO, T. (2021). Prosody and Corpora. *Cadernos de Linguística*, v. 2, n. 1, e385.

ABSTRACT

This paper focuses on the experience of spoken corpora compilation and discusses the relevance of prosody in this type of endeavor, as well as in the study of spoken language in its several possibilities. Through the voices of scholars associated with four different projects (CorpAfroAs, Mohawk Corpus, LABLITA, C-ORAL-BRASIL), the steps considered of utmost relevance in both the compilation and research potential of spoken corpora are presented; additionally, perspectives for the field in the future are pointed out.

RESUMO

Este artigo dedica-se à apresentação da experiência de compilação de corpora orais e discute a relevância da prosódia neste tipo de empreendimento, bem como para o estudo da fala em suas diversas possibilidades. Através da narrativa de pesquisadores associados a quatro diferentes projetos (CorpAfroAs, Mohawk Corpus, LABLITA, C-

ORAL-BRASIL), os passos considerados essenciais para a compilação e para pesquisas baseadas em corpora orais são apresentados; adicionalmente, perspectivas futuras para a área são apontadas.

KEYWORDS

Spoken Corpora; Prosody; Speech Segmentation; Information Structure.

PALAVRAS-CHAVE

Corpora Orais; Prosódia; Segmentação da Fala;
Estrutura Informacional.

INTRODUCTION

This paper presents the results of a roundtable organized during the *ABRALIN ao Vivo* series and partially maintains the structure of the original roundtable format. The purpose was to exchange ideas about spoken corpus compilation among the participants, focusing mainly on the role of prosody in such an enterprise. All the participants have lengthy career experience in dealing with the compilation of spontaneous speech corpora and prosodic studies. Two of the authors have compiled corpora of endangered or underdescribed languages, namely Mohawk (an indigenous language of northeastern North America) and different Afro-Asiatic languages of Africa. The other authors have experience with larger corpora of more widely spoken languages, namely Italian and Brazilian Portuguese (BP). All the authors share the basic assumption that prosodic phenomena are essential for the interpretation of speech in many ways: phrasing, syntax, information structure, illocution, and much more. Therefore, the corpora discussed include prosodic segmentation and the analysis of different aspects of linguistic structure through prosodic cues.

The point of departure for this panel was the common ground shared by all the participants about the necessity of analyzing real, unplanned speech data, taking into account prosody as a defining feature of speech. Spoken data can be approached in several different ways, depending on variables that range from the language being documented to the equipment available, funding, and the project team, among others. The panelists have compiled corpora for their diverse purposes. Corpus is a Latin word for 'body', so any body of texts (or other) in whatever form is technically a corpus (and that is how the term is most frequently used outside of linguistics). The only requirement is that it should constitute some sort of unit that caters to its purpose. In linguistics, a corpus is a language resource consisting of a usually large and structured set of texts (although there are small corpora for specific studies). Most corpora are presently stored and processed electronically. In corpus linguistics, corpora are defined even more narrowely and should follow a substantial set of prerequisites that qualify them for computerized searches, allowing statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules.

This paper is organized to introduce the reader to the experiences of the authors in their corpous compilation. It is not our goal here to provide step-by-step guidelines for spoken corpus compilation. Rather, we would like to show what it is possible to achieve in different language scenarios. Each new spoken corpus project should begin with the compilation of a set of initial questions to be revised, through trial and error, learning from mistakes, and continuous reassessment throughout the course of the work. It is our belief that it is better to study language through empirical data, even when the dataset available cannot be considered a corpus under the narrower definitions typically shared by corpus linguists. As

technology and methodology are evolving, corpus compilation is becoming not only necessary, but easier, in the sense that young scholars now have a half century of literature behind them that addresses issues pertinent to the field. Despite advances in technology, languages can lose their first-language speakers. Their documentation, even if not carried out with state of the art technology and methodologies, can create a crucial record of an essential part of the heritage of the community and of us all as human beings. Documentation should be done with whatever resources are available to researchers. For the discussion of spoken corpus compilation guidelines we suggest Himmelmann (2006), Mello (2014), Niebuhr and Michaud (2015), among many available references in the area.

This paper introduces the panelists' corpora in Table 1, which summarizes the main characteristics of the most important resources provided by each research group. Then in sections 2 through 5, authors go more deeply into the description of the resources produced by their teams, also focusing on different methodological aspects of spoken corpus compilation, with special emphasis on prosody, discussing the research possibilities offered by the resources, and concluding with some remarks about future perspectives for the field.

	CorpAfroAs	Mohawk Corpus	C-ORAL-BRASIL	LABLITA
Size	small	small	medium	large
Language(s)	lesser-described Afro-Asiatic languages	lesser-described North American Indigenous language	Brazilian Portuguese (BP)	Italian
Sound-aligned	yes	yes	yes	yes
Transcription	IPA, phonological words	community orthography	additional criteria added to regular orthography	additional criteria added to regular orthography
Granularity	pause type and length, overlaps, production phenomena	pauses, pause length, etc.	overlapping, filled pauses	overlapping, filled pauses
Software	Elan-CorpA	ELAN	WinPitch	WinPitch
Segmentation	prosodic	prosodic	prosodic	prosodic
Annotation	morphosyntactic words, morpheme-level glossing	free translation	PoS tagging, syntactic parsing, information structure for a sample of the corpora	PoS tagging, information structure for a sample of the corpus
Searchability	regular expressions, query engine online searchability	by character sequences in the Mohawk and the English translations.	metadata, PoS, prosodic boundaries and information tag queries, online searchability	metadata, PoS, prosodic boundaries and information tag queries, online searchability

Table 1. Resources summary.

The different resources and their specifications are presented in the following order: CorpAfroAS by A. Mettouchi (section 2), Mohawk by M. Mithun (section 3), LABLITA Lab by A. Panunzi (section 4), and finally C-ORAL-BRASIL projects by H. Mello and T. Raso (section 5).

1. THE RESOURCES PRODUCED UNDER THE COORDINATION OF A. METTOUCHI

1.1. THE CORPAFROAS CORPUS OF SPOKEN AFROASIATIC LANGUAGES: GENERAL PRESENTATION

When the CorpAfroAs project (<https://corpafroas.huma-num.fr/≥) was submitted to the French Agence Nationale de la Recherche (ANR) in 2006, very few spoken corpora of underdescribed languages were available. Among them, the great majority were archives, whose main purpose was conservation. The creators of the resource carried out their annotation (when there was one), without necessarily aiming at systematicity or consistency, or at fostering cross-linguistic research on their data.

In this context, CorpAfroAs brought an innovative perspective to the compilation of corpora in underdescribed languages, by building a methodology for the treatment of fieldwork textual data in underdescribed languages, from data gathering to automatic searches on the corpus. The constitution of the corpus was based on the linguistic analysis of the prosodic and morphosyntactic structure of the selected languages.

The outcome of the scientific work conducted on the project was the compilation of a pilot-corpus, which was made accessible online to the community of researchers in 2012. The term "corpus" implies that the aim was not simply to compile an archive, but starting from the theoretical analysis of spoken data gathered in the field, to create a body of systematically unified transcriptions, accompanied by morphosyntactic annotations, with sound indexed to text. One of the innovative aspects of the project was the choice to segment the spoken data into prosodic units, as opposed to syntactic or discursive ones.

The main deliverable of the project consists of transcribed, translated, and morphosyntactically annotated spoken narrative and conversational data in twelve languages, accompanied by metadata concerning all aspects of the recording sessions: human, contextual and technical. The languages in the project are: Kabyle, Tamashek (Berber), Beja, Gawwada, Ts'amakko (Cushitic), Wolaitta (Omotic), Hausa, Zaar (Chadic), Libyan and Moroccan Arabic, spoken Hebrew (Semitic), and Juba-Arabic (Arabic-based creole). Language sub-corpora are linked to a corresponding online grammatical sketch or list of glosses providing information on the grammar of the language, and the glossing labels and definitions used in the annotation of the sub-corpus.

Besides the elaboration of the corpus itself, a new version of the annotation software Elan, named Elan-CorpA, was developed within CorpAfroAs, together with methodological and scientific documents designed to help other researchers build comparable corpora in other languages: a methodological manual (https://corpafroas.huma-

num.fr/fichiers/manual.pdf), and a complete list of glosses adapted from the Leipzig Glossing Rules (http://bit.ly/Glosses-CorpAfroAs).

One of the long-term aims of CorpAfroAs is to generate similar initiatives for various language families, and to facilitate the development of quantitative corpora by language family or by language. Another aim is to contribute to corpus-based typological studies.

CorpAfroAs was followed by CorTypo (<https://cortypo.huma-num.fr/≥), another ANR-funded project for which Amina Mettouchi was also the project leader. Most of the corpora in CorTypo were segmented and annotated according to the CorpAfroAs template and principles, but CorTypo additionally features a typological database interfaced with the corpus.

1.2. CORPUS DESIGN OF CORPAFROAS

The languages of the corpus are not representative of the composition of branches and sub-branches of Afroasiatic - the only representativeness lies in the fact that there is at least one language per branch of the phylum. All subcorpora are spoken and represent unscripted speech.

The size of a written corpus is straightforwardly conveyed by the number of orthographic words it contains. The size of a spoken corpus is not that easy to describe, because results can vary substantially depending on whether the material is a monologue or a multilogue, on the speech rate, and on the way we define “words” and choose the unit of calculation. Table 2 shows different measurements for the Kabyle subcorpus of CorpAfroAs.

	DURATION (mn)	MORPHEMES	WORDS (morphosyntactic)	INTONATION UNITS	PAUSAL UNITS
FOLKTALE 01	13:29	6639	1803	614	392
FOLKTALE 02	12:16	6044	1748	546	372
RECOUNT 03	15:20	7302	2502	794	365
Total Monologues	41:05	19985	6053	1954	1129
CONVERSATION	8:06	3351	1384	680	-

Table 2. Size of the Kabile subcorpus of CorpAfroAs

Table 2 underlines the fact that the number of morphemes per word is an important piece of information for a morphosyntactically-annotated corpus. Let us compare Zaar (Chadic) and Kabyle (Berber) in that respect. For two folktales of a similar length (NARR01 in Zaar (12:45) and NARR02 in Kabyle (12:16)), Zaar has 2,052 words corresponding to 2,617 morphemes, and Kabyle has 1,748 words corresponding to 6,044 morphemes. This is exemplified in the two examples below:

(1) tá ɬə nə́ ʧa : -kə́nì dŭ:m //
FUT go for collect-nmlz honey //
'They have gone to gather honey' (SAY_BC_NARR01_SP1_013)

(2) i-ddm=dd t-a-tffaħ-t /
SBJ3.SG.M-grasp:PFV=PROX F-ABSL.SG-apple-F.SG /
'He took an apple' (KAB_AM_NARR02_029)

(FUT= future; NMLZ= nominalizer; SBJ= subject; SG= singular; M= masculine; PFV= perfective; PROX= proximal; F= feminine; ABSL=absolute)

Transcription is not orthographic, even if some of the languages of the corpus have more or less stable orthographic conventions. The choice was not only for homogeneity across the whole corpus, but mainly because orthography is designed for writing one's language, in the sense of producing, creating a written text. Transcription is supposed to be faithful to the recording, and in that respect, if we are to capture not only the gist of the meaning, but also the details of how that language is organized, then a transcription in IPA is desirable. Moreover, the mere decision to choose IPA triggers a number of exciting scientific questions, among them how to delimit words, and what kinds of words. This is addressed in more detail in Mettouchi (2013) and Izre'el & Mettouchi (2015).

