Challenging the perceptual relevance of prosodic breaks in multilingual spontaneous speech corpora: C-ORAL-BRASIL / C-ORAL-ROM

Massimo Moneglia [1], *Tommaso Raso*[2], *Maryualê Malvessi-Mittmann*[2], *Heliana Mello*[2]

[1] LABLITA, University of Florence, Italy
[2] Faculty of Letters, UFMG, Brazil

moneglia@unifi.it, tommaso.raso@gmail.com, maryuale@gmail.com, heliana.mello@gmail.com

Abstract

A Corpus of Brazilian Portuguese (BP) will join C-ORAL-ROM [1] adopting the same corpus design and prosodic annotation schema. The inter-rater agreement concerning the annotation of terminal and non terminal breaks by both experts and non experts is studied and compared with the early C-ORAL-ROM results [2]. Although the overall prominence of prosodic breaks is confirmed (K > 0.80) the inter-rater agreement for terminals turns out satisfactory only for the experts (0.76). Moreover the annotation of non terminal breaks shows low reliability and suffers of language specific factors connected to the rhythmic structure of BP [3:179-184]. The paper focuses on the qualitative analysis of the language contexts types determining the low inter-rater agreement and highlights both language specific and general factors which interact with perceptual prominence of prosodic breaks in BP.

1. Introduction

The Evaluation of perceptual relevance of prosodic breaks in natural language performance can profit from the results of almost one decade of studies. Within the ToBI annotation schema, inter-rater agreement tests have regarded first non-spontaneous speech resources [4] [5] [6]. These studies show a substantial consistency among listeners in their perception of prosodic cues and specifically for prosodic boundaries (phrasal boundaries and boundary tones). More recently, the annotation of phrasal boundaries and boundary tones has been tested also on spontaneous speech resources. A reasonable level of agreement has been recorded on American English [7] [8]. The K value turns out, however, more satisfactory for phrasal boundaries (K =0.66) with respect to the annotation of boundary tones and confirms that the agreement values obtained on boundary tones are in general higher than those obtained for prominence [9] [10]. However ToBI schema strictly requires expert labelers.

On the other hand, the capacity of naïve competent speaker to retrieve prosodic breaks in spontaneous speech data have been used successfully within large language corpora collection initiatives, with pretty coherent findings that can be roughly compared provided that agreement is recorded with comparable metrics [9] [11] [12][1] in accordance with de facto standards [13].

It has been shown in [14] that, when natural speech is presented to native Russian speakers, the K agreement rate [15] [16] is 0.92 (Cohen) 0.76 (Fleiss). In [1] has been shown that non-expert transcribers, after a brief training, can obtain a substantial consistency of inter-annotator agreement rates for the annotation of "weak" and "strong" prosodic breaks in a sub-corpus of the spoken Dutch corpus (K-coefficient between 0.61 and 0.80).

Important studies have been accomplished during the preparatory studies of the CoSIH corpus of spoken Hebrew. [17] finds that 80% of the break positions of a short spontaneous narrative have been agreed by all annotators.

The prosodic tagging of the multilingual Romance corpus C-ORAL-ROM has been evaluated by a third party [2] in a large scale evaluation performed by mother tongue non expert annotators. Data show, at the cross-linguistic level, a strong agreement in the Romance languages specifically for the annotation of terminal breaks. Terminal breaks have been confirmed by evaluators over 94% of the time for all languages and also K (Cohen realistic) reaches high values in all resources (>8 in all Romance languages, except French) with better scores for what regards dialogic contexts with respect to monologic ones in spontaneous speech performances. This high agreement shows that the annotation of the utterances in terms of their prosodic breaks is able to capture relevant perceptual facts, furthermore considering that the resource is multilingual it appears that the proposed coding scheme can be applied across languages.

Prosodic breaks cannot be determined solely on the basis of acoustic data without a perceptual validation. The correlation between acoustic analysis of prosodic boundaries set against human perception have shown that the perception of breaks is a function of the simultaneous activation of various cues: (1) final lengthening; (2) initial rush; (3) pitch reset; (4) fall of intensity (5) pause (6) rhythm ([11] [3]; [18] among many others). However the problem with the perceptive relevance of prosodic boundaries is the underdeterminacy of the reasons which determine the perception of a break, which eventually does not depend on the previous set of prosodic cues, but rather on their syntactic and informational correlations. Perceptual judgments may be sometimes not coherent with acoustic data [17]and for this reason the study of those language contexts in which the inter-rater agreement is critical is also relevant.

