An Experiment in Paratone Detection in a Prosodically Annotated EAP Spoken Corpus

Adrien Méli, Nicolas Ballier, Achille Falaise, Alice Henderson

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

An Experiment in Paratone Detection in a Prosodically Annotated EAP Spoken Corpus

Adrien Méli[1], *Nicolas Ballier*[1], *Achille Falaise*[2], *Alice Henderson*[3]

[1]Université de Paris, CLILLAC-ARP, F-75013 Paris, France
[2]Université de Paris, LLF, F-75013 Paris, France
[3]UGA, LIDILEM, 38400 Saint-Martin-d'Hères, France
adrienmeli@gmail.com, {nicolas.ballier, afalaise}@u-paris.fr,
alice.henderson@univ-grenoble-alpes.fr

Abstract

This article describes an experiment in paratone detection based on a spoken corpus of English for Academic Purposes (EAP) recently automatically re-annotated with prosodic information. The Momel and INTSINT annotations were carried out using SPPAS. The EIIDA corpus was chosen as it offered long uninterrupted stretches of speech of academic presentations. We describe the clustering method adopted for automatic detection, contrasting a supervised and an unsupervised method of paratone boundary detection. We showcase the relevance of the annotation scheme followed for this corpus and contribute to the investigation of the phonostyle of lecture delivery. We discuss the relevance of clustering methods applied to the labels of the pitch targets for the analysis of paratones.
Index Terms: discourse intonation, EAP phonostyles, clustering

1. Introduction

Brown and Yule proposed the term "paratone" to refer to "structural units of spoken discourse which take the form of 'speech paragraphs'" [1, pp.100-101], as when people who are asked to read written text aloud use certain intonational cues to mark boundaries between paragraphs. Previous research on automatic paratone detection was designed to foster information retrieval on audio documents [2]. A paratone detection classifier was trained on the Boston Directions Corpus [3] manually labeled for intonational boundaries using the Tones and Break Indices (ToBI) transcription convention [4].

The performance of the system is not indicated and the ToBI conventions do not a have specific convention for paratones. [3] used a classifier with the same data. The manual annotation of one speaker (spontaneous and read speech) followed the Grosz and Sidner (1986) theory of discourse structure [5]. Three types of labels reached sufficient inter-rater agreement: "segment-initial (SBEG), segment-final (SF), and segment-medial (SCONT, defined as neither SBEG nor SF)" [3]. Agreement was deemed to be satisfactory, but the annotation, to the best of our knowledge, is not publicly available.

Following on from the BASE [6], MICASE [7], and JSCC corpora [8], among others, the number and variety of EAP spoken corpora continues to increase, with university lectures having received the most attention [9]. The Aix-MARSEC database [10] offers several types of speech genres but a limited number of tokens for conference delivery styles. Overall, spoken corpora on EAP offering more than orthographic transcription are rare, with the prosodic transcription of the Hong Kong Corpus of Spoken English (HKCSE [11, 12, 13]) being a notable exception [14]. Moreover, treating spoken corpora like written text misses out on what Pickering and Byrd term the "acoustical realizations" of authentic spoken discourse [14, p.115]. With these corpus limitations in mind, we have re-annotated an existing EAP corpus in order to facilitate the analysis of prosodic structures and their variability in lecture delivery. The rest of the paper is organized as follows: section 2 presents the EIIDA corpus [15] and the prosodic annotation scheme. Section 3 outlines the two methods used for the detection of paratone boundaries. Section 4 presents the results. Our final section discusses the results and concludes.

2. The EIIDA Corpus

2.1. EIIDA project

The EIIDA corpus ("Études Interdisciplinaires et Interlinguistiques du Discours Académique", *i.e.* Interdisciplinary and Cross-linguistic Academic Discourse) is one of the first multilingual spoken corpora of specialized academic language. It can be used to carry out comparative linguistic analyses on written and spoken academic discourse (research articles vs. conference presentations), in two languages (English and French) and two fields (geochemistry and linguistics). The spoken corpus exists in a written form, and is searchable online through the ScienQuest interface[1] [16]. The corpus totals 180 texts (written and spoken) including approximately 900,000 lexical tokens. The spoken corpus (300,000 tokens) corresponds to roughly 20 hours of audio recordings. Our experiment is based on the English component of the corpus (77,000 tokens, 15 talks from 12 women and 3 men). In the current study we used only recordings by native speakers.