For most CorpAfroAs subcorpora, there are two lines of text. One (tx) is in broad IPA and phonological words; it reflects the corresponding recording as faithfully as possible. The other (mot) is linked to the morphosyntactic analysis and annotation; it is in morphosyntactic words. The following example shows how different those two lines are (the colours underline morphophonological rules) and gives an idea of the type of annotation template chosen for CorpAfroAs.

(3) tx pʷintəd səβʕaθzəðmin gəsɣarən /
mot wwinntdd sbʕa tzdmin n jsɣarn /
mb wwin-nt=dd sbʕa t-zdm-in n j-sɣar-n /
ge bring\PFV-SBJ3.PL.F=PROX seven F-bundle\ANN-F.PL GEN ANN.M-firewood-M.PL/
rx V14-PRO=PTCL NUM N.OV PREP N.OV /
'They brought seven bundles of firewood' (KAB_AM_NARR_01_0790)

(PL= plural; GEN= genitive; SBJ= subject; PRO= pronoun; M= masculine; PFV= perfective; PROX= proximal; F= feminine; ANN=annexed; PTCL= particle; OV= overt, NUM= numeral; PREP= preposition)

Initially, the workflow consisted of entering the transcription and segmentation in Praat, then annotating in Toolbox, and then exporting into ELAN[1]. Later, a new version of Elan was developed within the project, which integrated a semi-automatic annotation module (CHANARD, 2015). However, the first stage in Praat remained the privileged entry point for the transcription, because the main focus of the work was prosodic segmentation, and this could be done more easily and accurately in Praat.

1 The procedure is described in the manual: https://corpafroas.huma-num.fr/fichiers/manual.pdf

The annotation template is the following:

ref	identifier for the annotation unit (time-associated)	
tx	transcription in broad phonetics into phonological words (SA)	
	mot	intermediary tier with segmentation into morphosyntactic words (SS)
		mb morphophonological transcription into morphemes (SS)
		ge morpheme-by-morpheme gloss of mb according to the Leipzig Glossing Rules, expanded within the project (SA)
		rx part-of-speech and other information relevant for retrieval purposes (SA)
ft	free translation into English (SA)	

SA: symbolic association. SS: symbolic subdivision

Figure 1. Annotation template of CorpAfroAs.

This type of annotation schema is necessary for conducting automatic searches reflecting formal encoding. For instance, in Kabyle, nouns are marked for number (singular, plural), gender (masculine, feminine) and state (absolute, annexed). Depending on the morphology and origin of the noun, the state can be marked overtly or covertly. Annotating all (overt and covert) absolute or annexed states in the "ge" line, and indicating by COV or OV in the "rx" line the overtness of the marking, allows to automatically search ("find ABSL in ge and COV in the corresponding rx cell") how many nouns in the corpus are, e.g. covertly in the absolute state vs how many are overtly so. This in turn can lead to interesting investigations on cognitive processing, overtness and markedness. Automatic searches can be conducted within Elan-CorpA and on the online corpus, using regular expressions.

1.3. PROSODIC SEGMENTATION IN CORPAFROAS

The purpose of CorpAfroAs was to provide a pilot corpus, carefully segmented and annotated, with accompanying software and methodological documentation. Concerning prosody, the main question was: what are the relevant units of speech for those languages, and more specifically:

- Are those units of a different nature depending on the prosodic systems (accentual / tonal) of the languages under investigation?
- How are prosody and morphosyntax articulated (especially in terms of information structure)?

The unit chosen for the segmentation of the corpus was the intonation unit (IU), defined as a stretch of speech forming a homogeneous and coherent intonation contour (CHAFE, 1994; DU BOIS *et al.*, 1992; 1993; TAO, 1996). Segmentation was implemented in collaboration with native speakers, for each language subcorpus.

For Kabyle, two native speakers were first trained by exposure to stretches of monologues containing typical IUs from particularly regular excerpts of folktales, and were made to understand that what was asked of them concerned the melodic and rhythmic contour of the unit, not its lexical, grammatical or pragmatic contents. This procedure relies on the assumption, supported by cultural practice, that folktales in Kabyle culture play a crucial role in oral education, not just to convey contents (cultural and social knowledge), but crucially, to teach forms (typical and varied grammatical and stylistic patterns).

Then the recording was played using Praat, and they were asked to tell where they would insert boundaries in the flow of speech; they indicated that by a beat of the hand on the table. For each beat, a boundary marker was inserted into the Praat textgrid corresponding to the sound file. The units thus delimited were additionally checked using Praat in order to look for measurable acoustic cues whenever there was disagreement between the two native speakers. A study of such IUs showed that four main perceptual and acoustic cues for boundary recognition were used: final lengthening, initial rush, pitch reset, pause. This is consistent with cross-linguistic findings (cf. CRUTTENDEN, 1997; DU BOIS *et al.*, 1992; HIRST; DI CRISTO, 1998). More details can be found in Izreel & Mettouchi (2015). The segmentation process was non-aprioristic concerning the function of those units.

Later, there was the addition of units for silent pauses over 200 ms (with exact duration in milliseconds coded by a number inside the cell), and for breath intakes (coded as BI, followed by the duration of the intake in milliseconds). The following were also added: careful transcription and annotation of production phenomena and dysfluencies (false starts, hesitations etc), and distinction between boundary tones: non-terminal (annotated /, rising tone, perception of non-finality, non-conclusiveness, expected continuity), terminal (annotated //, steep fall or rise, perception of finality, conclusiveness) and truncated or suspended boundary (annotated ##, (self-) interruption, or abandoned stretch of speech).

1.4. PROSODIC RESEARCH IN CORPAFROAS

Several studies of prosody were conducted on the corpus by members of the project. Some of them can be found in the book edited by Mettouchi *et al.* (2015). Caron *et al.* (2015) found that whereas thetic clauses and topics had similar intonational profiles across the four languages of the study (Zaar, Tamashek, Tripoli Arabic and Juba-Arabic), the expression of focus differed both in morphosyntactic and prosodic patterns.

Yatsiv-Malibert & Vanhove (2015) were able to propose the following typological generalizations, based on Beja, Zaar, Juba-Arabic and spoken Hebrew: (a) lack of a complementizer in a language correlates with prosodic integration of speech reports within the quotative frame, (b) non-clitic complementizers tend to be prosodically integrated within the quotative frame (and not within the reported discourse), (c) the prosodic boundary tends

to appear at the beginning of the speech report for OV languages, and at the end of the speech report for VO ones.

Prosodic segmentation was also crucial for studies on individual languages, such as Kabyle. The findings also depended on fine-grained transcription and annotation of the corpus.

In Mettouchi (2018), I found that the presence of a prosodic boundary was part of the definition of the direct object in Kabyle, thus integrating prosody and syntax as interacting coding means rather than as different layers of analysis. Following corpus analysis, direct objects in Kabyle can be defined as nouns in the absolute state (morphology), directly following the verb (syntax) in the same intonation unit (prosody), or possibly separated from it by a noun in the annexed state (= nominal subject), an adverb, a postverbal negator. Some apparent counterexamples involving a prosodic boundary between the verb and the object proved to actually be additional evidence, as the presence of dysfluencies or stylistic highlighting underlined the fact that those boundaries were not supposed to be there in the default situation.

Research conducted since 2011 on nouns pronominally indexed on verbs in Kabyle, and summarized in Mettouchi (2018a), has also underscored the importance of prosodic boundaries in the computation of grammatical relations (and information structure constructions).

In Kabyle, nominal subjects and objects can only be computed unambiguously within the prosodic group of the verb (an IU containing a verb): a noun is a nominal subject if and only if, within the prosodic group of the verb:

- the verb has no clitics other than the subject affix AND the noun occurs before the verb, and is in the absolute state; and
- the noun occurs after the verb (immediately or not) and is in the annexed state.

Several information structure constructions involve the presence of prosodic boundaries (noted [..]) in their very definition[2]:

(a) [V_{sbj} (N_{abs})] = (sub)topic continuation

(b) [V_{sbj} N_{ann} (N)] = introduction of a new episode in a narrative or a new subtopic in a conversation

2 V_{sbj} indicates a verb with a pronominal subject; N_{abs} is a noun in the absolute state, N_{ann} is a noun in the annexed state; parentheses indicate an optional element, square brackets mark the prosodic boundaries of the intonation unit.

(c) [N V_{sbj} (N)] = recapitulation of a preceding series of situations/events, serving as background for the following discourse

(d) N_{abs} [V_{sbj} (N) (N)] = 'contrastive comments', going against a presupposition about the topic which was built in the preceding context

(e) [V_{sbj} (N) (N)] N_{ann} = reactivation of a participant for topic promotion.

The interaction between prosody and syntax is also central in the cleft construction in Central-Western Kabyle (METTOUCHI, 2021): it is a construction whose function is to express narrow focus, and which is characterized by:

The juxtaposition of:

- a phrasal constituent (NP, ADV, QNT+N, PREP+N,...) preceded by a copula when its head is nominal,
- and a clausal constituent introduced by the relativizer i (*realis/de re*) or ara (*irrealis/de dicto*), where the relativizer bears the main prosodic prominence of the structure, and the relationship between the two parts of the cleft is marked by:
- an F0 peak on the relativizer,
- a lowering of F0 on the clefted constituent, proportional to the degree of prominence of the relativizer, that proportion expressing degree of contrastiveness,
- a single intonational contour for both parts of the cleft.

Another domain where prosodic segmentation is key is codeswitching. A paper based on a Kabyle/French codeswitching subcorpus (METTOUCHI, 2008) showed that there are fewer bilingual IUs than bilingual Complementizer Phrases, and that IUs tend to consistently start in the same language, with occasional switches, so that we get /L1... /L1... /L1...(switch)/L2... /L2.... Prosodic boundaries therefore consistently align with language choice.

Dysfluencies and production phenomena, which were crucial as evidence in the papers on clefts and direct objects, also brought insights for a better understanding of the role of pausing in relation to genre and oral performance.