The evaluation accomplished on the Brazilian Portuguese (BP) corpus aims to test the overall hypothesis that prosodic breaks, especially terminal ones, have strong perceptual salience. However the evaluation will also assess the reliability of the annotation schema adopted in the C-ORAL-ROM when applied to speech corpora of BP, verifying whether its specific prosodic features may lead, in specific language contexts, to differences in perceptual relevance of these notions.

2. Inter-annotator agreement CORAL-BP

The inter annotator agreement test on the C-ORAL-BRASIL corpora presented here has been accomplished by two groups of 3 annotators, respectively qualified as expert and non expert. Annotators of Group 1 were Linguistics students at the Master or Doctoral levels, while Group 2 were

Linguistics students at the undergraduate level. All annotators attended an 8 hour workshop on speech segmentation. This took into account tonal units and the recognition of terminal and non-terminal breaks. During this course, they undertook three segmentation practices. Annotators of Group 1 took a supplementary 15 hour minicourse and 8 hour workshop on the relation between prosody and informational structure.

Two continuous dialogs (marked “dl” in the tables) and two continuous monologs (marked “mn” in the tables) were used with both groups for a total of 22 minutes ofspeech. The selection of language context was intended to ensure representativeness of the C-ORAL-BRASIL sampling. The first dialog (822 words) takes place between a costumer and a sales person at a shoe store. The second dialog (719 words) takes place at a university office. This dialog was featured by two people of rural provenance. The first monolog (855 words) features a woman who narrates her giving birth. The second monolog (784 words) features a middle aged man narrating the medical consequences of an accident and have opposite speed characteristics.

Each transcriber worked autonomously. The chore was to segment the texts which were already transcribed segmentally, marking the *terminal* and *non-terminal* breaks based on perception, without any access to the prosodic curves. In order to do this, they had to listen to a stretch of the sound signal through a sound editor, experiment with short stretches in order not miss the intermediary terminal breaks which within the speech flow could be missed. For each text segmentation, the annotators had 4 days for presenting the results of work.

Word boundaries (possible positions for prosodic breaks) are classified for the purpose of the evaluation into the following classes: 1) no break (tagged as 0); non-terminal break (tagged as /); terminal break (tagged as //).

The specifications given to annotators are those adopted in the C-ORAL-ROM validation and make reference to the sole ability to mark prosodic breaks on the basis of their perceptual relevance [2].

Fleiss’s K has been calculated to record the agreement respectively on *terminal*, on *non terminal*, and on all breaks, independently of their qualities. The K values have been derived in two ways considering all possible position for a break and alternatively considering as baseline those positions that have been marked by at least one annotator (realistic). The algorithm was run for each group separately and joining the two groups, for each text, for each text type and for the whole collection.

3. Discussion

3.1. K-score

The inter annotator agreement test on the C-ORAL-BRASIL The strong perceptual relevance of prosodic breaks in spontaneous speech corpora is confirmed for BP. The inter-annotator agreement test performed on the basis of C-ORAL-BRASIL allows to verify that the detection of this salience is independent from both language contexts and rater’s skills.

Table 1 shows that when only the agreement on breaks is concerned there is little difference between the K scores reached by the two groups of annotators, and only a minimal difference is observed when the text typology varies. In all cases K is over 0.8. The only exception (monologue in group 2 K= 0.79) have low relevance. As a whole, this confirms that the detection of prosodic boundaries is a primary evidence of language perception.