2.2. Prosodic Annotation

All sound files for the academic presentations in the EIIDA corpus were normalized for sample rate (11,000 Hz), format (wav) and channels (mono) using ffmpeg ([17]). Volume normalization across files was carried out by applying a filter implementing the R128 algorithm recommended by the European Broadcasting Union. A routine was used to split long sound files into smaller segments to feed SPPAS [18]. A Praat [19] script reintegrated the small SPPAS-generated TextGrids into a main TextGrid for the original long sound file.

The TextGrids obtained are exemplified in Figure 1. The first tier (Paratone) contains the the paratone boundaries. The second tier (Momel) shows the F0 targets identified, using the

[1]English: https://corpora.aiakide.net/?c=EIIDA-en
French: https://corpora.aiakide.net/?c=EIIDA-fr

http://dx.doi.org/10.21437/Interspeech.2021-294

Momel (Modelling Melody) algorithm [20] (refered later as "Momel values") based on asymmetric modal quadratic regression. The third tier (INTSINT), shows the INTSINT values (International Transcription System for Intonation) [21], which consists in 8 labels representing the annotation: T (Top), M (mid), B (bottom), H (Higher), L (Lower), S (Same), U (Upstepped) and D (Downstepped). The fourth tier (TokensAlign) is the token-aligned tier.

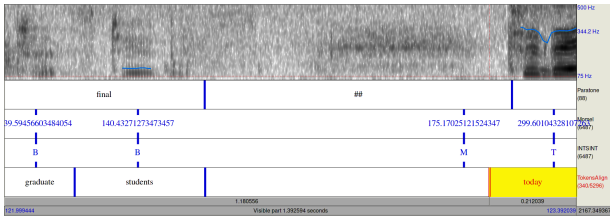

Figure 1: *Pitch reset and paratone boundary (talk #2)*

Momel and INTSINT values are automatically generated by SPPAS, but they require preliminary processing steps that are described below. For SPPAS to generate Momel and INTSINT tiers on a TextGrid as well as associated values, a Praat PitchTier, *i.e.* a list of F0 values at certain predefined time-steps, is required. The default values of the built-in function in Praat set the minimum pitch value at 75Hz, the maximum pitch value at 600Hz, and the time-step at 10ms. Following [22], a new PitchTier for each long sound file was generated with new pitch floors and ceilings. The new pitch floor is equal to 0.75 times the first quartile of all the F0 values obtained with default Praat values; the new pitch ceiling is equal to 1.5 times the third quartile. Given that each long sound file was split into shorter sound files of approximately 20 seconds, and that SPPAS was executed on these short files for more precise alignment rather than on the long sound file, for Momel and INTSINT TextGrids to be generated from the short sound files, a new PitchTier also had to be generated in Praat. This was done by using the updated pitch floor and ceiling calculated from the main PitchTier. One of the benefits of INTSINT is that it is not pre-empted by a theoretical framework applied to a given language. It has been successfully applied to several languages in Hirst and Di Cristo's volume [23], meaning that our procedure could later be applied to the talks in French from the EIIDA corpus.

3. Methods for Detection of Paratone Boundaries

For discourse analysis, it is worthwhile investigating speech at macrolevel, resorting to the largest prosodic unit in the prosodic hierarchy. For dialogical situations, [1] give examples of "identifying features" which they formulate in terms of acoustic cues and expected content: "Alternatively, the speaker can use a summarising phrase, often repeating the introductory expression, not necessarily low in pitch, but also followed by a lengthy pause. The most consistent paratone-final marker is the long pause, normally exceeding one second" [1, p.100]. These are dialogical features that may need to be refined in monologues.