In Mettouchi (2019), I showed, by comparing two types of monologues (folktales and a personal account) that while standalone audible breath intakes preceded by non-terminal boundary tones characterize the account, the folktales are marked by a high number of complex pauses, involving a silent pause preceding an audible breath intake, preceded by terminal boundary tones. As a conclusion to the study, I suggest that the

general function of the audibility of breath intakes, as opposed to silent or semi-silent breathing, is to regulate interaction by indicating to the interlocutors that the speaker is monitoring the discourse or narrative.

This qualitative study underlines the importance of segmenting various pause types, and annotating them, as well as annotating the terminal vs non-terminal nature of the boundary.

The interaction between segmentation and gesture is another domain that is worth investigating, if the video recording for the corpus is available. In a study of open-hand palm gestures conducted with G. Ferré (FERRÉ; METTOUCHI, 2020), on Kabyle folktales compared to English and French ones, the statistical analysis reveals a strong correspondence between Kabyle and the use of both palm-away and palm-down gestures, which the speakers align preferentially with the end of a major IU (terminal boundary tone), and with a pitch reset on the following IU.

1.5. PERSPECTIVES

In order to work on gestures, the current annotation of the CorpAfroAs Kabyle subcorpus had to be enriched with gestural annotations (palm-away, palm-up, palm-down, palm-on-side, palm toward self), and a number of other prosodic annotations (tonal contours, focal accents, pitch resets). This poses the question of the evolution of corpus annotation. Annotated corpora can evolve over time, not only because there are corrections to be made, but also in order to add new layers of annotation. It is therefore important that this should be made possible by the initial template, and the software into which the data are entered.

The reflection that this triggers, if phrased in terms of what would be sound advice to a young researcher starting a corpus, would be that it is good to have an evolving corpus based on a very simple starting point: a finely transcribed text aligned with the recording, and segmented into intonation units, thanks to Praat, and then imported into Elan and given an aligned translation.

This results minimally in a two-tier .eaf file (prosodically segmented transcription and translation), with a corresponding wav file, and the initial Praat Textgrid, can be the basis for several prosodic investigations.

Afterwards, depending on the type of investigation to be undertaken, it is always possible to enrich the initial "basic corpus" with several additional tiers and annotations.

Another piece of advice is methodological: consistency is key at all levels, in the transcription and segmentation process, as well as in the annotation. Because what you annotate is what you get, and generalizations, be they qualitative or statistical, need to be based on transparent, accurate and consistent annotations. This also implies that the

segmentation and annotation processes be documented and explicited in the presentation of the corpus, and the annotations clearly defined.

2. THE RESOURCES PRODUCED UNDER THE COORDINATION OF M. MITHUN

This section focuses on the Corpus of Spoken Mohawk: Kanien'kehá:ka. Mohawk is a language of the Iroquoian family, indigenous to the North American Northeast. There are six major communities, each with distinctive dialects, some more different than others. The language is still spoken well by skilled first-language speakers, though these are disappearing rapidly. There is, however, great interest in the language in all of the communities, and astonishingly impressive fluent second-language speakers are emerging.

2.1. THE CORPUS

The Mohawk corpus consists of audio and video recordings, with transcriptions and translations spanning nearly 50 years. Most have been done by a single researcher, working with first-language speakers on the transcription and translation. A few very valuable recordings were contributed by community members who had recorded family members at earlier times. The corpus consists of over 300 recordings varying in length from a few minutes to a few hours, totaling together just over 60 hours. 76 speakers are represented, of whom 30 are still alive.

It was important to record speech from all six communities. All of the dialects are descended from the speech of a community living in what is now eastern New York State until the 17th century. Some groups have been apart for the past 350 years, others less than 150 years. Some have been in contact since then, others not. The dialects differ in relatively minor phonological detail, but they vary significantly in idiomaticity. The language is highly polysynthetic, and the morphological structure is exactly the same across all dialects. But conventionalized lexical items and phrases differ noticeably.

A variety of genres is represented in the corpus. There is a lengthy, ornate traditional ceremonial speech, and some recordings from over 40 years ago of a very few young children just acquiring the language. Slightly more than half of the material consists of interactive conversation, involving anywhere from two to ten speakers. Language has always played an important role in Mohawk culture: Mohawks enjoy and appreciate their language; they admire and cultivate linguistic virtuosity in all areas from formal oratory through well-told stories to snappy repartee. Gatherings are typically full of laughter and

rapid-fire interaction, which can add to the challenges of transcription, but also to the value of the record.

The transcription is in the community orthography, which was devised over the past several decades at the request of community members. The system has a historical basis in writings of French-speaking missionaries who first arrived in the region in the seventeenth century, but it has been updated to represent all of the distinctions inherent in the language, including tone. While the same system is used for all of the dialects, each is transcribed as spoken. It can be written using a basic European keyboard, with grave and acute accents for tone, and apostrophe for glottal stop. Transcriptions and translations are entered into ELAN software, which produces transcripts searchable by sequences of characters in both the Mohawk and the English translations, as well as time stamps and the duration of prosodic units and pauses. A separate set of tiers is set up within ELAN for each speaker in a conversation, allowing display of overlaps.

2.2. THE PURPOSE IN BUILDING THE CORPUS

This corpus was constructed to create a record of the language as spoken by skilled, first-language speakers while this is still possible, both as a reference for future generations and as a basis for an extensive descriptive grammar. The language is much more than a list of distinctive sounds, morphological templates for verbs and nouns, and some schemata for basic syntactic constructions. It is what speakers choose to say and how they choose to say it. In the case of Mohawk, this can be quite different from French and English counterparts. Traditional ways of interacting, of packaging thoughts into concepts and linking them together, can be highly susceptible to language contact effects. Even when a language is still spoken fluently and grammatically by first-language speakers, these more subtle differences can erode without notice. This is of course not necessarily bad, but the Mohawk communities care deeply about a record of their traditions. Fortunately, speakers with traditional skills were able to contribute to the corpus. This has made it possible to build the descriptive grammar from spontaneous speech, usually interactive in context, each point illustrated with one or more examples from each community. Each example is accompanied by an identification of the speaker and community, a choice made by each speaker.

2.2.1. THE ROLE OF PROSODY

All transcription is based on prosodic segmentation, the segmentation of the speech signal into intonation units (prosodic phrases) and prosodic sentences. Prosody was foundational from the beginning. At the outset, the researcher first divided each recording into intonation units, then did a preliminary transcription. This was then brought to work with a speaker for fuller transcription and translation. Collaboration with a speaker was crucial in every case.

Speakers can of course identify what is said even in situations of reduced audio information, as when people are talking rapidly over one another. Excellent speakers have noted that the task is easier if they were recently participants in the conversation. And they can contribute invaluable information about what is being said: subtle details about the social implications of certain choices of expression, background about community events, the histories of relationships among participants and others mentioned, , as well as further examples of particular grammatical constructions and appropriate contexts of use. During these sessions, the intonation unit has proven to be the easiest to work with, the amount of information one can easily hold in the mind at a time.

The segmentation into intonation units has another important advantage: it reveals structuring not necessarily accessible from syntactic structure alone.

2.2.2. PRINCIPLES UNDERLYING SEGMENTATION

The features used to segment speech into intonation units fall into three major groups: pitch, timing, and phonation. For Mohawk the most salient is pitch. Intonation units are typically characterized by a coherent intonation contour, usually beginning with an initial pitch reset and some kind of final boundary intonation. These cues are often but not always paralleled by timing: potential pauses at boundaries, a possible initial rush (greater speed at the outset), and a possible final lag. There may also be cues from non-modal phonation, particularly final creaky voice. Intonation units may cluster into larger prosodic sentences, the whole beginning with an initial pitch reset, with smaller pitch resets on each successive intonation unit, and each of these ending with a non-terminal contour until the last, which ends with a terminal contour (usually a definitive fall). These are essentially the criteria proposed by Wallace Chafe in a series of works spanning the decades from the 1970's through the first two decades of the 21st century. Some summaries can be found in Chafe (1994) and (2000).

2.2.3. PROSODIC SEGMENTATION AS A THEORETICAL CHOICE

Segmentation into intonation units and prosodic sentences has not only served as a practical tool; it also reveals a kind of structure not discernible from morphological and syntactic structure alone. Because of the potentially elaborate morphological structure of the language, much of what is said in multi-word sentences can be conveyed with a single verb in Mohawk, typically with the addition of various discourse particles. Such a structure can be seen in (4) where the speaker noted that as children they knew not to bother their grandmother.

(4) *Iah* *kwi'* *tha'-t-a-iakhi-'nikonhnh-à:r-en-'*
not in.fact.TAG CONTR-DV-IRR-1PL>FI-mind-hang-BEN-PFV
not in fact you know would we mind-hang her
'We wouldn't bother her.'

(CONTR = contrastive, DV = duplicative, IRR = irrealis, 1PL = 1st person plural agent, FI = feminine indefinite patient, BEN = benefactive applicative, PFV = perfective aspect.)

Basic stress, tone, and vowel length are determined purely phonologically within the word: stress is basically penultimate (with additional principles involving epenthetic vowels), stressed open syllables are automatically lengthened, and distinctive tone is descended from laryngeals. Because so much information is contained within a word, constituent order is largely governed by information structure, and it is here that prosody plays a significant role.

2.2.4. PRACTICAL IMPLEMENTATION

Something not often mentioned in discussions of the implementation of prosodic segmentation, particularly when analysts are not working with their own languages, is that learning to hear prosodic structures can be much like learning to hear phonemically in a new language, as one becomes accustomed to paying special attention to cues that are significant for that language and perhaps even for that speaker, and disregarding others.

2.3. QUESTIONS AND FINDINGS

The Mohawk corpus has been the empirical foundation for essentially all my own work on the language, from basic word and sentence prosody, to lexicalization patterns, clause structures, complex sentence structures, discourse structures, and patterns of interaction. It combines the advantages of prosodic information and spontaneous speech in a larger linguistic and extra-linguistic context, without the intermediary of translation. It has been possible, among other things, to trace the relation between prosodic structure and the packaging of information. As seen in Mithun (2021), for example, speakers generally utter one new idea at a time, each in a separate intonation unit. Thus, the first time a significant referent is introduced, it is typically presented in an intonation unit of its own, and perhaps with a low content verb, but once an established part of the scene, it becomes part of longer units.

The integration of prosody into the record also makes it possible to distinguish constructions which might appear to be the same when viewed simply in written form on paper or a screen. Mohawk, like many languages, has distinct prosodic contours for topic shifts, various kinds of focus, and cleft constructions, though all may appear in writing as simply a nominal phrase followed by a verb.