Table 1. *Kappa scores for prosodic break agreement.*

GROUP 1 (3 raters)							
	All positions			**Realistic**			**Agr. on break**
	overall	terminal	non-term.	overall	terminal	non-term.	overall
ALL	0.81	0.84	0.67	0.56	0.76	0.50	0.87
DIAL	0.82	0.88	0.62	0.57	0.79	0.49	0.87
MON	0.79	0.79	0.70	0.53	0.72	0.47	0.87
GROUP 2 (3 raters)							
	All positions			**Realistic**			**Agr. on break**
	overall	terminal	non-term.	overall	terminal	non-term.	overall
ALL	0.74	0.77	0.59	0.46	0.68	0.37	0.81
DIAL	0.76	0.79	0.57	0.47	0.65	0.40	0.85
MON	0.71	0.74	0.59	0.41	0.68	0.31	0.79
GROUPS 1 and 2 (6 raters)							
	All positions			**Realistic**			**Agr. on break**
	overall	terminal	non-term.	overall	terminal	non-term.	overall
ALL	0.76	0.79	0.60	0.54	0.69	0.49	0.83
DIAL	0.78	0.83	0.59	0.55	0.69	0.52	0.85
MON	0.73	0.74	0.61	0.49	0.65	0.43	0.82

Data regarding the perceptual relevance of the attribute “terminal” face to “non-terminal” deserves a more complex analysis. This statistics must be considered on the positions marked with a break by at least one annotator (realistic) rather than referring to all positions, considering the former the baseline that excludes the positions that are not reasonable candidates for a prosodic break.

Considering the two groups, the realistic k score on the inter-annotator agreement on terminal breaks reaches 0.67, which is significantly lower than the agreement achieved by Group 1 for terminal breaks (0.76 in realistic data and 0.84 considering all positions). Roughly speaking, this means that the perceptual relevance of the distinction between terminal and non-terminal prosodic breaks allows a sufficient level of break detection. However, this score is lower with respect to the expectations, considering the total agreement on terminal breaks reached in C-ORAL-ROM, and reflects interesting variation across annotators. More specifically, the group of more expert annotators gets significantly better results (0.76) with respect to the second group (0.68). No relevant variation emerges in monologues and dialogues.

The different scores correlate with the nature of terminal prosodic breaks. Their identification is a function of a judgment that requires direct perception but is not limited to it, since it goes in parallel with the accomplished of a speech act.

Data regarding the consensus on the annotation of non-terminal breaks are the more surprising. The realistic k turns out unsatisfactory for all groups and for all typologies (agreement on non-terminal positions). Although it is well known that a certain number of non-terminal breaks may be weak, this datum deserves a closer analysis since it departs from the previous results recorded for the recovery of non-terminal breaks in spoken Romance corpora.

3.2. Detailed Analysis

Considering the positions marked by at least one annotator with a terminal break, the total agreement (all annotators agree on the presence of a terminal break) is very relevant if the first group of annotators is considered (67.5%), while it is under 50% in the second group. This confirms that the recovery of terminated sequences varies in accordance with the skills developed by the annotators. However, this percentage varies more consistently in accordance to the language contexts. In the first group, the agreement on terminal breaks in dialogues corresponds to 74.8%, while in monologues, only 57.7% of the positions are of total convergence. This difference is even more pronounced in the second group, with 64.9% of total agreement in dialogues and only 28.1% in monologues.

Table 2. *Terminal position in %.*

Text and agreement type	Group 1		Group 2	
DIALOGUES	**freq.**	**%**	**freq.**	**%**
3 terminal breaks	238	74.8%	318	64.9%
2 terminal vs 1 non-terminal	39	12.3%	61	12.4%
1 terminal vs 2 non-terminal	32	10.1%	73	14.9%
2 terminal vs 1 no-break	6	1.9%	28	5.7%
1 terminal vs 2 no-breaks	3	0.9%	10	2.0%
Total	318	100.0%	490	100.0%
MONOLOGUES	**FREQ**	**%**	**FREQ**	**%**
3 terminal breaks	138	57.7%	111	28.1%
2 terminal vs 1 non-terminal	46	19.2%	36	9.1%
1 terminal vs 2 non-terminal	45	18.8%	145	36.7%
2 terminal vs no-break	4	1.7%	52	13.2%
1 terminal vs 2 no-breaks	6	2.5%	51	12.9%
Total	239	100.0%	395	100.0%
ALL TEXTS	**FREQ**	**%**	**FREQ**	**%**
3 terminal breaks	376	67.5%	429	48.5%
2 terminal vs 1 non-terminal	85	15.3%	97	11.0%
1 terminal vs 2 non-terminal	77	13.8%	218	24.6%
2 terminal vs no-break	10	1.8%	80	9.0%
1 terminal vs 2 no-breaks	9	1.6%	61	6.9%
Total	557	100.0%	885	100.0%

In summary, the differences in agreement on terminal break perception crosses the groups and reflects qualitative difference in monologues, where terminal breaks are less easy to be detected than in dialogic contexts even by expert annotators.