3.1. A semi-supervised approach

Following Brown and Yule's pioneering observations on paratones, we manually analysed plausible candidates for paratone boundaries in relation to the SPPAS-generated annotations in the English recordings. Two annotators (specialists of phonetics) annotated one file by a male and one by a female speaker for paratone boundaries, to minimally cater for gender variation. The basic hypothesis is that the initial pitch reset of paratone is likely to be expressed by a T (top value), presumably followed by a lower value (L, B or D) and it should follow a sequence corresponding to final declination of the end of the preceding paratone (L, D or B being the expected labels). The manual annotation was first done independently of the INTSINT targets.

Figure 1 shows an example of speech reset, where the speaker resumes her main argumentation after a parenthetical about her latest publication ("competence in academic conversation skills for graduate students"). The spectrogram in Figure 1 shows how she resumes the presentation of her outline ("Today, I wanna connect three ideas ") with a pitch reset (shaded selection). The underlying question is the possibility to detect the boundaries of paratones on the basis of the observed Momel/INTSINT features. The succession of BBBB INTSINT labels is a likely candidate for paratone ending (and it should be noted that the preceding label was D, signalling a downstep). Conversely, the pitch reset at the beginning of the following paratone is marked by a T (top) to correlate with a paratone initial boundary. It should be noted that the algorithm captured a pitch variation during the inspiration (M), so that potentially parasitic targets like M (probably corresponding here to the speakers' medium range) may need to be taken into account in the pre-phonation zone.

We listed the candidates observed when agreement was reached among the two annotators and searched the corresponding patterns in INTSINT vectors (tiers) of the other recordings.

3.2. An unsupervised (clustering) approach

Due to gender imbalance in the corpus (12 women, 3 men), we did not distinguish female and male speakers for the unsupervised clustering analysis. In this approach, we search recurring patterns in the INTSINT tiers of the TextGrid files. As with most n-gram based studies, we set the threshold value to 3 INTSINT targets. The ceiling value was an empirical question, with an important theoretical proviso. The Obligatory Contour Principle [24, 25, 26] predicts that alternation is required in order for contours to be perceived. To respect this alternation, we set the upper limit of identical INTSINT targets to 7, and checked that no more than 6 identical successive INTSINT targets could be found. We used the tokenize_character_shingles() function from the R {tokenizers} package [27] to investigate the sequencing of the INTSINT representation. The function acts like an n-gram tokenizer over characters (namely, our sequences of T (Top), M (Mid), B (Bottom), H (Higher), L (Lower), S (Same), U (Upstepped) and D for Downstepped).

4. Results

4.1. Detection of the boundaries with the manual approach

Sharp rise in intensity and pitch proved useful to assess initial boundaries (cues more reliable than pause duration, sometimes under 0.45 s.). On 1,750 seconds of speech, the two annotators disagreed over 4 paratones (81 paratones detected in total). Inter-rater disagreements mostly occurred over filled pauses, followed by more nuanced pitch resets (HL) and whether the filled pause should be a kind of pre-head of the paratone or included in the inter-paratone break. The window of the clustering (a 3-gram, a 4-gram or 5-gram cluster of labels) and the combinatorial patterns with pitch extraction still needs to be fine-

tuned, but a sample of our observations reveals a pattern, as shown in Table 1.

Table 1: *Frequent INTSINT labels at paratone boundary (talk #2 and # 16, female and male speaker)*

Final sequence of labels (talk #2)	**Pitch reset labels**
BULHB	MT
DBBBB	MT
LBBUB	TL
BBBBB	TL
Final sequence of labels (talk #16)	**Pitch reset labels**
TLHLL	TLH
TLTDL	TLS
TLSDS	TLL
TLLTL	TLD
TLTDD	TLT
TLULU	TLS
TLSLT	TLD
TLTDU	TLS

In our two recordings, intra-speaker variation seemed to prevail over gender variability. The patterns are not strictly identical across the two speakers, though patterns emerge with 4 succesive targets instead of 6. 4-grams proved more conservative, due to optional repetitions or intermediary labels (L between T and B or D), and mostly correspond to 3-grams. Keeping in mind that S stands for same, a majority of downstepped, bottom and lower pitch targets are observed at the end of the paratone, whereas the following paratone begins with T or M (not unlike the labels "high head" and "low head" in the British tradition).