2.4. TRANSCRIPTION, SEGMENTATION, AND WRITING: SOME CHOICES

Because there was not a well-established written tradition for Mohawk, it was possible for us to establish principles as we worked. We now generally punctuate written material according to the prosody, with commas corresponding to non-terminal pitch contours, and periods corresponding to final contours. These are not generally at odds with major syntactic units.

Various kinds of disfluencies, false starts, truncated intonation units (those without a terminal contour), hesitation particles, pauses, and repetitions of the type that might not appear in a formal written document in English or French are transcribed as spoken. These, too, are rarely random and can tell us much about the processing of speech.

One kind of choice that arises when establishing transcription principles is the segmentation of speech into words. There are three lexical categories in Mohawk, defined in terms of their internal morphological structure: verbs, nouns, and particles. Verbs and nouns each have clear, largely templatic internal structures. Word boundaries for each are completely clear, on both phonological grounds (basic penultimate stress) and morphological grounds. Particles are by definition monomorphemic, though they may be compounded, and herein lies potential uncertainty. Frequently-recurring sequences of particles can come, over time, to be processed as single chunks, in line with common processes of grammaticalization. One result of such processes can be seen in example (4) above, in the particle *kwi'*. This is the result of an amalgamation of a particle *ki'* 'in fact, actually', which indicates that this comment is pertinent to something in the preceding context, and the tag *wáhi'* "you know, isn't it, right', which might elicit some reaction from the listener. Such amalgamation is gradual. In many cases, transcription and writing force a decision. A certain sequence may sound like a single word in the speech stream, but some speakers may paraphrase it with the sequence of particles from which it evolved, while others may no longer be aware of the component parts.

2.5. LESSER DESCRIBED LANGUAGES

Mohawk could be characterized as a lesser-described language. Nearly all first-language speakers are now above 65 years of age. Every scrap of their speech is precious, as is the insight they can contribute about it. While the value of documentation of a wide variety of genres, occurring in the course of a vast range of activities, is indisputable, some opportunities are no longer available. The size of this Mohawk corpus is necessarily limited: nearly all material was recorded, transcribed and translated (in close collaboration with speakers), and entered into the database by a single researcher. Certain kinds of statistical studies are thus possible, but for others the sample is simply not comparable to those for larger languages. The social context is not one in which laboratory experiments are

generally appropriate. The priority is clearly on documentation of unscripted speech, on as many topics as possible, in as many genres as possible, and interactive to the extent possible, complemented by discussions about it with the speakers.

2.6. LOOKING TO THE FUTURE

Considerations for a linguist beginning a new corpus project will necessarily vary with the situation of the language to be documented. For a lesser-described language, particularly one with relatively few speakers, it can be crucial to consider what kinds of speech can be documented. It can be valuable to document speech during various activities, as speakers make things, play games, cook, eat, and much more. In the Mohawk Corpus compilation experience, speakers have been happy to be recorded, but invariably ask what they should talk about. It has been important to have a long list of topics to suggest. Often the speakers have ended up talking about entirely different things, but the suggestions were what got them started. In drawing up lists of topics, it can be useful to pay close attention to what the speakers tend to talk about in their everyday lives when they are not being recorded: neighbors, politics, gardening, pets, events, what they hope younger generations will know.

3. THE LABLITA ITALIAN CORPORA

3.1. THE CORPUS OF ITALIAN AND OTHER RESOURCES

The LABLITA corpus of Italian has been collecting data on spontaneous spoken Italian since 1965. The material has been transcribed using LABLITA-CHAT format (CRESTI; MONEGLIA, 1997), ensuring the annotation of terminal and non-terminal prosodic breaks as the basic segmentation level of spoken language. The format for the representation of speech is based on the model of the *Language into Act Theory* (L-AcT), as in Cresti (2000) and Moneglia & Raso (2014), that assumes a strict correspondence between the prosodic execution of speech and the expression of specific pragmatic values (namely, illocution and information structure).

Currently the LABLITA collection is comprised of about 700,000 words and 420 recording sessions, allowing the representation of spontaneous speech variation across Channels, Regulative aspects, Types of interaction, Social Contexts. Each transcript has been aligned *per utterance* with the corresponding audio source. The corpus is available online in the Orfèo platform, ensuring real time access to both the audio and the textual information (http://corpus.lablita.it/). Metadata are accessible as well. For each session, all the multimedia and annotated files can be downloaded, specifically:

- Transcription (rtf and txt format)
- Audio file (wav)
- PoS tagging (CONLL format)
- Alignment (WinPitch and Praat formats)
- Metadata (TEI and CHAT formats)

LABLITA has also coordinated the building of the C-ORAL-ROM corpus (CRESTI; MONEGLIA, 2005), a multilingual collection of four reference corpora for spoken Romance languages: Italian, French, Spanish, and Portuguese (<http://www.elda.org/en/proj/coralrom.html>). Each corpus has been collected using the same design, criteria and sampling techniques, thus ensuring comparability between the resources. Each sub-corpus includes formal and informal speech recorded in a large variety of contexts, with different dialogue/monologue structures, typologies/genres, and semantic domains.

The corpus comprises a total of 1,258,170 words (about 300,000 for each language) and 772 recording sessions. The entire multilingual corpus is available in a multimedia format, allowing simultaneous access to aligned acoustic and textual information. Transcriptions include the annotation of perceptual prosodic breaks, disfluencies, and overlappings. Each transcription is headed with metadata about speakers and recording situation.

DB-IPIC database (PANUNZI; GREGORI, 2012) is a freely accessible online XML database specifically designed for the study of information structure in spontaneous spoken language (<http://lablita.it/app/dbipic/>). DB-IPIC was designed to host the informal section of the Italian C-ORAL-ROM Corpus (74 recorded sessions; 124,735 transcribed words). Later, 3 mini-corpora of similar size (between 30,000 and 40,000 words) were added to the database in order to allow comparison between Italian, Brazilian Portuguese - derived from C-ORAL-BRASIL (PANUNZI; MITTMANN, 2014), and Spanish - derived from C-Or-DiAL (NICOLÁS; LOMBÁN, 2018).

All resources included in DB-IPIC have been manually annotated adopting the L-AcT framework, assuming that:

- major prosodic breaks signal utterance boundaries;
- the internal segmentation of the utterance in prosodic units reflects its information structure;
- each utterance is characterized by the expression of an illocutionary force, conveyed by the Comment information unit.

The online database permits direct access to the sound files and their download in mp3 format. Besides these main resources, the LABLITA collection includes a corpus of early language acquisition. Moreover, new multimedia collections of both adult and child language are in development, including video data and gesture analysis.

3.2. THE PURPOSE OF THE LABLITA CORPORA

The development of spoken corpora is a necessary step for the usage-based study of language *tout court*. The study of free conversation and of any linguistic production in its natural context allows us to observe the actual state of affairs regarding a language. This is true because spoken language is, in contrast to written language, the most basic and natural kind of linguistic interaction between humans, from both the ontogenetic perspective and the phylogenetic one. Using Levinson's words: "face-to-face interaction is not only the context for language acquisition but the only significant kind of language use in many of the world's communities, and indeed until relatively recently in all of them" (LEVINSON, 1983, p.44). From this perspective, the purpose of building a corpus of spoken language is a very general one, which is to capture the real dynamics of the verbal communication and language use.

At a more specific level, the objective of the corpora collected by the LABLITA group is to represent the interplay between prosody and pragmatics in a vast series of interactional contexts, in accordance with the L-AcT framework. The general ideas that lead and inspire the LABLITA collections are that spoken language is governed by pragmatic principles (Speech Act production and Information Patterning), and that prosody is the main means of expression of such principles. For these reasons, prosody is the starting point of the transcriptions, annotations and analyses.

The LABLITA corpora document a major language (Italian) with wide diatopic variability, mainly from North to South, and dozens of regional and local varieties. This fact played a crucial role in the definition of the criteria used for the LABLITA collection. The choice was to ensure a substantial variety of recording contexts and situations, focusing on diaphasic and diastratic variation, in order to maximize the likelihood of finding different types of interactions and Speech Acts. On the other hand, the geographical area of collection has been taken as a fixed standpoint. The result is a corpus that mainly represents the conversational Italian spoken in the Tuscan area, and more specifically Florence and its hinterland. Many people who live in this area are actually from other parts of Italy, so the corpus contains varieties from different parts of the country. The Tuscan variety, however, is by far the best documented one. It is worthwhile underlining the fact that the Tuscan variety played a special role in the development of standard Italian, being the source of the literary norm since the 13th and 14th centuries.

3.3. MAIN METHODOLOGICAL ISSUES

3.3.1. THE ROLE PROSODY PLAYED IN CORPUS DESIGN

As just mentioned, in the perspective on language we take, prosody definitely constitutes a core element of the linguistic system. Prosody is present in any language, with specific patterns, and it is one of the first linguistic elements to which babies are sensitive even from the first months of life (MEHLER *et al.,* 1988; MOON *et al,.* 1993). Moreover, prosody plays a central role both in segmentation and in the pragmatic interpretation of the utterances used in spontaneous communication, and notably in the encoding of Speech Acts. For these reasons, when we design, collect and process a corpus, prosody is central for at least two main reasons. The first is related to linguistic variation. Ensuring contextual variation allows us to represent different language uses, in which different Speech Act series and interplays occur. This also allows us to document how prosody spontaneously varies in different communication events. When we have collected a vast repertoire of prosodic profiles, we can identify and verify *a posteriori* their natural forms and functional correlations, in a corpus-driven perspective. Secondly, prosody is crucial in order to give the basic form of representation of spoken language in transcripts, which are also aligned to the audio source through the perceptual identification of tonal units and conclusive breaks.

In the L-AcT framework, the segmentation of the acoustic signal is done during the process of transcription and alignment. At least two trained annotators work on the same text. Intonation units are not defined in a strict formal way, nor identified via instrumental measures, but based on perceptual cues. They correspond to various phenomena, such as pauses, pitch resets, initial rushes or final lengthening. More holistically, the units manifest an overall coherent contour. From this perspective, the measurements of phonetic parameters should be called upon to explain what perception does naturally, i.e. the segmentation of the speech flow in tonal constituents. In the C-ORAL-ROM corpus we also measured the consensus between non-expert annotators in identifying prosodic breaks (MONEGLIA *et al.,* 2005). Global results scored a Kappa coefficient (FLEISS, 1971) between 0.766 and 0.920 on all the four languages (French, Italian, Portuguese, and Spanish).

Segmentation is a crucial cue in high level semantic and pragmatic interpretation of utterances. In fact, the same sequences of phonemes and morphemes can assume completely different values if segmented in different ways. The syntactic interpretation of an utterance is then guided by the prosodic segmentation. If we assume the perspective of the hearer, i.e. the *decoding* point of view, segmentation is then the mark through which we can recognize the higher level of organization of the speech flow. From the point of view of the speaker, segmentation could be instead considered a consequence of the mental organization of speech in pragmatically interpretable sequences, corresponding to Speech

Acts, and in their structuring in units of information. For these reasons, it is very relevant that work on spoken corpora takes prosodic segmentation into account.

