Journal of Speech Sciences 7(2): 01-08. 2018.
Available at: http://revistas.iel.unicamp.br/joss

# SPEECH SEGMENTATION IN DIFFERENT PERSPECTIVES: DIACHRONY, SYNCHRONY, DIFFERENT DOMAINS, DIFFERENT BOUNDARIES, CORPORA APPLICATIONS

**MELLO, Heliana**[1]
**RASO, Tommaso**[2]

[1]UFMG-CNPq-Fapemig
[2]UFMG-Fapemig

## 1 Speech Segmentation: a wide framework of discussions

This special issue of JoSS is one of a series of initiatives undertaken by the Lab of Empirical and Experimental Linguistic Studies (LEEL) at the Federal University of Minas Gerais (UFMG) together with several international partners on the important topic of speech segmentation, primarily segmentation applied to spontaneous speech. The other main initiatives are the following:

1. The "IX LABLITA and IV LEEL International Workshop: Units of reference for the analysis of spontaneous speech and their correlations across languages". This workshop took place in Belo Horizonte, Brazil, at UFMG in August 2015. Several teams from eight different countries organized a two-year interaction before and after a face-to-face meeting at the workshop. Each team segmented and tagged, according to their own theoretical framework, the same two English texts extracted from the Santa Barbara Corpus of Spoken American English (Du Bois et al. 2000-2005). During the workshop, each team presented their proposal, both on the English texts and on texts chosen by each scholar from a different specific language (Brazilian Portuguese, French, Hebrew, Italian, Japanese, Central Pomo, Russian, Upper Kuskokwim), focusing on different levels of linguistic analysis, but always using texts prosodically segmented. The discussion allowed the comparison of different approaches among teams that had independently reached some common conclusions on the importance of speech segmentation for the individualization of the main functional linguistic units. After the workshop, the interaction went on for several months and each group proposed a revised version of their segmentation after the common discussion. The results of this process, together with a database that allows both the listening and the reading of segmented and tagged texts will be published soon in (cf. 7 below). During the workshop, selected papers were also presented by several scholars. They were published in the item listed as 5 below.

2. The Workshop "V LEEL International Workshop: Spoken Corpora and Speech Segmentation" organized by the LEEL Lab at UFMG, with the participation of A. Mettouchi, from the École Pratique des Hautes Études (Paris), and several Brazilian scholars. In this meeting, the major topic of discussion was the automatic segmentation and tagging of spoken corpora.

3. The Workshop "Spoken Corpora advances: prosody as the crux of speech segmentation, annotation and multilevel linguistic studies" organized by T. Raso and H. Mello in Cape Town within the ICL20 in July 2018. In this opportunity, scholars from six countries discussed various aspects and applications of speech segmentation for one day.

4. The "X LABLITA and VI LEEL International Workshop: Prosody and Gesture: Corpus compilation, prominences and phrasing", organized by the LEEL Lab, at UFMG, in August 2019 with the participation of three international and five Brazilian invited scholars.

After the workshop, two courses about gesture structure and prosodic segmentation were organized.

5. Special Issue of the Journal *Chimera: Romance Corpora and Linguistic Studies* entitled *Approaching Diversity in Speech Studies: New Methodologies under Empirical Perspectives*, edited by G. Bossaglia, H. Mello and T. Raso, 2016. https://revistas.uam.es/index.php/chimera/issue/view/614

6. Special Issue of the Journal *Revista de Estudos da Linguagem* entitled *Speech segmentation*, edited by P. Barbosa and T. Raso (2018a): <http://periodicos.letras.ufmg.br/index.php/relin/issue/view/Speech%20Segmentation>, with an introduction by the editors (Barbosa and Raso, 2018b) that offers an in-depth survey about various theoretical and methodological aspects of research on speech segmentation, found at: <http://periodicos.letras.ufmg.br/index.php/relin/article/view/14303/pdf_1>

7. The forthcoming volume *In Search of a Basic Unit of Spoken Language: A Corpus-driven Approach*, edited by Sh. Izre'el, H. Mello, A. Panunzi and T. Raso, for John Benjamins. This volume brings a rich introduction about the phonetic and linguistic aspects of the study of both prosodic boundaries and the functionality of the units demarcated by them (Izre'el, Mello, Panunzi and Raso, forthcoming, a and b).