4.2. Prediction of the boundaries with a semi-supervised approach

Based on an auditory analysis of one of the speeches, we spotted the following candidates for the paratone final signatures : "BULHB", "DBBBB", "UDLSL", "SDLSB", "DSDUD", "LBBUB", "BBBBB". The 4-gram shows variation in boundary-initial and boundary-final position, but common features can be retained, with MTL a possible prototype for the pitch reset signature and B(U)B a frequent signature for boundary-final position. After our manual inspection, it seems that the prototypical signature of the initial boundary is TL, with a frequent pre-phonation (M) detected. As further evidence of our prototypical (M)TB(or L) initial signature, 2 answers in talk 2 were labelled for INTSINT and began with TTTTBH (repetitions of T corresponding to initial hesitation) and TLHU, validating our TB / TL signature for pitch resets.

The clustering analysis on the remaining files validated the presence of our candidate for clusters in the INTSINT annotation as seen in Table 2.

4.3. Prediction of the boundaries with an unsupervised approach

There are hardly more than 5 occurrences of identical 7-grams (BBBBBBU 5 occ., BBBBBBB 4, BBBBBUB 4) and 6-grams do not entirely capture the succession of the boundary final signature immediately followed by the boundary initial signature (potential candidates are : BBBBTL 5 occ., BUBTBB 5, BBBBBT 4, BBMTLL 4). Table 2 lists the most frequent 4-grams found in the whole corpus. If TL is the most reliable correlate of pitch reset, how valid is this signature for initial paratone detection? On 20 minutes of speech, the patterns MTL, TTL and TLS correspond to 24, 39 and 27 (respectively) potential paratone boundaries for an estimated number of 60 paratones based on the duration of the file.

Table 2: *Most frequent INTSINT labels at paratone boundary in the corpus*

Frequent sequences of INTSINT labels	**Frequency**
final (candidates)	
BBBB	52
BBBU	26
BUBB	21
BUBU	19
BBTB	18
TBBB	18
BBUB	16
initial (candidates)	
TLTL	23
BBBT	17
TLTB	15
UBBB	14
BTBB	13
DTLD	13
TDTL	13
TTLL	13

5. Discussion

More experiments need to be carried out to figure out the optimal criteria for the potential detection of paratone boundaries, whether based on raw Momel pitch extractions or symbolical INTSINT labels. We have not taken into account "intra-paragraph features" as reported in [28] but we spotted potential candidates. Explicit enumeration discourse markers ("first", "second", "third") were not necessarily realised as autonomous initial paratone boundaries. Conversely, when speech is explicitly structured with titles for subsections, the paratone is preceded by a pre-paratone header that isolates the topic of the paratone (the title of the subsection) and is then followed by a pitch reset (M)TLDS. In these cases where the paratone is preceded by a pre-paratone header that isolates the topic of the paratone, it is then followed not necessarily by a pause but by a pitch reset (M)TLDS. These pre-paratones headers are the shortest units detected.

5.1. Reliability of the annotation

As explained in section 2.2, alignment errors[2] are limited in scope with the 20-second window that has been used. Candidates most prone to errors are long pauses which are not detected or not labeled as silent pauses. For less than 10% of the data, the INTSINT annotation did not work for some subsections of the file. We intend to measure the cost of our global estimation for the PitchTier. A solution would be to be isolate the introduction of the speaker done by a chair at the beginning

[2][29] reported imprecision for phone alignment for French vowels and [30] reported imprecision for phone alignment on non-native English vowels.

of the talk and to measure the inter-annotation agreement with a PitchTier estimated on the whole file or on the speaker only.

5.2. Interpretability of the INTSINT annotation

The detection method is dependent on the interpretation to be given to the system. In that respect, S (for same) is troublesome for our clustering technique since its optional insertion may cause a sequence not to be counted. A systematic substitution of S by the preceding label would alter the initial INTSINT scheme. The interpretation of INTSINT is not only symbolical, it is also dependent on the signal. If the relevance of the pitch level assigned by the label is guaranteed by the association with the Momel pitch target [31] and its estimation in Hertz, any voiced speech event can potentially be captured and transcribed in INTSINT labels. A case in point is laughter, which results in pitch targets picked up by the system. For example in Figure 2, the UTL sequence actually corresponds to laughter and is "inserted" between the final BBB and the initial TTL pitch reset sequence of the following paratone. Such a paratone boundary is not detected by a concatenation of the final signature and of the final signature (BBBTTL), suggesting that initial boundary detection is more robust and that 4 is probably the optimal upper value for INTSINT sequences for initial/final boundary detection.