Training is a complex process necessary to ensure an acceptable level of interpersonal agreement during the segmentation task. I was a member of the Italian team working in the C-ORAL-ROM Corpus in the early 2000s. We used a three-stage procedure in order to accomplish a transcription: a first annotator did the initial work, then a revisor checked the first segmentation and highlighted the problems and inconsistencies from her/his point of view, and finally a supervisor made the decision on controversial points.

3.3.2. TRANSCRIPTION AND SEGMENTATION: THEIR RELATION TO WRITING

The LABLITA corpora are transcribed using orthographic criteria, which already contain a certain (let's say high) level of abstraction with respect to the data source, which is a continuous audio streaming. On the other hand, orthographic transcription allows easy access to the corpora for studies that are not only limited to the phonetic or prosodic levels, as well lexical, syntactic, semantic and textual research. Orthographic transcription also allows the processing of corpora with standard computational tools (PoS tagging, first and foremost), usually calibrated on the written standard.

Moreover, Italian orthographic norms have been adapted for the representation of non-standardized forms or sequences occurring in spontaneous speech. In Italian, strong diatopic variation influences the regional variants at each linguistic level. Not to mention the dialectal variation, which is much bigger (the corpora, however only rarely include dialectal varieties. It is worth noticing that, even if the LABLITA corpora focus mainly on the Tuscan variety of Italian, specific conventions have been then adopted in order to represent all the variants contained in the resources.

Another point that distinguishes the transcription rules from the standard orthography is, of course, the marking of tonal breaks and the absence of punctuation. Tonal breaks and other signs for the representation of the linear (e.g. interruptions, fragmentations, retracting) and non-linear (e.g. overlapping) phenomena of speech constitute a text mark-up level needed in addition to the standard orthographic norms, in order to ensure a minimal representativity of the audio source.

3.3.3. TEXT-TO-SPEECH ALIGNMENT

Alignment is necessary to access audio from the text queries, and in the LABLITA corpora it is done based on the units that are considered relevant to the segmentation of speech, i.e. the prosodically terminated sequences. This choice also has a theoretical impact, since it implicitly states that prosody is the central element for the processing of speech.

Without the information given by alignment, it would not be possible to consider the real data recorded in speech interaction, and the oral corpora would then consist of a collection of transcriptions with no links to the primary data source.

More generally, each level of annotation requires specific and dedicated treatment when we process speech, and not written language. As already mentioned, the standard tools used for text computation are traditionally trained on written language, and they perform much more poorly on spoken data. From this perspective, prosodic segmentation constitutes primary data and should inform each level of annotation, starting from morphosyntactic tagging (PANUNZI *et al.*, 2004; BICK *et al.*, 2012).

3.4. MAJOR FINDINGS AND THEIR RELATION WITH THE CORPUS ARCHITECTURE

The main question we want to address with the research on spoken language is the relationship between prosodic forms and pragmatic values, within a framework that considers language as an interpersonal exchange based on individual acts of speech. The role of corpora is central, since L-AcT is a usage-based theory of language, in which communicative mechanisms are built starting from the speaker's affective state, through the interaction between participants.

From this perspective, the corpus drives the research and leads the researcher to new findings, in the sense that we are called to provide an explanation for phenomena as they are manifested in actual use. As TOGNINI BONELLI (2001, p.84) notes, "in a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence". The analysis of prosody always starts from examples in natural contexts, so having considerable variation in the collected contexts gives us better chances of finding a comprehensive repertoire of forms and phenomena.

Precisely in this sense, one of the most recent findings that will lead to further research in the near future is the extension of the definition of parenthetical structures in speech. In L-AcT, parentheticals are defined as information units occurring inside an utterance, introducing information with a metalinguistic value and a specific modality; they are prosodically characterized by a jump to a lower f_0 and intensity level (MONEGLIA; RASO, 2014). As has been recently noted, departing from corpus analysis (SACCONE, 2021), the parenthetical strategy in speech seems to go beyond the utterance level, characterizing wider portions of speech. The phenomenon could then be better treated as a general textual strategy, rather than within the limits of the information patterning framework.

3.5. PERSPECTIVES

Spoken corpus building is becoming an ever more complex task, which involves skills ranging from sound recordings, text and audio processing and querying, acoustic analysis, statistical testing. There are many skills needed to shepherd a linguistic resource from its collection to its exploitation. First, we must not forget that we are collecting data, so the methodologies of data collection play an important role in the process of building a resource. In this respect, there is a potential contradiction between two important points. On one hand, we need the best acoustic quality to produce a resource that is analyzable in a reliable way by the computational tools used for speech processing and annotation. On the other hand, if the goal is to document free conversation in different situations and pragmatic contexts, we need to record real interactions, which usually do not occur in a laboratory or in an anechoic chamber, with an optimal setting. The researcher has to keep these aspects in mind, and needs to balance them with respect to her or his goal.

Fortunately, nowadays the technology available allows us to record different situations with relatively less effort and good acoustic quality. But again, the availability of new technologies can increase the level of the challenge. As is well known, free conversation typically happens face to face, and the multimodal aspects of communication are increasingly capturing the attention of many researchers. It is very likely that the new generation of spoken corpora will be multimodal, the source will be video, and therefore the methodologies of collection will also include skills needed for filming the interactions. Moreover, new levels of analysis and processing will be added to traditional ones, namely gesture analysis, video processing and making multimodal resources available.

Next generation linguists dealing with speech data need to manage all of the aspects connected to the treatment of these kinds of data. Since it is not likely that a single researcher would master all the skills required to gather, compute and analyze multimodal corpora, it will be necessary to do work in well-structured teams, even larger than the ones within which we currently work.

4. THE C-ORAL-BRASIL CORPORA: BRAZILIAN PORTUGUESE AND OTHER LANGUAGES

4.1. THE CORPORA

The main C-ORAL-BRASIL corpora are medium-sized, prosodically annotated, and text-to-speech aligned. All the corpora feature PoS tagging and syntactic parsing (BICK, 2012; 2014) as well as a set of measurements and statistics about segmentation parameters and

speakers' metadata. Besides those features, the corpora always try to guarantee high acoustic quality in many different natural contexts. The corpora so far compiled are:

1. C-ORAL-BRASIL I (RASO; MELLO, 2012; RASO; MELLO, 2014; MELLO, 2014), dedicated to informal spontaneous speech, is made up of 218,130 words covering 139 texts (1/3 monologues, 1/3 dialogues and 1/3 conversations, i.e. dialogues with more than two main speakers); the texts contain approximately 1,500 words each; the corpus features 306 speakers. This corpus can already be queried through the DB-CoM platform at < www.c-oral-brasil.org/db-com> (MELLO, to appear), that allows queries considering the linguistic data, PoS annotation, besides all the sociolinguistic metadata information about speakers.

2. C-ORAL-BRASIL II (RASO *et al.,* to appear) features three different corpora: (i) formal in natural context; (ii) media; (iii) telephone. The whole C-ORAL-BRASIL II comprises 289,921 words. The formal in natural context corpus features 121,396 words in 74 texts, representing the following semantic domains: preaching, teaching, conferences, professional explanations, political speech, political debate, business and law. The media corpus (TV and radio) features 139,396 words in 101 texts. It is made up of texts of the following kinds of shows: interviews, meteorology, sports, news, reportages, scientific press and talk shows. An extra section provides 24,776 additional words from all different domains. They are separated from the rest because they would make the proportions of the corpus non-comparable with those of C-ORAL-ROM. The telephonic corpus provides 31,308 words in 79 texts. For the formal and the media corpora, the texts are differentiated between dialogic and monologic interactions. For the telephonic one, texts are differentiated as for private and public interactions.

The project has also compiled other corpora:

1. a corpus of Brazilian learners of English (COBAI-Lindsei-BR; <http://c-oral-brasil.org/cobai_lindsei_br.html>), which integrates the Lindsei project at the University of Louvain (https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html) (MELLO *et al.*, 2012). The corpus comprises fifty recordings and their transcriptions, which follow the Lindsei guidelines. The transcription guidelines include a code for each recording, speakers' turns, and the marking of several speech features, such as: overlapping, pauses, backchannelling, contractions, truncation, among others. The recorded informants were university, high intermediate to advanced level students of English as a second language. The recordings covered three different tasks: a narrative about a set topic chosen by the informant, free discussion with the interviewer and the

description of a pictured scene. Each recording is on average twenty minutes long and features quasispontaneous speech patterns. For each recording there is an accompanying learner profile that covers the learner's language history and other elements that might have contributed to her/his process of language acquisition, besides having information about the interviewer and the actual interview itself. There is no sound to text alignment thus far; however, we plan to have it in the future, along with the same scheme for prosodic boundary segmentation adopted by the C-ORAL corpora.

2. Corpus Oral de Língua Portuguesa Indígena - COLPI (<http://c-oral-brasil.org/colpi.html>) or Indigenous Portuguese Oral Corpus is a small sized oral corpus, which documents Brazilian Portuguese spoken as a second language by Brazilian Indigenous peoples (MELLO; MELLO, 2016). This corpus represents a first step in the attempt to document and make available data that so far has been scattered and not accessible to researchers. The recordings were made by an anthropologist in the course of her fieldwork and mostly document narratives, therefore portraying monologic texts. COLPI comprises twenty recordings, featuring 28,319 words and approximately 190 minutes of recording. These recordings are an excerpt of a much larger body of recorded data, which could not all be used in the corpus due to poor recording quality. The transcription guidelines followed those established for the C-ORAL-BRASIL corpora, with some necessary adaptations. The recorded texts represent a number of different indigenous ethnicities sharing stories. The recording topics are as follows: a. Kaxinawá/HuniKwin (provenance: Western Amazonia): oral traditions; b. Aweti, Kalapalo, Kamayurá, Kuikuro, Mehinaku, Waurá and Yawalapiti (provenance: Northern Mato Grosso): foundational myth also known as Kwarup in the High Xingu reservation area, where the recordings were carried; c. Baniwa, Desana and Tariano (provenace: Northern Amazonia): traditions and rites of passage; d. Guarani, Kaingang and Xetá (provence: Northern Paraná): collective interviews about their cultural traditions and current living conditions; e. Fulni-ô (provenance: Southern Pernambuco): cure rituals and practices.