The consistency of perceptual relevance of terminal breaks, however, can be better evaluated considering the number of positions in which two annotators on three agree on the presence of a terminal break: 85% of convergent positions in dialogues and 77% in monologues in the first group, and 77% in dialogues and only 37% in monologues in the second group. These data show that the agreement on terminal break perception increases significantly also in monologues if expert annotators are in charge of the work.

Table 3. *Non terminal position in %*

Text and agreement type	Group 1		Group 2	
Dialogues	**freq.**	**%**	**freq.**	**%**
3 non-terminal breaks	98	35.1%	125	28.4%
2 non-terminal vs 1 terminal	32	11.5%	73	16.6%
1 non-terminal vs 2 terminal	39	14.0%	61	13.9%
2 non-terminal vs 1 no-break	48	17.2%	79	18.0%
1 non-terminal vs 2 no-breaks	62	22.2%	102	23.2%
Total	279	100.0%	440	100.0%
Monologues	**freq.**	**%**	**freq.**	**%**
3 non-terminal breaks	187	46.8%	171	34.5%
2 non-terminal vs 1 terminal	45	11.3%	145	29.3%
1 non-terminal vs 2 terminal	46	11.5%	36	7.3%
2 non-terminal vs 1 no-break	49	12.3%	69	13.9%
1 non-terminal vs 2 no-breaks	73	18.3%	74	14.9%
Total	400	100.0%	495	100.0%
All Texts	**freq,**	**%**	**freq.**	**%**
3 non-terminal breaks	285	42.0%	296	31.7%
2 non-terminal vs 1 terminal	77	11.3%	218	23.3%
1 non-terminal vs 2 terminal	85	12.5%	97	10.4%
2 non-terminal vs 1 no-break	97	14.3%	148	15.8%
1 non-terminal vs 2 no-breaks	135	19.9%	176	18.8%
Total	679	100.0%	935	100.0%

The fact that terminal breaks are prominent is reinforced by the result of only 1.6% positions in which one of the annotators signed a terminal break while the others do not perceive a break.

If the set of positions where at least one annotator marks one non-terminal break is considered, the data show opposite tendencies. No matter if expert or non expert annotators, or if they work on dialogues or monologues, in around 20% of these cases the other two annotators do not mark any break. Therefore, a good lot of non-terminal breaks are week. However, it must be taken into account that the number of strong non-terminal breaks in which all annotators agree is the double, and range as a whole from 42% to 31 % (respectively for experts and non experts). The number of convergent positions, where two annotators mark a non terminal break, is extremely similar for expert and non experts (47% and 44% in dialogues, and 58 % and 63 % in monologues).

With respect to these positions, we must consider that a lot of them (from 40% to 50%) produce strong tagging uncertainty. However, this uncertainty, roughly speaking, 3 cases out of 4, regards the presence of a break itself (/ -/- 0), rather than the presence of the attribute "terminal" (/ -/- //)

4. Types of disagreement positions in *week* non terminal breaks

***Discourse markers* and Co-articulated strong syntactic breaks.** Discourse markers are not syntactically bound to the contiguous words in the utterance. In Romance languages like Italian and in Spanish this syntactic break occurs, almost necessarily, in parallel to a non terminal prosodic break. This may be not always the case in BP. High frequency discourse markers like "aì" are very frequently coarticulated to the subsequent syllable, as in (2), where "aì" forms a rising diphthong with the following vowel (here and below relevant contexts are surround in a box in the WinPitchPro spectrogram and underlined in the text). Annotators may for this reason perceive or not a break after the discourse marker. This language specific lack of correspondence between syntactic boundary and prosodic boundary in high frequency discourse markers is probably one of the main reasons for the lower rate on non terminal breaks in CORAL-BRASIL.