8. The last initiative we want to mention is a project, coordinated by T. Raso together with P. Barbosa, which aims at building an automatic tool capable of segmenting large corpora of spontaneous speech. For partial results so far, see Teixeira and Mittmann (2018); Teixeira (2018); Teixeira, Barbosa and Raso (2018a); Teixeira, Barbosa and Raso (2018b).

These initiatives have in common the subject of speech segmentation and a general empirical background but, at the same time, they host a very wide perspective allowing an in-depth look at the methodological, phonetic and linguistic problems that the topic offers, as well as at the applications that it allows. The participants bring with them the knowledge and the experience of studies for different languages, including minor and less studied languages like Central Pomo, Upper Kuskokwim and Kabyle. All the participants of the different initiatives work on empirical data, and most of them on big corpora of spontaneous speech from different languages. The main level of segmentation that has been treated at the workshops and the publications listed above is that of the intonation unit (or prosodic unit, or tone group, also known with other similar names), but different levels of segmentation are also dealt with in several presentations and papers, mainly those related to stress groups, syllables and Vowel-to-Vowel units. On one hand, some of the scholars who have contributed to these mentioned initiatives are more interested in the linguistic analysis of the units marked by the boundaries that separate intonation units; on the other hand, others are more interested in analyzing the phonetic features that mark the production and perception of the boundaries or in the elaboration of tools that may make automatic processes for speech segmentation possible. Therefore, diverse and complementary perspectives contribute to a better understanding of speech segmentation.

## 2 The papers in this volume

In this volume, six different papers are portrayed. They encompass the following themes: diachronic interest (Schweitzer), the application of the Language into AcT Theory (previously tested on English and many Romance languages) on Japanese (Cresti and Moneglia), the analysis of prosodic breaks between two different functional information units also adopting the L-AcT framework (Saccone, Vieira and Panunzi), spoken corpora segmentation (Bossaglia and

Ferrari), an algorithm to capture stress groups in French (Martin) and the correlation between breath intakes and terminal and non-terminal boundaries in Kabyle storytelling (Mettouchi).

The paper by Claudia Schweitzer, entitled "Etude sur le chant baroque français par segmentation accentuelle et intonative", studies the prosody of French Baroque music, using Piet Mertens' (1992) model and subsequent developments up to Mertens (2008) and a corpus of eight recitatives by five musicians. The paper analyses both the rhythmic and the intonational structure of the recitatives, concluding in agreement with Martin (2015 and 2018), that the prosodic structure is independent from the syntactic one.

The following three papers constitute a small group of researches that are homogeneous from a theoretical point of view, since all of them follow the Language into Act Theory (Cresti 2000; Moneglia & Raso, 2014), an extension of Austin's Speech Act Theory that integrates illocution in a wider prosodic-informational framework. However, the third in the sequence, presents a series of resources, useful also outside this paradigm.

The paper by Emanuela Cresti and Massimo Moneglia, entitled "Prosodic segmentation and functional correlations: the case of Japanese", tests the Theory with a non-Indo-European language, namely Japanese, for the first time. The results confirm the prosodic and informational predictions of the Theory and allow making very interesting syntactic observations, due to the particular canonical word order of Japanese in respect to the already well-studied languages (mainly English and Romance languages).

The paper by Saccone, Vieira and Panunzi, entitled "Complex illocutive units in the Language into Act Theory: an analysis of non-terminal prosodic breaks of Bound Comments and Lists" presents a study that compares the behavior of prosodic boundaries in two different complex structures according to the Language into Act Theory, namely Bound Comment (a processual, non-patterned sequence of illocutions) and List (a patterned sequence of illocutions) using two languages, Brazilian Portuguese and Italian. The study is carried out through the implementation of automatic analysis and statistical measurements, building the methodological basis for other studies of this type. The conclusion is that durational measurements seem necessary to catch the distinction between breaks from these two kinds of structures. While this could be possible for BP, since for this language normalized durations are available, it is still not feasible for Italian; therefore, duration was not considered in this study.