3. Different informationally tagged minicorpora of roughly 30,000 words and 20 texts each, covering informal Italian (with texts extracted from the C-ORAL-ROM (CRESTI; MONEGLIA, 2005), American English (CAVALCANTE *et al.*, 2019; CAVALCANTE; RAMOS, 2016), extracted from the Santa Barbara Corpus of Spoken American English (DU BOIS *et al.*, 2000-2005), infomal Brazilian Portuguese (RASO *et al.*, 2018) and telephone Brazilian Portuguese (RASO *et al.*, 2019).

Besides these corpora, the project is currently compiling other resources: (i) a corpus of spontaneous speech by patients of schizophrenia, the C-ORAL-ESQ (FERRARI *et al.*, to appear), which foresees at least 40 interactions between patients and physicians, following the same transcription and segmentation criteria adopted in the C-ORAL-BRASIL corpora; (ii) a minicorpus of Angolan Portuguese extracted from 27 long recordings in varied natural situations (ROCHA *et al.*, 2018); (iii) new informationally tagged minicorpora from different sections of the main corpora.

The informational tagging follows the Language into Act Theory (L-AcT) (CRESTI, 2000; MONEGLIA; RASO 2014; CAVALCANTE, 2020). The minicorpora are crucial for research about the interplay bertween information structure and prosody. These minicorpora, along with a Spanish tagged minicorpus tagged at the LABLITA Lab (NICOLÁS; LOMBÁN, 2018), allow the analysis of information structure under a comparative perspective taking Romance languages and English into account. The published resources can be freely downloaded from the site <www.c-oral-brasil.org>.

The DB-CoM search query interface (<www.c-oral-brasil.org/db-com>) focuses on the incremental development and implementation of a multilevel search and query tool as well as a database portraying several spontaneous speech corpora. So far, the C-ORAL-BRASIL I corpus and its associated informationally annotated minicorpus are available for public searches. The interface allows for different types of multilevel queries, which profit from the rich PoS annotation scheme for the whole C-ORAL-BRASIL I corpus, as well as its metadata documents (interaction type, participants' profiles, etc). The queries performed in the minicorpus allow information structure to be taken into consideration in its full spectrum of sophistication, following the L-AcT tag set.

4.2. WHAT WAS THE PURPOSE IN BUILDING THE RESOURCES?

The primary purpose was to build corpora that could allow pragmatic and prosodic studies, without excluding other objects of study, such as the lexicon, syntax, phonetic aspects, among others. More precisely, we were interested in identifying how we could describe speech acts and information units using prosodic and pragmatic criteria; and how different kinds of interactions (mainly monologues, dialogues and conversations with more than two main participants) are structured from a pragmatic point of view, i.e., how they are structured in terms of speech acts and information structure.

Of course, this structural variation depends not only on the type of interaction, but also on more granular diaphasic variation. This is why the C-ORAL-BRASIL corpora reduce the amount of repetition of the same contextual setting and try to collect data from the widest range of situations. Some examples are: supermarket and other shopping activities, soccer game, broker showing an apartment to client, cards or other table games, dinner cooking,

engineer and worker at the work site, a band organizing a show, waiter at a party interacting with guests, etc. We tried to avoid repetitions of chats without any actional activity and interviews (which of course are more common in spoken corpora because they are easier to record), since they repeat the same pragmatic organization of speech.

A major limitation to studies of spoken structure in general is the lack of diaphasic variation, i.e., what we do while speaking in a certain setting and in a certain type of interaction. Concisely, our main goal is to show how speech changes as a function of diaphasic variation. However, we also tried to control diastratic variation (schooling, age and gender) in one specific diatopic area (the metropolitan area of Belo Horizonte) and avoided featuring the same speaker in too many texts.

4.3. THE ROLE PROSODY PLAYED IN CORPUS DESIGN

The role of prosody in spoken corpora is methodologically constitutive and central to several research interests. At least two aspects for which prosody is essential can be mentioned: (1) prosodic segmentation of speech into intonation units, distinguishing between boundaries that convey terminal and non-terminal functions. This issue will be focused on later; (2) study of prosodic forms of each intonation unit, establishing a direct relationship between prosodic features and the informational function of the unit. In order for this to be done, the acoustic quality of recordings must be sufficiently good. In fact, following L-AcT, the C-ORALBRASIL project assumes that there is substantial isomorphy between intonation unit and information function, and that the illocutionary function carried by a specific information unit (the Comment) is the core and mandatory unit of the utterance, conveyed by prosodic features.

4.3.1. PROSODY AS A THEORETICAL CHOICE

The main reason for segmenting speech prosodically is that intonation unit boundaries seem to delimit the compositional scope of word sequences. It would be easy to exemplify cases in which the same lexical sequence would generate completely different interpretations depending on the prosodic segmentations installed (IZRE'EL *et al.*, 2020). Different segmentations may give rise to a different number of utterances or to a different functional relation among intonation units in the same utterances. When we interpret a sequence of words (or when we produce them), the first step in their interpretation is dealing with prosodic segmentation. Only after that, other prosodic features, together with pragmatic and cognitive parameters (RASO; ROCHA, 2016), intervene to convey the illocutionary or the informational value of the unit; additionally some other prosodic features convey minor semantic information, like local lexical prominences. Of course, the syntactic and semantic interpretation of a speech sequence also depends on the segmentation.

4.3.2. TECHNICAL PROCEDURES FOR SEGMENTATION

The data are segmented based on perception of prosodic disjunctures in the speech flow. A group of annotators is trained to perceive prosodic ruptures; when they reach a Kappa agreement (FLEISS, 1971) of at least 0.8 they start the text transcriptions and segmentation. The annotators with the highest Kappa agreement carry the revisions. There are theoretical and practical aspects involved. To define an intonation unit is not an easy task (MELLO; RASO, 2018), since it seems there is no formal cue that allows its definition (as for example, higher voicing peak for a syllable nucleus, or stress for a stress group). Certainly, intonation is not the only feature that contributes to mark an intonation unit. If we try to define the intonation unit using its boundaries, we meet the same obstacle, since many features are involved in signaling boundary (and they are not always salient). It seems that the intonation unit is a very perceivable domain, but it is hard to define in a formal way. Usually, it is associated with some unit (phrase, clause, information unit). We could look at it as an interface between prosodic phenomena and a domain for linguistic functions. It is also probable that other factors, such as memory or motor functions, are responsible for some constraints of the intonation unit.

Many practical problems have been dealt with during the compilation of the C-ORAL-BRASIL corpora. Counting on the previous experience of the Lablita Lab and the C-ORAL-ROM corpora compilation (CRESTI; MONEGLIA, 2005) was very advantageous. Relying on this previous methodological basis, some new criteria were also implemented. That was done with basically three goals in mind: to simplify the annotation, when we judged that some criteria did not contribute to the project's mission; to improve segmentation reliability through the creation of a protocol that could reach more coherence and that could be statistically evaluated; to pay more attention to the acoustic quality of the recordings, which is crucial for prosodic studies. This last effort was facilitated by technological improvements that came through after the C-ORAL-ROM recordings had been concluded. The experience suggests that some steps should be followed in order to achieve reliable segmentations: (i) training of annotators. We trained the group that eventually segmented the corpus for some months. Before they were ready to segment, they had to reach an agreement of at least a Fleiss Kappa of 0.8 (FLEISS, 1971). Usually some annotators reach this agreement earlier than others; (ii) after the first round of segmentation, we foresaw a second round of revisions. During that second round, only the annotators that reached the best agreement in a new Kappa statistic were recruited. In the C-ORAL-BRASIL I, the average agreement was 0.86; (iii) a third round of revision was carried after the alignment.

4.4. THE MAJOR FINDINGS AND THEIR RELATIONSHIP TO THE CORPUS ARCHITECTURE

The C-ORAL-BRASIL corpora were built in order to study illocutions and information structure, but they can also be used for other kinds of work, such as lexical, morphosyntactic or phonetic studies. The corpus architecture privileges diaphasic variation, in order to allow the emergence of the highest possible variation of illocutions and information structures. Thanks to data collected following these guidelines, we were able to harvest a great variation of illocutions and information patterns, and to study them based on their prosodic form. These data have allowed and will allow much varied research in the interface between information structure (following the L-AcT model) and prosody. One of the most recent findings (CAVALCANTE, 2020) was the modeling of the three forms of the Topic information unit that had been described earlier (RASO *et al.*, 2018). Now we are working on the modeling of those units that we call "short information units" (just one phonological word isolated in an intonation unit). Previous studies (RASO; VIEIRA, 2016; RASO; GOBBO, 2019; RASO; SANTOS, forthcoming) have shown that we can differentiate several types of discourse markers through their prosodic forms; that also applies to parentheticals and other information units that in the literature are often mixed and confused, since they are analyzed without close attention to their prosodic form. The goal is to show that prosody, and not the lexicon or syntax, is responsible for conveying informational functions. These findings would not be possible without resource to prosodically annotated corpora, carrying very good acoustic quality and designed in order to represent diaphasic variation.

4.4.1. TRANSCRIPTION CRITERIA

The transcriptions follow orthography-based criteria, but many phenomena that are possibly undergoing processes of grammaticalization or lexicalization are recoverable because of some graphic conventions followed. Some of the phenomena that are exceptions to the orthographic criteria are (MELLO *et al.*, 2012): loss of verbal paradigm variability, contraction of prepositions, form reduction in pronouns and aphaeretic forms. This allows quantitative studies of spoken grammar. One of the major findings, thanks to the transcription criteria adopted, was to show the grammatical reason for pronoun reduction in BP (cf. *você ~ cê*) (FERRARI, 2015). It does not matter whether it is the full or the reduced morphological form of the pronoun that is used; what is relevant is its duration, which in BP is the main feature that marks stress: the stressed forms must be used in post-verbal position or when there is a pragmatic motivation, while the non-stressed forms are always syntactic subjects. Nevertheless, both full and reduced morphological forms can be used with any function, since both can be stressed or unstressed. Transcriptions must consider readability, computability, in addition to those speech phenomena that can account for lexical and morphosyntactic speech characteristics.

4.4.2. IMPORTANCE OF SOUND-TO-TEXT ALIGNMENT

Alignment is a mandatory tool for studying speech. Since speech is a process, and writing is a product that lacks all the information carried by the acoustic signal, in order to study speech we need to organize the spoken process and turn it somehow into an object that can be studied. This does not mean that speech becomes a product through alignment, but rather that alignment renders it reproducible, allowing us to easily repeat the process as many times as necessary in order to observe all the information conveyed. Aligning sound and transcription we can always have both channels (the written text and the acoustic features), we can listen to the sound repeatedly, if needed. Alignment software also allows several kinds of analyses. We believe that without alignment, it is not possible to study speech truthfully, since its main characteristics (all the phonetic information, prosody in the first place) cannot be systematically captured and quantitatively analyzed appropriately through speech analysis software (and scripts already available or the ones that can be created). Besides this, we need to say that speech is multimodal. The ideal scenario would be to also have the video aligned with text and sound. This would offer the possibility of understanding contexts better, and of studying co-speech gestures and facial expressions. However, technology still does not allow video recordings in different situations concurrently at present. For instance, it is very hard to record speech interactions in which speakers move around a lot. Nevertheless, we are trying to implement video recordings along with speech whenever it is possible.