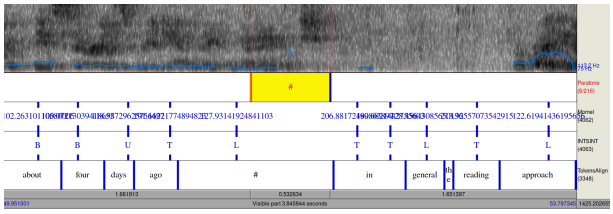

Figure 2: *"Parasitic" UTL sequence caused by audience laughter (talk#16)*

5.3. Biclustering and other clustering algorithms

We have relied only on INTSINT coding, while other phonetic and discourse correlates could be taken into account. Duration of breaks between paratones could be used, and the duration of paratones could be analysed to check the plausibility of the paratones postulated by the clustering. Then tokens could be incorporated in the analysis. Our qualitative analysis of the paratones confirmed the relevance of discourse markers. Discourse markers like "Now" and "So" proved useful to delimit initial paratone boundaries. To take into account this multidimensional (and multi-tier) investigation of paratones, a biclustering method could be applied. Using R packages such as textgRid [32], it could be possible to exploit the other tiers for biclustering, using an R package like multiClust [33]. These technological possibilities would allow us to combine features from discourse (a paratone boundary is unlikely just after "and"), intensity peaks, duration of the paratones and of the inter-paratone intervals. None of these cues seem to be enough on its own to be a discriminant feature. A more systematic investigation of the clustering algorithms would have required a comparison of the relevance of agglomerative (bottom-up) algorithms like AGNES or divisive (top-down) algorithms like DIANA, in order to disentangle the optimal number of pitch targets to take into account in our clustering analysis. Clearly, several alternative clustering methods are applicable to the INTSINT data.

5.4. Future Research

Although the EIIDA metadata is quite complete for the conditions of the recording, one dimension that is not taken into account is the existence of a prepared script for the talks. Some speakers actually read their speech, as evidenced by repairs typical of misreadings and a faster speech rate. For talk 2, the speech rate (number of syllables as detected by de Jong and Wempe (2009) [34] algorithm divided by the total duration of the file) is 4.26 and the articulation rate (number of syllables as detected by the algorithm divided by phonation time) was 5.12. It is tempting, then, to resume the investigation of paratones taking into account speech rate and this feature (read speech/improvised speech with notes) and its consequence on the plausible number of INTSINT labels retained for the characterisation of paratone boundaries. The difference in speech delivery was not taken into account for an across-speaker investigation of paratones. In this light, machine learning analysis of paratones with this type of data can be envisaged. One of the obvious applications of this line of investigation would be for text-to-speech systems when processing scientific papers.

Another related line of investigation for EAP would be to try to see the connections between these bigger units, the paratones, and the scientific moves that have been analysed in scientific journals. In this sense, analyzing conference speeches may be interesting for some specific delivery modes or their conventional subparts, such as the initial joke very often cracked at the beginning of the talk. Supplementing the resource with a prosodic layer of annotation will enable linguists to analyse phonostyles [35], which have seldom been investigated for academic discourse *per se* (but see [36] and [37] for hedges, or [38] for collocates).

6. Conclusion

In this paper, we have presented the annotation workflow for an EAP spoken corpus and its use for paratone detection. For automatic detection, our method might prove more robust than a detection based on the duration of silent pauses. One of our findings is that the duration of the pauses between paratones is much shorter than we initially assumed. For the female speaker, the median duration of the inter-paratone breaks is 615 ms and a fourth of them is below 430 ms. If final copyright clearance is granted by the corpus owners, we plan to upload the resulting TextGrids to the Ortolang Speech and Language Data Repository [39]. Our TextGrids could be used to foster research on paratones. We hope to have shown the interest of the method to detect pitch resets with the initial (M)TL pattern, thus vindicating Daniel Hirst's claim about INTSINT: "The possibility of extracting a symbolic representation of an intonation pattern automatically from the acoustic data opens a number of interesting perspectives for future research." [21, p.40].

