# Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech "C-ORAL-ROM"

**Morena Danieli[1]; Juan María Garrido[2]; Massimo Moneglia[3]; Andrea Panizza[1]; Silvia Quazza[1]; Marc Swerts[4]**

[1]LOQUENDO
Via Nole 55, 10149 Torino, Italy
Silvia.Quazza@LOQUENDO.COM

[2]Telefónica Investigación y Desarrollo
Ocata, 1
08003 Barcelona, Spain
jmgarri@tid.es

[3]LABLITA, Università di Firenze,
piazza Savonarola 1, 50132 Firenze, Italy
moneglia@unifi.it

[4]Tilburg University, Faculty of Arts
P.O. Box 90153 , 5000 LE Tilburg, The Netherlands
m.g.j.swerts@uvt.nl

## Abstract

C-ORAL-ROM, Integrated Reference Corpora For Spoken Romance Languages, is a multilingual corpus of spontaneous speech delivered within the IST Program. Corpora are tagged with respect to terminal and non terminal prosodic breaks. Terminal breaks are considered the most perceptively relevant cues to determine the utterance boundaries in spontaneous speech resources. The paper presents the evaluation of the inter-annotator agreement accomplished by an institution external to the consortium and shows the level of reliability of the tagging delivered and the annotation scheme adopted. The data show, at cross-linguistic level, a very high K coefficient (between 7.7 and 9.2, according to the language resource). A strong level of agreement specifically for terminal breaks has also been recorded. The data thus show that the annotation of the utterances identified in terms of their prosodic breaks is able to capture relevant perceptual facts, and it appears that the proposed coding scheme can be applied in a highly replicable way.

## 1. C-ORAL-ROM Prosodic Tagging

The C-ORAL-ROM Project (IST2000-26228) provides four multimedia corpora of spontaneous speech for French, Italian, Portuguese and Spanish, (Cresti and Moneglia, forthcoming). They consist of different speech materials, sampled following an explicit set of parameters (***formal*** versus ***informal***, ***dialogue*** versus ***monologue***; ***media*** versus ***natural context***), applied in such a way that the proportional representation of the different speech styles is equal in the four languages.

In C-ORAL-ROM the textual string is presented as short consecutive chunks of words in orthographic transcription, which are separated by prosodically motivated tags: these are based on the occurrence of terminal and non terminal breaks in the speech waveform. The presence of a terminal break is considered the main cue for the detection of *utterances* (Austin, 1962); that is, the linguistically relevant information units in the domain of spontaneous speech (Biber et al., 1999; Cresti, 2000). The motivation for this is that terminal breaks are assumed to mark the utterance limit (Karcevsky, 1931; Crystal, 1975).

The rough equivalence between utterance units and sequences separated by terminal breaks is based on the idea that competent speakers are extremely sensible to intentional prosodic variation ('t Hart et al., 1990) and that the voluntary accomplishment of a speech act is always accompanied by such variations.

The notion of *Prosodic break* is specified as follows:

**Concept:** *Prosodic break*
**Definition:** Perceptively relevant prosodic variation in the speech continuum such as to cause the parsing of the continuum into discrete prosodic units.
**Concept:** *Terminal prosodic breaks*
**Definition:** Given one or more prosodic units, a prosodic break is said terminal if a competent speaker assigns to it the quality of concluding the sequence.
**Concept:** *Non-terminal prosodic breaks*
**Definition:** Given a sequence of one or more prosodic units, a prosodic break is said *non-terminal* if a competent speaker assigns to it the quality of being non conclusive.

In C-ORAL-ROM each word boundary (W) is considered a possible position for a break (within-word breaks are not considered) and necessarily has one of the following values: 1) no break (O); 2) terminal break (T); non-terminal (N).

Tagging is based only on perceptual judgments and does not require any specific linguistic knowledge, although the notion of *speech act* is always familiar to the expert transcribers (PhD and PhD students) who annotated the corpus according to the following procedure: 1) Tagging of prosodic breaks simultaneous to the transcription by a first labeler; 2) Revision of tagging by a different labeler in connection to the revision of transcripts; 3) Revision of tagging and specific check of terminal breaks by a third labeler during the alignment.

This process already ensures control on the inter-annotator relevance of tags and a maximum accuracy in the detection of terminal breaks. The accuracy with respect to non-terminal breaks is by definition lower. The current paper will report on additional evaluations of the prosodic tags by another group of independent annotators.