The paper by Giulia Bossaglia and Lucia Ferrari, entitled "The C-ORAL-BRASIL project: varied resources for the study of spoken Brazilian Portuguese", presents several resources dedicated to the study spontaneous speech, with a major focus on Brazilian Portuguese. These are: the corpora C-ORAL-BRASIL I (Raso & Mello, 2012), and C-ORAL-BRASIL II (Raso, Mello & Ferrari, forthcoming), that are prosodically annotated, as well as several informationally minicorpora tagged based on the Language into Act Theory framework, some of which (American English and Brazilian Portuguese) already downloadable, and two (Italian and Angolan Portuguese) still in progress.

The paper by Philippe Martin, entitled "Géneration automatique de la structure prosodique en français", presents a tool for generating the prosodic structure of French data. It is a useful tool in WinPitch software, and despite the assumption of the primacy of prosody over syntax, this does not seem to affect a purely phonetic interpretation of the output. Besides, the tool displays some flexibility that allows the users to correct errors and adjust parameters. The text presents a general discussion, which supports the steps of the algorithm, the description of the algorithm itself and the evaluation of these procedures.

The paper by Amina Mettouchi, entitled "Audible breath intakes in monologues", deals with audible breath intakes and their role in spontaneous speech. The study sheds new light on a phenomenon that has played a side role in prosodic studies and brings to the core the necessity

of the pursuit of further studies. The research addresses different issues related to the role of pauses and tone boundaries, and their correlation to audible breath intakes as expressive devices in folktales and recounts in Kabyle. The method of study of the data involved acoustic and perceptual analyses, along with tagging of different types of inbreath phenomena and their relation to pauses and tone boundaries.

## 3 The intonation unit

Since the main unit (but not the only one) treated in the whole set of initiatives listed in section 1 is the intonation unit (IU), we would like to present some reflection and very preliminary considerations about this kind of unit and its intrinsic particularity from a structural and a functional point of view[1].

When we consider minor prosodic domains like syllables (or V-to-V units), feet or stress groups, we easily find a specific element that defines the unit: the syllabic nucleus (with some difference in definition depending on the specific theory), the onset of a vowel, the duration, the stress. However, when we deal with the intonation unit we are working with a prosodic domain without a single specific feature that defines it. Regardless of the theoretical approach, there is no unique parameter that may account for what an intonation unit is: despite its denomination, intonation alone is insufficient to define the unit, since inside it and at its boundaries, at least duration (but often intensity too) plays a decisive role.

Moreover, if we try to define an intonation (or prosodic) unit (or any other way we might want to call this unit), it seems hard to avoid referring to its boundaries. Du Bois et al. (1992) and Chafe (1994) define the IU as a coherent f0 profile. However, this definition cannot be satisfactory at least for two reasons: (i) how do we define a coherent f0 profile? What interrupts the coherence of an f0 profile? A change of f0 movement direction? A change of f0 variation rate? An f0 reset or and f0 shift? In addition, for any of them, how strong the change must be in order for it to interrupt the coherence?; (ii) in many cases we clearly perceive boundaries without an evident change in the f0 profile. This means that we clearly perceive a change in IU, without any apparent f0 correlate marking this change. Therefore, we must admit that the IU may change independently of changes in the f0. This is the main reason that leads some scholars to define the IU with respect to its boundaries. Therefore, IUs are defined as the segmental material between two perceptible boundaries. This definition, thus, moves our question to what a perceptible boundary is, which is not an easier task, since, in a clearly circular way, the boundary is often defined as what delimits an IU.