4.5. BUILDING CORPORA OF BETTER-KNOWN LANGUAGES

The C-ORAL-BRASIL corpora feature a well-known language, which provided the opportunity for gathering as much data as possible. Despite the inherent difficulties involved in the compilation of spoken corpora, in the case of so-called better described languages, it is usually less complicated to approach speakers and get their permission to be recorded. No specific protocols, such as the ones necessary in indigenous community scenarios, are needed. Thus, researchers have a better chance of gathering larger corpora. In so doing, the types of statistics we apply can have larger range applicability in terms of being representative of certain phenomena in the language under study. The same goes for experimentation. Since the languages we study have been studied for a long time, and are frequently spoken by the researchers investigating them, it is not very difficult to conceive experiments that consider different cultural and interactional scenarios. The same can be said for experiments that deal with lexical, syntactic or even prosodic variation.

4.6. FUTURE PERSPECTIVES

New generations of linguists should receive very different training from that of linguists of past generations. Nowadays linguistics must be studied also in terms of big data and the necessary skills to tackle it. Even when quantitative studies are not the focus, there are software and scripts (also for phonetic studies) that allow us to capture important measurements in a more precise and reliable way. Therefore, a young linguist should have some training in programming and in statistics. Of course, now the knowledge available about how differently speech behaves from writing should lead to much more attention to spontaneous spoken data and, given the characteristics of speech, to prosody, which should not be treated any longer as a side, marginal discipline, but as the core constitutive aspect of speech.

5. CONCLUSION

Sections 2-5 above report on the resources developed for the study of spoken language by the authors of this paper and their teams. As a concluding note, we would like to summarize the major assumptions shared through the narrated experiences:

(a) natural spoken data collection requires planning, time and resources;

(b) data can only be collected if speakers have provided their permission to be recorded. Guidelines for ethical research should be followed and if applicable, projects should have been approaved by ethics committees before any data collection takes place;

(c) participants should play an active role in decisions made regarding the actual interpretation of data. This is even more relevant when the documented languages are minority or lesser-spoken languages;

(d) recordings should be done with the best possible equipment available to the team. For spontaneous speech corpora, wireless microphones and recorders allow for the documentation of a large range of situations;

(e) transcription criteria should be created for specific corpora. Standardized orthographies are not always the best choice;

(f) spoken corpora should provide sound to text alignment, in addition to the audio files and the transcription files;

(g) spoken corpus annotation depends heavily on the corpus goals; nevertheless, if made publicly available, corpus materials should always provide both the annotated and unannotated versions;

(h) when this does not go against ethical considerations concerning speakers' protection and respect for their privacy, data should be openly and freely shared with the scientific community at large. For this, citation rules should be clearly advocated and enforced by publishers, so that the huge amount of work done by the corpus creators is acknowledged. There should also be clear guidelines concerning re-use of existing corpora, especially concerning acknowledgement of e.g. additional annotations layers (e.g. by adding the annotator as co-author of the initial corpus), and concerning the enforcement of initial ethical rules regarding speaker anonymization etc. (so that for instance speakers' names, if anonymized in the initial corpus, are not disclosed when the corpus is re-used for other purposes; or so that initial conditions for access (e.g. involving registration) are duplicated for all subsequent public sharing of the data, including when the annotations have been enriched or modified within projects different from the initial corpus compilation, but using the same recordings).

REFERENCES

BICK, E. A anotação gramatical do C-ORAL-BRASIL *In:* RASO, Tommaso; MELLO, Heliana (eds.). *C-ORAL-BRASIL I*: corpus de referência de português brasileiro falado informal. Belo Horizonte: UFMG, 2012, p. 223-254.

BICK, E. The grammatical annotation of speech corpora: techniques and terspectives. *In*: RASO, Tommaso; MELLO, Heliana (eds.). *Spoken corpora and linguistic studies*. John Benjamins Publishing Company, 2014, p. 106-128. https://doi.org/10.1075/scl.61.04bic

BICK, E.; MELLO, H.; PANUNZI, A.; RASO, T. The annotation of the C-ORAL-BRASIL spoken corpus using an adaptation of the Palavras Parser. *In*: CALZOLARI, N. *et al. Proceedings of LREC 2012*, Paris: ELRA, 2012, p. 3382-3386.

CARON, B.; LUX, C.; MANFREDI, S.; PEREIRA, C. The intonation of topic and focus in Zaar (Nigeria), Tamasheq (Niger), Juba Arabic (South Sudan) and Tripoli Arabic (Libya). *In*: METTOUCHI, A; VANHOVE, M.; CAUBET, D. (eds.), *Corpus-based Studies of Lesser-described Languages*: The CorpAfroAs corpus of spoken AfroAsiatic languages. Studies in Corpus Linguistics 68. John Benjamins: Amsterdam-Philadelphia, 2015, 63-115. https://doi.org/10.1075/scl.68.03car

CAVALCANTE, F. *The information unit of Topic*: a crosslinguistic, statistical study based on spontaneous speech corpora. Ph.D. Dissertation. Belo Horizonte, UFMG, 2020.

CAVALCANTE, F.; RAMOS, A. The American English spontaneous speech minicorpus: architecture and comparability. *CHIMERA*: Romance Corpora and Linguistic Studies 3, vol. 2. 2016, p. 99-124.

CAVALCANTE, F.; RASO, T.; RAMOS, A. *Minicorpus de Inglês Americano*. 2019. <www.c-oral-brasil.org >

CHAFE, Wallace. *Discourse, Consciousness, and Time:* The Flow and Displacement of Conscious Experience in Speaking and Writing. Chicago: University of Chicago Press. 1994.

CHAFE, Wallace. Verbs and their objects and the One New Idea Hypothesis. *In*: MELBY, Alan K.; LOMMEL Arle R. (eds.), *LACUS Forum XXVI,* The Linguistic Association of Canada and the United States, 2000, p. 5-18.

CHANARD, C. (2015) Lexicon-aided annotation in ELAN. *In*: METTOUCHI, A; VANHOVE, M.; CAUBET, D. (eds.), *Corpus-based Studies of Lesser-described Languages*: The CorpAfroAs corpus of spoken AfroAsiatic languages. Studies in Corpus Linguistics 68. John Benjamins: Amsterdam-Philadelphia, 2015, p. 311-332. https://doi.org/10.1075/scl.68.10cha

CRESTI, E. *Corpus di italiano parlato*. Firenze: Accademia della Crusca, 2000.

CRESTI, E.; MONEGLIA, M. L'intonazione e i criteri di trascrizione del parlato adulto e infantile. *In*: MACWHINNEY, B. (ed.), *Il progetto CHILDES: strumenti per l'analisi del linguaggio parlato*, , Pisa: Edizioni del Cerro, vol. II, 1997, pp. 57-90 .

CRESTI, E.; MONEGLIA, M. (Eds.). *C-ORAL-ROM*. Integrated reference corpora for spoken Romance languages. DVD + vol. Amsterdam: John Benjamins, 2005. https://doi.org/10.1075/scl.15

CRUTTENDEN, A. *Intonation*. Second edition [Cambridge Textbooks in Linguistics]. Cambridge: Cambridge University Press. 1997.

DU BOIS, John W.; CUMMING, Susanna; SCHUETZE-COBURN, Stephan; PAOLINO, Danae. Discourse Transcription. *Santa Barbara Papers in Linguistics* 4. Santa Barbara, CA: Department of Linguistics, University of California, Santa Barbara, 1992.

DU BOIS, John W.; CUMMING, Susanna; SCHUETZE-COBURN, Stephan; PAOLINO, Danae. Outline of Discourse Transcription. *In*: EDWARDS, Jane A.; LAMPERT, Martin D. (eds), *Talking Data:* Transcription and Coding in Discourse Research. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1993, p. 45-89.

DU BOIS, J. W., CHAFE, W.L.; MEYER, Ch., THOMPSON, S.A., ENGLEBRETSON, R., MARTEY, N. *Santa Barbara corpus of spoken American English*, Parts 1-4. Philadelphia: Linguistic Data Consortium, 2000-2005.

FERRARI, L. *Aspectos prosódicos e sintáticos dos pronomes clíticos em português do Brasil e no vernáculo florentino*, Ph.D. Dissertation, Belo Horizonte: UFMG, 2015.

FERRARI, L.; ROCHA, B.; RASO, T. *The C-ORAL-ESQ Corpus*. To appear.

FERRÉ, G.; METTOUCHI, A. A Cultural Study of Open-Palm Hand Gestures and their Prosodic Correlates, *Proceedings of 10th International Conference on Speech Prosody* 2020, 25-28 May 2020, paper 21 (Online), Tokyo, Japan. 2020. <https://www.isca-speech.org/archive/SpeechProsody_2020/pdfs/21.pdf> and https://youtu.be/Ws3nydKYjLE . https://doi.org/10.21437/SpeechProsody.2020-58

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76.5, 1971, p. 378-382. https://doi.org/10.1037/h0031619

GOBBO, O. *Marcadores discursivos como unidades informacionais marcadas prosodicamente*. MA Thesis, Faculdade de Letras, UFMG, Belo Horizonte, 2019.

HIMMELMANN, N. P. Prosody in Language Documentation. In: GIPPERT, J.; HIMMELMANN N.P.; MOSEL ,U. *Essentials of Language Documentation* Berlin/New York: de Gruyter, 2006, p.163-181.

HIRST, Daniel; DI CRISTO, Albert (eds). *Intonation Systems:* A Survey of Twenty Languages. Cambridge: Cambridge University Press, 1998.

IZRE'EL, Sh.; METTOUCHI, A. Representation of speech in CorpAfroAs: Transcriptional strategies and prosodic units. *In*: METTOUCHI, Amina; VANHOVE, Martine; CAUBET, Dominique (eds.), *Corpus-based Studies of Lesser-described Languages*: The CorpAfroAs corpus of spoken AfroAsiatic languages. Studies in Corpus Linguistics 68. John Benjamins: Amsterdam-Philadelphia. 2015, p. 13–41. https://doi.org/10.1075/scl.68.01izr

IZRE'EL, Sh.; MELLO, H.; PANUNZI, A.; RASO, T. (eds.). *In search of basic units of spoken language*: a corpus-driven approach. Amsterdam: John Benjamins, 2020. https://doi.org/10.1075/scl.94

LEVINSON, S. *Pragmatics*. Cambridge University Press. 1983. https://doi.org/10.1017/CBO9780511813313

MEHLER, J.; JUSCZYK, P.; LAMBERTZ, G.; HALSTED, N.; BERTONCINI, J.; AMIEL-TISON, C. A precursor of language acquisition in young infants, *Cognition*, 29.2, 1988, p.143-178. https://doi.org/10.1016/0010-0277(88)90035-2

MELLO, H. Methodological issues for spontaneous speech corpora compilation: the case of the C-ORAL-BRASIL. In: RASO, T.; MELLO, H (Eds.). *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins, 2014, p. 29-68.