(1)*REG: aí entrou todo mundo em pânico //
[then, everybody panicked //]
Alternative: aí / entrou todo mundo em pânico //

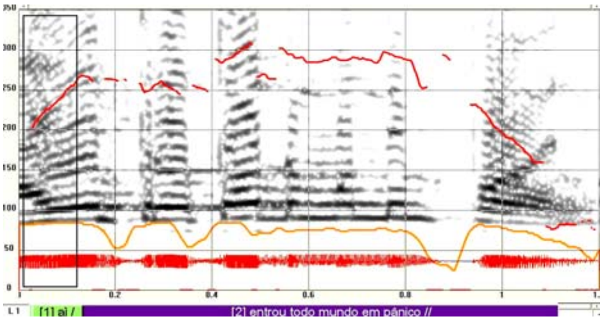

Figure 1: *Co-articulated Discourse marker in BP*

Pragmatic break in high speed co-articulated sequences. *Reported speech* like records high frequency in spontaneous narratives. (3) is an instance of two reported utterances within a reported discourse. Both the syntactic and the pragmatic levels find a boundary in the position into object, which marks the boundary between two utterances. As the spectrogram shows the break is neither segmental nor suprasegmental. As a matter of facts "ção – já" is a co-articulated sequence. Therefore the absence of an objective prosodic boundary face to a pragmatic and syntactic break leads to the underdeterminacy of perception. Such a case may occur in BP in which the high speed of speech allows a sequence of reported utterances to be performed without break.

(2) *REG: começou dar contração já vou te dar anestesia //
[you start having contractions / I'll give you anesthesia]
Alternative: começou dar contração / já vou te dar anestesia //

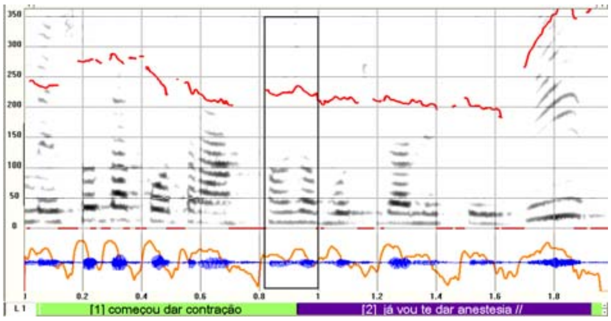

Figure 2: co-*articulated sequence of utterances in BP*

Syntactic constituents in a Rhythmic structure. Specifically in the rural variety, speech may present a strong accentual structure. This allows the performance of long sequences of phrases that are bound by syntactic relations, within a rhythmic sequence. For instance in (4) the sequence is characterized by high intensity and longer duration of the stressed syllables, but it is not scanned by perceptively relevant prosodic movements. In connection to the stressed syllable the end of a prosodic unit can be perceived or not. A large percentage of disagreement on non terminal breaks in BP depends on this language specific property.

(3) *REG: aí eu falei assim / que Roupa que cê tem da Bruna aí / qualquer coisa que cê tiver de neném lavada cê manda / porque minhas coisa tá tudo sem lavar e preciso de roupa // [then I said / do you have any of Bruna's clothes there / any clean baby clothes that you have you send me / because all my stuff is unwashed and I need clothes //]
Alternative: qualquer **coi**sa / que cê **ti**ver / de ne**ném** / lav**ada** / cê **ma**nda

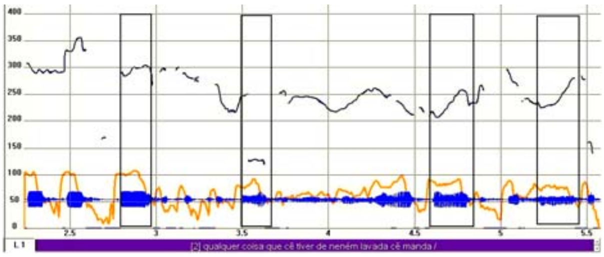

Figure 3: *Rhythmic Syntactic constituents in BP*

5. Conclusions

Although the overall perceptual prominence of prosodic breaks is confirmed in BP, the inter-rater agreement for terminal breaks turns out satisfactorily only for experts, so confirming that the notion of terminal breaks is not a function of a sole perceptual judgment, but it requires the access to linguistic information. The agreement decreases, both for experts and non experts, in non terminal positions. In BP this tendency, already recorded in the C-ORAL-ROM validation, is maximized. The work highlights the contexts of disagreement regarding non terminal breaks: a) contexts in which the annotators perceived a break within co-articulated sequences in connection to a strong syntactic break; b) prosodic reasons; i.e the accentual structure of some language varieties of BP, that make the assignment of non terminal breaks underdetermined by the sole perceptual prominence.