We will not address here the various problems related to prosodic boundaries: how to define them according to different theoretical approaches or just based on perception; how to consider the multiple variables that can affect how boundaries are marked (difference among speakers, registers, boundary functions, etc.). What is important is that, no matter the approach or the variables, it is clear that prosodic boundaries depend on a varied set of parameters: silent pause, f0 (reset or f0 shift, variation rate, change of movement direction), change in articulation rate, change in intensity, change in duration (usually lengthening of the last syllable(s) before the boundary and shortening of the first syllable(s) after the boundary, when unstressed), among other known and perhaps unknown features. Except for silent pause, it seems that no feature alone guarantees the perception of a boundary. Moreover, even pause, which is not very

[1] We thank Marcelo Vieira for bringing this problem to our attention and for discussing it during a series of talks he gave to the LEEL members.

common in spontaneous speech (especially informal spontaneous speech) does not help when we need to understand the function of a boundary. This means that we perceive a lot of strong boundaries when there is no pause at all, and that even when we have a pause, we cannot say much about the unit that is demarcated by a boundary characterized by pause among its features. More information about the very complex question of how we can face the physical and functional composition of prosodic boundaries, in addition to a very rich bibliography about them can be found in Barth-Weingarten (2016), Raso & Barbosa (2018) and Izre'el et al. (forthcoming-b), among others.

This premise is to say that the IU seems to be a different kind of domain if compared to other prosodic domains. Syllable (or V-to-V unit), foot and accentual phrase (or stress group), all of them being characterized by a specific feature that marks the domain itself, can be easily defined with respect to one specific feature, while the IU cannot. On the other hand, this does not mean that we should place under discussion the legitimacy of such a domain, since its perceptual relevance is highly recognized in different theoretical frameworks, with a very high inter-rater agreement. Nevertheless, it does mean that we should better understand what motivates this kind of unit and how we can define and understand it. What follows does not have the presumption of offering a definition of IU, but it suggests a general direction to look for it and for its motivation in language in a very preliminary way.

We propose that the IU is not a domain strictly pertaining to the prosodic structure; it is a sort of result of the interface between the prosodic "space" and the linguistic functions it hosts (and perhaps some extra-linguistic and deeper process-related aspects). As for prosodic "space" we mean one level that hosts and integrates information not only related with a linguistic level but also information dealing with processes that include aspects of general cognition and biomechanical constraints. Prosody constitutes one of the several levels in which whatever material must be integrated and organized in some way. However, in the case of the IU, it seems especially hard (if at all possible) to define the domain just through formal means. It seems that we always need to refer to some function to fully understand the domain we are referring to. In some way, we can say that IUs are a "prosodic space" generated by the integration of several kinds of information (of a different nature). This specific space seems to be more easily definable as a necessary space to host major linguistic functions. Whatever the framework we adopt, it is generally necessary to refer to an IU as the prosodic space of what we can call *sentence* or *illocution* or *utterance*, but also of what can be called *information unit*. It seems that the IU cannot be defined only by a strictly prosodic structural point of view, and that, whatever formal definition we give, we always have some functional counterpart that is not structural. If it is possible to define the syllable or a stress group solely from a structural point of view, without entering their functional counterpart, it does not seem possible to do the same thing for the domain we call IU.

This peculiar aspect of the IU as a domain that is so difficult to be defined from an exclusively structural point of view, which at the same time is perceptually so salient and is so necessary when we face major linguistic functions, can help in a better understanding of some of the problematic issues we encounter when we study minor prosodic disjunctures from a perceptual point of view. Let us observe a few of the methodological difficulties and see some possible correlations with our view. (i) As we have seen, it is very hard to explain how a prosodic boundary is produced and perceived: what are the cues, or more precisely, the multiple combinations of cues, that convey this perception? Certainly, they are due to a fair amount of prosodic features, and probably also to some segmental ones and, maybe, when needed, some syntactic/semantic clues. This already seems to point to the direction we propose: the structural and the functional levels apparently interact very strongly; we do not find a true structural