MELLO, H. DB-CoM: a query interface for the study of spoken corpora. To appear.

MELLO, H.; RASO, T. Speech segmentation in different perspectives: diachrony, synchrony, different domains, different boundaries, corpora applications. *Journal of Speech Sciences*, 7:2, 2018, pp. 1-8. *https://econtents.bc.unicamp.br/inpec/index.php/joss/article/view/14997*, *https://doi.org/10.20396/joss.v7i2.14997*.

MELLO, H.; RASO, T.; MITTMANN, M.; VALE, H.; CÔRTES, P. Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. *In*: RASO, T.; MELLO, H. (Eds.). *C-ORAL-BRASIL I*. Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG, 2012, p. 125-176.

MELLO, H.; AVILA, L.; NEDER NETO, T.; ORFANO, B. M. . LINDSEI-BR: an oral English interlanguage corpus. In: *Proceedings of the VII GSCP International Conference*: Speech and Corpora. Florença, Itália: Firenze University Press, 2012. v. 1. p. 85-86.

MELLO, G. B. R.; MELLO, H. COLPI: Compilation of an Indigenous Brazilian Portuguese L2 corpus. *CHIMERA*: Romance Corpora and Linguistic Studies, v. 3, 2016, p. 125-132.

MELLO, H.; RASO, T. Roundtable on Prosody and Corpora Compilation (guests M. Mithun, A. Mettouchi, A. Panunzi), *Abralin ao Vivo*, 20 May 2020: https://youtu.be/OmfrqPnr30Q

METTOUCHI, A. Audible breath intakes in monologues. *Journal of Speech Sciences*, 7:2, 2019, p. 93-106. https://econtents.bc.unicamp.br/inpec/index.php/joss/article/view/14999; https://doi.org/10.20396/joss.v7i2.14999

METTOUCHI, Amina; VANHOVE, Martine; CAUBET, Dominique (eds.), *Corpus-based Studies of Lesser-described Languages*: The CorpAfroAs corpus of spoken AfroAsiatic languages. Studies in Corpus Linguistics 68. John Benjamins: Amsterdam-Philadelphia. 2015. vi, 332pp+index. https://doi.org/10.1075/scl.68

METTOUCHI, A. Kabyle/French Codeswitching: a case study. *In*: LAFKIOUI, M.; BRUGNATELLI, V. (eds), *Berber in Contact:* Linguistic and Sociolinguistic Perspectives, Köln: Rüdiger Köppe, 2008, p.187-198. https://llacan.cnrs.fr/pers/mettouchi/pub/Codeswitching-Kabyle-French.pdf

METTOUCHI, A. Segmenting spoken corpora in lesser-described languages: new perspectives for the structural analysis of speech, Plenary talk at the 46th Annual Meeting of the Societas Linguistica Europaea, Split (Croatia) 18-21 September, 2013. https://llacan.cnrs.fr/pers/mettouchi/pub/SLE-2013-Split-Mettouchi-Plenary-Part11.pdf

METTOUCHI, A. The Interaction of state, prosody and linear order in Kabyle (Berber): Grammatical relations and information structure. *In*: TOSCO, Mauro (ed), *Afroasiatic:* Data and Perspectives, CILT, John Benjamins: Amsterdam-Philadelphia, 2018a, p. 261–285. https://llacan.cnrs.fr/pers/mettouchi/pub/Mettouchi-State-Prosody-and-Word-Order-in-Kabyle.pdf https://doi.org/10.1075/cilt.339.14met

METTOUCHI, Amina. Prosodic Segmentation and Grammatical Relations: The Direct Object in Kabyle (Berber) / Segmentação prosódica e relações gramaticais: o objeto direto em kabile (berbere). REVISTA DE ESTUDOS DA LINGUAGEM, [S.l.], v. 26, n. 4, p. 1571-1599, 2018b. http://doi.org/10.17851/2237-2083.26.4.1571-1599

METTOUCHI, A. From a corpus-based to a corpus-driven definition of clefts in Kabyle (Berber): Morphosyntax and Prosody. *Faits de Langue* 52/1, 2021.(To Appear).

MITHUN, Marianne. The Multidimensional Organization of Speech: Syntactic and Prosodic Structure. *Cadernos de Linguística*, v. 2, n. 1, p. 01-23, 26 Feb. 2021. https://doi.org/10.25189/2675-4916.2021.v2.n1.id287

MONEGLIA, M.; FABBRI, M.; QUAZZA, S.; PANIZZA, A.; DANIELI, M.; GARRIDO, J.M.; SWERTS, M. Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM corpus. *In*: CRESTI, E.; MONEGLIA, M. (eds), *C-ORAL-ROM*. Integrated Reference Corpora for Spoken Romance Languages, Amsterdam: Benjamins, 2005, p. 257-276. https://doi.org/10.1075/scl.15.09mon

MONEGLIA, M.; RASO, T. Notes on the Language into Act Theory. *In*: RASO, T.; MELLO, H. (eds), *Spoken Corpora and Linguistics Studies*. Amsterdam, Benjamins, 2014, p. 468–494. https://doi.org/10.1075/scl.61.15mon

MOON, CH.; PANNETON-COOPER, R.; FIFER, W. P. Two-day-olds prefer their native language. *Infant Behavior and Development*, 1993, v. 16, p. 495-500. https://doi.org/10.1016/0163-6383(93)80007-U

NICOLÁS MARTÍNEZ, C.; LOMBÁN SOMACARRERA, M. Mini-Corpus del español para DB-IPIC. *CHIMERA*: Romance Corpora and Linguistic Studies, 2018, v.5, n. 2, p. 197-215. https://doi.org/10.15366/chimera2018.5.2.002

NIEBUHR, O.; MICHAUD, A. Speech data acquisition: the underestimated challenge. *Kieler Arbeiten zur Linguistik und Phonetik (KALIPHO)*, v. 3, p. 1-42, 2015.

ORFÉO. https://www.projet-orfeo.fr/

PANUNZI, A.; GREGORI, L. DB-IPIC. AN XML Database for the Representation of Information Structure in Spoken Language. *In*: MELLO, H.; PANUNZI, A.; RASO, T. (eds), *Pragmatics and Prosody*: Illocution, Modality, Attitude, Information Patterning and Speech Annotation. Firenze, Firenze University Press, 2012, p. 133–150.

PANUNZI, A.; MITTMANN, M. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. *In:* RASO, T.; MELLO, H. (eds), *Spoken corpora and Linguistic Studies*. Amsterdam: Benjamins, 2014, p. 129–151. https://doi.org/10.1075/scl.61.05pan

PANUNZI, A.; PICCHI, E.; MONEGLIA, M. Using Pi-Tagger for lemmatization and PoS tagging a spontaneous speech resource: C-ORAL-ROM Italian. *In*: LINO M.T.; XAVIER, M.F.; FERRAIRA, F.; COSTA, R.; SILVA, R.. *Proceeding of LREC 2004*, Paris: ELRA, 2004, p. 563-566.

RASO, T.; CAVALCANTE, F.; BOSSAGLIA, G. *Minicorpus Português Brasileiro*. 2018. www.c-oral-brasil.org > corpora

RASO, T.; CAVALCANTE, F.; MITTMANN, M. Prosodic forms of the Topic information unit in a cross-linguistic perspective: a first survey. *In*: DE MEO, Anna; DOVETTO, Francesca Maria. *Proceedings of the SLI-GSCP International Conference*, 13-15 June, 2016. Roma: Aracne Editrice, 2018.

RASO, T.; FERRARI, L. Uso dei Segnali Discorsivi in corpora di parlato spontaneo italiano e brasiliano. In: FERRONI, R.; BIRELLO, M. (eds.) *La competenza discorsiva a lezione di lingua straniera*. Roma: Aracne, 2020.

RASO, T.; MELLO, H. (eds.). C-ORAL-BRASIL I: corpus de referência de português brasileiro falado informal. Belo Horizonte: UFMG, 2012.

RASO, T.; MELLO, H. C-ORAL-BRASIL: Description, Methodology and Theoretical Framework. In: BERBER SARDINHA, T. and SÃO BENTO FERREIRA, T. (eds.). *Working with Portuguese Corpora*, Bloomsbury, 2014;

RASO, T.; MELLO, H.; FERRARI, L. (eds.). *C-ORAL-BRASIL II*: corpus de referência de português brasileiro falado formal, media e telefone. www.c-oral-brasil.org, to appear.

RASO, T.; ROCHA, B. Illocution and attitude: on the complex interaction between prosody and pragmatic parameters. *Journal of Speech Sciences*, 2016, v.5 n.2, p.5-27. https://doi.org/10.20396/joss.v5i2.15062

RASO, T.; SOARES, E; AZEVEDO MIRANDA, I. *Minicorpus telefônico*. 2019. www.c-oral-brasil.org > corpora.

RASO, T.; VIEIRA, M. A description of Dialogic Units/Discourse Markers in spontaneous speech corpora based on phonetic parameters. *Chimera*: Romance corpora and linguistic studies, 2016, v. 3, n. 2, p. 221-249.

RASO, T.; SANTOS, S. Short information units: a corpus-based prosodic study on the lexeme assim in Brazilian Portuguese. *Journal of Speech Sciences*, forthcoming.

ROCHA, B.; MELLO, H.; RASO, T. Para a compilação do C-ORAL-ANGOLA: um corpus de fala espontânea informal do português angolano. *Filologia e Linguística Portuguesa,* 2018, 20(esp.), 139-157. https://doi.org/10.11606/issn.2176-9419.v20iEspecialp139-157

TAO, Hongyin. *Units in Mandarin Conversation*: Prosody, Discourse, and Grammar [Studies in Discourse and Grammar 5]. Amsterdam: John Benjamins, 1996. https://doi.org/10.1075/sidag.5

YATSIV-MALIBERT, I.; VANHOVE, M. Quotative constructions and prosody in some Afroasiatic languages: towards a typology. *In*: METTOUCHI, Amina; VANHOVE, Martine; CAUBET, Dominique (eds.), *Corpus-based Studies of Lesser-described Languages*: The CorpAfroAs corpus of spoken AfroAsiatic languages. Studies in Corpus Linguistics 68. John Benjamins: Amsterdam-Philadelphia. 2015, p. 117-169. https://doi.org/10.1075/scl.68.04mal