6. Acknowledgements

This paper received financial support from CNPq and FAPEMIG research grant programs and is written within the Agreement between the University of Florence and UMFG.

7. References

[1] Cresti, E, Moneglia M. (eds) (2005) Integrated Reference Corpora for Spoken Romance Languages. Benjamins:Amsterdam

[2] Danieli M, , Garrido J. M.; Moneglia M.; Panizza A, Quazza S., Swerts M. (2004) "Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech "C-ORAL-ROM" in M.T Lino, M.F. Xavier, F. Ferraira, R. Costa, R. Silva (eds) Prococeedings of the 4th LREC Conference, ELRA, Paris, vol. 4 pp. 1513-1516

[3] De Moraes, J. A. (1998). Intonation in Brazilian Portuguese. In D. Hirst, and A. Di Cristo (eds) Intonation Systems: A Survey on Twenty Languages. Cambridge: Cambridge University Press.

[4] Pitrelli, J. F., Beckman, M. E. & Hirschberg, J. (1994), "Evaluation of Prosodic Transcription Labeling Reliabilityin the ToBI Framework". In Proceedings of ICSLP,Yokohama: 123-126

[5] Grice, M., Reyelt, M., Benzmuller, R., Mayer, J., Batliner, A. (1996). "Consistency in Transcription and Labelling of German Intonation with GtoBI. In Proceedings of Int. Conf. On Spoken Language Processing, vol. 3 (pp. 1716--1719). Philadelphia.

[6] Syrdal, A. and McGory, J., (2000) "Inter-transcriber reliability of ToBI prosodic labeling," Proc. of the Intl. Conf. on Spoken Lang. Proc., Beijing: China, 235-238.

[7] Dilley, L., Breen, M., Gibson E., Bolivar, M., Kraemer, J. A (2000) "Comparison of Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices)" papers/2000_ICSLP

[8] Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M., (2004)"Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," Proc. of the Intl. Conf. on Spoken Lang. Proc., Nara: Japan, 2729-2732.

[9]Buhmann, J., Caspers, J., van Heuven, V., Hoekstra, H. Martens, J-P., Swerts, M., (2002). "Annotation of prominent words, prosodic boundaries and segmental lengthening by no-expert transcribers in the spoken Dutch corpus". In Proceedings of LREC 2002 (pp 779--785). Paris: ELRA.

[10] Mo, Y. Cole, J., Lee,E-g (2008): "Naive listeners prominence and boundary perception", In SP-2008, 735-738

[11] Du Bois, J. W., Schuetze-Coburn, S., Cumming, S. & Polino D. (1993), "Outline of discourse transcription". in Edwards J. A. & Lampert M. D. (Eds.). Talking Data: Transcription and Coding in Discourse Research, Hillsdale, New Jersey, Lawrence Erlbaum Associates, 45-89.

[12] Cheng, W. Greveaves, C. & Warren, M. (2005), "The creation of prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic)", International Computer Archive of Modern English (ICAME) Journal 29, 5-26.

[13] Isard, A., Carletta, J. (1995). "Replicability of transaction and action coding in the Map Task corpus". In J. Moore et al. (Eds.), Empirical Methods in Discourse Interpretation and Generation, Working Notes of the AAAI Spring Symposium Series, Stanford, Stanford University, 60--66.

[14]. Nesterenko, I.(2006): "Prosodic boundaries in spontaneous Russian: perceptual annotation and automatic classification", In SP-2006, paper 110.

[15] Cohen, J. A. 1960 . "A coefficient of agreement for nominal scales". Educational and Psychological Measurement 20: 37-46.

[16] Fleiss, J. L. 1971. Measuring nomial scale agreement among many raters. Psychological Bulletin. 76, 3780382.

[17] Amir N. Silbert-Varodz, V. & Izre'el, S. (2004), "Characteristics of intonation unit boundaries in spontaneous spoken Hebrew: Perception and acoustic correlates", SProSIG, 677-680.

[18] Izre'el S. (2005) Intonation Units and the Structure of Spontaneous Spoken Language: A View from Hebrew. In: Proceedings of the IDP05 on Discourse-Prosody Interfaces.