principle, nor immediately see the meeting point of many features to perform a task. (ii) In phonological models like Pierrehumbert's (1980, 2000) and in ToBI (Silverman et al. 1992) notational system, it is easy to find a high agreement about the presence of a boundary, but it is very hard to find an acceptable agreement on the specific degree of disjuncture; this happens also in different theoretical frameworks when the goal is to distinguish among different strengths of boundaries, but not between the presence or absence of a boundary. On the other hand, much better results seem to have been achieved when the annotators were asked to distinguish between different functions of perceived boundaries, i.e. between terminal and non-terminal boundaries (Moneglia et al, 2005; Mello et al, 2012), that is, boundaries that convey the perception of conclusion of something (*sentence? Illocution? Something complete or interpretable?*) and boundaries that convey that the hearer still needs more information, in order to accept the sequence as concluded. It seems that some specific prosodic feature strongly conveys, very independently from the syntactic and semantic composition of the segmental material of the IU, the perception of continuity (or, in contrast, of terminality), which is functionally decisive to the interpretation of the message; should not it also be seen as a clue of the strict interaction between the prosodic (still, not clearly defined) level and its functional goal? In this case, it seems to us that it is not possible to follow a path that departs from structure to arrive at function; it is only the functional result that can help, guide us in finding which the complex formal features that may account (maybe in a flexible way) for the functional result are. (iii) Some authors, among them, in a very clear way Barth-Weingarten (2016), claim that prosodic boundaries are gradient and not categorical. This position is also confirmed by all the inter-rater agreement tests like Kappa statistic tests (Fleiss 1971), where even if it is easy to reach high or very high levels of agreement (easily above 0.8), there is also a significant percentage of disagreement; besides this, any scholar who tried to segment a spontaneous speech text in IUs knows that sometimes it is impossible to make a trustable choice without a theory-bounded decision. This seems to us to be caused by two possible reasons at the same time: one of them is that other clues are necessary to take a decision, clues that do not depend anymore from the prosodic structure (again, a prosodic structure that already places together many different features), but also from other linguistic and extra-linguistic levels; the other reason is that sometimes it simply does not make a real difference from a communicative point of view to place or not to place a boundary in a certain position. We think this might be another clue from the fact that the IU cannot be considered strictly as a level of prosodic structure, and should be considered as a sort of merge point were many structural and functional processes, and probably also other processes (like memory, articulatory limits etc.) interact to give form to a major linguistic message.

It is important to make it clear that this initial reflection should not be interpreted as the obvious common necessity in linguistics to match form and function, but as something different when we deal with the IU. In this case, it seems much more difficult to identify the IU as a specific formal level if we do not consider it as a construct that is somehow coordinated by functional needs. As we said, whatever formal definition we give for the IU, it still remains something that cannot be accounted for without a functional reference, both when we look at the unit itself and when we look at its boundaries.

The authors of this volume are grateful to Fapemig for financing the research.

**REFERENCES**

1. Barbosa P, Raso T. (Eds.) *Speech segmentation*. Special Issue of Revista de Estudos da Linguagem, 2018a.

2. Barbosa P, Raso T. Spontaneous Speech Segmentation: Functional and Prosodic Aspects With Applications for Automatic Segmentation / A segmentação da fala espontânea: aspectos prosódicos, funcionais e aplicações para a tecnologia. In: *Revista de Estudos da Linguagem*, 2018b.

3. Barth-Weingarten D. *Intonation Units Revised: Cesuras in talk-in-interaction*. Amsterdam/Philadelphia: John Benjamins, 2016.

4. Bossaglia G, Mello H, Raso T. (Eds.) *Approaching Diversity in Speech Studies: New Methodologies under Empirical Perspectives*. Special Issue of *Chimera: Romance Corpora and Linguistic Studies*, 2016.

5. Chafe W. *Discourse, consciousness and time: The Flow and displacement of Conscious Experience in Speaking and writing*. Chicago: University of Chicago Press, 1994.

6. Cresti E. *Corpus di Italiano parlato*. v. 1. Firenze: Accademia della Crusca, 2000.

7. Du Bois J, Cumming S, Schuetze-Coburn S, Paolino D. *Discourse Transcription*, Santa Barbara Papers in Linguistics 4. Santa Barbara: Department of Linguistics, University of California, 1992.