7. Acknowledgements

We thank the reviewers for comments on an earlier version of this paper. Thanks are due to Taylor Arnold for helping us with clustering. The prosodic annotation of the corpus was funded by a grant to Alice Henderson's lab from CORLI, the French Linguistic Consortium for Corpus, Language and Interaction, via their 2019 call for projects dedicated to corpus delivery (*Appel à Finalisation de corpus*).

8. References

[1] G. Brown and G. Yule, *Discourse analysis*. Cambridge University Press, 1983.

[2] J. Choi, D. Hindle, F. Pereira, A. Singhal, and S. Whittaker, "Spoken content-based audio navigation (scan)," in *Proceedings of the ICPhS*, vol. 99. Citeseer, 1999.

[3] J. Hirschberg and C. H. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues," in *34th Annual Meeting of the Association for Computational Linguistics*, 1996, pp. 286–293.

[4] K. E. A. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "TOBI: a standard for labeling English prosody," in *The Second International Conference on Spoken Language Processing, ICSLP 1992, Banff, Alberta, Canada, October 13-16, 1992*. ISCA, 1992. [Online]. Available: http://www.isca-speech.org/archive/icslp_1992/i92_0867.html

[5] B. Grosz and C. L. Sidner, "Attention, intentions, and the structure of discourse," *Computational linguistics*, 1986.

[6] BASE, "British Academic Spoken English Corpus," 2008. [Online]. Available: http://www.helsinki.fi/varieng/CoRD/corpora/BASE/

[7] MICASE, "Michigan Corpus of Academic Spoken English," 2007. [Online]. Available: https://lsa.umich.edu/eli/language-resources/micase-micusp.html

[8] JSCC, "John Swales Conference Corpus," 2006. [Online]. Available: http://jscc.elicorpora.info/

[9] B. Paltridge, *Genre and English for Specific Purposes*. Wiley Blackwell, 2012.

[10] C. Auran, C. Bouzon, and D. Hirst, "The AixMARSEC project: an evolutive database of spoken English," in *In Bel, B. Marlien, I. (eds) Proceedings of the Second International Conference on Speech Prosody*, 2004, pp. 561–564.

[11] HKCSE, "Hong Kong Corpus of Spoken English," 2008. [Online]. Available: http://rcpce.engl.polyu.edu.hk/HKCSE/

[12] W. Cheng, C. Greaves, and M. Warren, "The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic)," *ICAME Journal*, vol. 29, pp. 47–68, 2005.

[13] ——, *A Corpus-driven Study of Discourse Intonation*. John Benjamins Publishing Company, Nov. 2008. [Online]. Available: https://doi.org/10.1075/scl.32

[14] L. Pickering and P. Byrd, *Investigating connections between spoken and written academic English: Lexical bundles in the AWL and in MICASE*. University of Michigan Press, 2008, ch. 6, pp. 110–132.

[15] EIIDA, "*Etudes Interdisciplinaires et Interlinguistiques du Discours Académique*," Interdisciplinary and Cross-linguistic Academic Discourse, 2017. [Online]. Available: http://www.transfers.ens.fr/eiida-etudes-interdisciplinaires-et-interlinguistiques-du-discours-academique

[16] A. Falaise, A. Tutin, and O. Kraif, "Une interface pour l'exploitation de corpus arborés par des non informaticiens: la plate-forme ScienQuest du projet Scientext," *Traitement Automatique des Langues*, vol. 52, no. 3, pp. 241–246, 2011.

[17] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[18] B. Bigi, "SPPAS - Multi-lingual approaches to the automatic annotation of speech," *The Phonetician*, vol. 111-112, pp. 54–69, 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01417876

[19] P. Boersma and D. Weenink. (2019) Praat: doing phonetics by computer [Computer program]. Version 6.1.07, retrieved 26 November 2019 from http://www.praat.org/.