8. Du Bois J, Chafe W, Meyer C, Thompson S, Englebretson R, Martey N. *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium, 2000-2005.

9. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5, 1971: 378–382.

10. Izre'el, Mello H, Panunzi A, Raso T. (Eds.). *In Search of Basic Units of Spoken Language: A Corpus-Driven Approach*. Amsterdam-Philadelphia: John Benjamins, forthcoming-a.

11. Izre'el, Mello H, Panunzi A, Raso T. In Search of a Basic Unit of Spoken Language: Segmenting Speech. In: Izre'el, H. Mello, A. Panunzi and T. Raso (Eds.). *In Search of Basic Units of Spoken Language: A Corpus-Driven Approach*. Amsterdam-Philadelphia: John Benjamins, forthcoming-b

12. 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. R. (Ed.) *C-ORAL – Brasil I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG, 2012. p. 125-176.

13. Martin Ph. *The Structure of Spoken Language. Intonation in Romance*. Cambridge: CUP, 2015.

14. Martin Ph. *Intonation, structure prosodique et ondes cérébrales*. *Introduction à l'analyse prosodique.* London: ISTE, 2018.

15. Mertens P. L'accentuation de syllabe contiguës. *ITL*. 1992: 95/96: 145-165.

16. Mertens P. Syntaxe, prosodie et structure informationnelle: une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de linguistique.* 2008 : 56 : 97-124.

17. Moneglia M, Fabbri M, Quazza S, Andrea, Panizza A, Danieli M, Garrido JM, 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. (Ed.). *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins, 2005. p. 257-276.

18. Moneglia M, Raso T. Notes Language into Act Theory (L-AcT). In: Raso, T.; Mello, H. (Ed.). *Spoken Corpora and Linguistic Studies*. Amsterdam: John Benjamins, 2014. p. 468-495.

19. Pierrehumbert J. *Phonetics and phonology of English intonation*. 1980. Ph.D. Dissertation. Massachusetts Institute of Technology, 1980.

20. Pierrehumbert J. Tonal elements and their alignment. In M. Horne (Ed.), *Prosody: Theory and experiment*. Kluwer, Dordrecht, 2000: 11-26.

21. Raso T, Mello H. (Eds.). *C-ORAL-BRASIL I: corpus de referência do português brasileiro falado informa*l. Belo Horizonte: UFMG, 2012.

22. Raso T, Mello H.; Ferrari, L. (Eds.). *C-ORAL-BRASIL II: corpus de referência do português brasileiro*. forthcoming

23. Silverman K, Beckman M, Pitrelli J. ToBI: A standard for labeling English prosody. *International Conference on Speech and Language Processing* (ICSLP), v. 2, 1992: 867-870.

24. Teixeira B. *Correlatos fonético-acústicos de fronteiras prosódicas na fala espontânea*, Master Thesis – Universidade Federal de Minas Gerais, Belo Horizonte, 2018.

25. Teixeira B. & Mittmann, M. Modelos acústicos para a identificação automática de fronteiras prosódicas na fala espontânea. In: *Revista de Estudos da Linguagem*, 2018: 1455-1488.

26. Teixeira B, Barbosa PA, Raso T. Para a segmentação automática de fronteira na fala espontânea a partir de parâmetros prosódicos. In: Maria José Bocorny Finatto; Rozane Rodrigues Rebechi; Simone Sarmento; Ana Eliza Pereira Bocorny. (Org.). *Linguística de corpus*: perspectiva. 1ed. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2018a: 425-446.

27. Teixeira B, Barbosa PA, Raso T. Automatic Detection of Prosodic Boundaries in Brazilian Portuguese Spontaneous Speech. In: Villavicencio A, Moreira V, Abad A, Caseli H, Ramisch C, Oliveira HG, Paetzold GH. (Org.). *Lecture Notes in Computer Science*. 1ed.: Springer International Publishing, 2018b: 429-437.