[20] D. Hirst and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic spline function." *Travaux de l'Institut de Phonétique d'Aix*, vol. 15, pp. 75–85, 1993.

[21] D. Hirst, "Phonetic and phonological annotation of speech prosody," in *(a cura di) Analisi prosodica. teorie, modelli e sistemi di annotazione, Atti del II Convegno Nazionale della Società Italiana di Scienze della Voce (AISV)*, R. Savy and C. Crocco, Eds. Torriana: EDK, 24-29: Università degli Studi di Salerno 30 novembre-2 dicembre 2005, 2006.

[22] ——, "The analysis by synthesis of speech melody: from data to models." *Journal of Speech Sciences*, vol. 1, no. 1, pp. 55–83, 2011. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01491728

[23] D. Hirst and A. Di Cristo, *Intonation systems: A survey of twenty languages*. Cambridge University Press, 1998.

[24] W. R. Leben, "Suprasegmental phonology." Ph.D. dissertation, Massachusetts Institute of Technology, 1973.

[25] J. A. Goldsmith, "Autosegmental phonology," Ph.D. dissertation, Massachusetts Institute of Technology, 1976.

[26] ——, *Autosegmental and metrical phonology*. Basil Blackwell, 1990.

[27] L. A. Mullen, K. Benoit, O. Keyes, D. Selivanov, and J. Arnold, "Fast, consistent tokenization of natural language text," *Journal of Open Source Software*, vol. 3, p. 655, 2018. [Online]. Available: https://doi.org/10.21105/joss.00655

[28] A. Peiró-Lilja and M. Farrús, "Paragraph prosodic patterns to enhance text-to-speech naturalness," in *Proc. 9th International Conference on Speech Prosody 2018*, 2018, pp. 612–616. [Online]. Available: http://dx.doi.org/10.21437/SpeechProsody.2018-124

[29] B. Bigi and C. Meunier, "Automatic segmentation of spontaneous speech / segmentação automática da fala espontânea," *Revista de Estudos da Linguagem*, vol. 26, no. 4, pp. 1489–1530, 2018. [Online]. Available: www.periodicos.letras.ufmg.br/index.php/relin/article/view/13026

[30] A. Meli, "A longitudinal study of the oral properties of the French-English interlanguage : a quantitative approach of the acquisition of the /I/-/i:/ and /U/-/u:/ contrasts," Theses, Université Sorbonne Paris Cité, Apr. 2018. [Online]. Available: https://tel.archives-ouvertes.fr/tel-02309362

[31] D. J. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation," in *Proceedings of the XVIth International Conference of Phonetic Sciences*, vol. 16, 2007, pp. 1233–1236.

[32] P. Reidy, *textgRid: Praat TextGrid Objects in R*, 2016, r package version 1.0.1. [Online]. Available: https://CRAN.R-project.org/package=textgRid

[33] N. Lawlor, P. Guan, A. Fabbri, K. Karuturi, and J. George, *multiClust: multiClust: An R-package for Identifying Biologically Relevant Clusters in Cancer Transcriptome Profiles*, 2018, r package version 1.12.0.

[34] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[35] Léon, Pierre, *Précis de phonostylistique : parole et expressivité*. Paris: Nathan, 1993.

[36] D. Poos and R. Simpson, "Cross-disciplinary comparisons of hedging," *Using corpora to explore linguistic variation*, pp. 3–23, 2002.

[37] A. Mauranen, "They're a little bit different," *Observations on hedges in academic talk. In Aijmer, K. & Stenström, AB (eds.), Discourse patterns in spoken and written corpora*, pp. 173–197, 2004.

[38] W. Cheng, C. Greaves, and M. Warren, *A corpus-driven study of discourse intonation: the Hong Kong corpus of spoken English (prosodic)*. John Benjamins Publishing, 2008, vol. 32.

[39] J.-M. Pierrel, "Ortolang. une infrastructure de mutualisation de ressources linguistiques écrites et orales," *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle*, vol. 11, no. 11-1, 2014.