### 1.1. Evaluation background

In the recent literature the evaluation of inter-annotator agreement with respect to various kinds of prosodic boundaries, mainly regards ToBI annotation (by trained labelers) of mostly non-spontaneous speech resources (Pitrelli, Beckman, Hirschberg, 1994; Grice et al., 1996; Syrdal & Mc Gorg 2000). The prosodic annotation in the Dutch corpus of spontaneous speech has also been recently verified by non-experts, who had received a minimal amount of training (Buhmann et al., 2002).

Roughly speaking such literature testifies a high degree of agreement on "boundary tones" (from 85 to 92%; for

ToBI annotation). In the Dutch corpus, a "substantial consistency" has been reported on the annotation of strong and weak prosodic breaks (K-coefficient between 0.61 and 0.80).

Although the prosodic labeling of the Dutch corpus is close to that of C-ORAL-ROM with respect to the nature of the resource (spontaneous speech) and the annotation unit (prosodic break), no specific test has been performed on the distinction between *terminal* and *non terminal breaks*. The annotation of breaks in the Dutch corpus may partially overlap those reported in C-ORAL-ROM, but it is not co-extensive. *Strong breaks* are defined as "severe interruptions of the normal flow of speech", while *Weak breaks* are defined as ""weak" but still clearly audible interruptions of the speech flow". It is very likely that all terminal breaks are perceived as severe interruptions of the speech flow, but also a remarkable number of non terminal breaks share this property. In other words a strong break may not have the functional value of terminal breaks (end of the utterance) and therefore have lower linguistic relevance.

### 1.2. Goals of the evaluation

The evaluation aims to test the hypothesis that prosodic breaks, especial terminal ones, have strong perceptual prominence and can be object of a reliable tagging in spontaneous speech corpora.[1] In parallel the evaluation will assess the reliability of the prosodic tagging of the C-ORAL–ROM speech corpora and the perceptual relevance of the coding scheme adopted in the Project when applied to different languages. Given the multilingual nature of the resource, this hypothesis can be tested at a cross-linguistic level, verifying whether language specific features may lead to differences in perceptual relevance.

## 2. Experimental Setting

Given the size of the resource (roughly 30 hours of speech for each of the four resources) the evaluation was performed only on a statistically significant portion. From each language corpus a subset was extracted, amounting to roughly 1/30 of its utterances (about 1300 utterances and around 1:30 hours of speech). The speech sections to be evaluated were automatically and randomly selected though ensuring the same distribution of speech types as in the overall corpus. Semantic and contextual coherence of the speech sections to be evaluated was guaranteed by choosing continuous series of utterances.

Two naïve mother tongue evaluators for each of the four languages were asked to evaluate the prosodic annotation of their language.[2] For each selected speech section, the procedure outputs an XML file ensuring text-audio alignment and a text file where each tagged utterance is reported twice (*validation copy*).

Each evaluator, independently of the other, had to examine the original annotation and possibly correct.

The evaluators received a two-days training. At the end of the training, a test was performed in order to assess the acquired competence of the evaluators and to ensure consistency between them in the evaluation[3].

The task was performed on Personal PC's, with the help of the speech software Winpitch, which allowed viewing the annotated text to be evaluated and listening to the corresponding aligned audio signal. The evaluator considered the existence of prosodic breaks at each word boundary position. If this perception did not match with the original tagging, he or she could modify the validation copy by inserting, deleting or substituting break marks. They were allowed to omit the evaluation of strings not clearly perceived, but this only occurred in a few cases

Each evaluator worked independently of the others and spent around sixty hours to accomplish his/her task, in four-hours daily sessions. None of the eight evaluators reported any difficulties in the evaluation and all of them could easily accomplish their task.

It is important to notice, however, that the 'exact replicability' of the scheme has not been tested during this evaluation, because the task proposed to the evaluators was not exactly the same as the one carried out by the expert annotators (non-experts had only to check, and not to annotate by themselves the speech material). For this reason the word 'evaluation' has been preferred to 'validation'. As a matter of fact, while non experts are the best candidates to test the perceptual prominence of a given cue, spoken language transcription and annotation cannot be easily replied by non-experts.[4]

## 3. Measures and Statistics

The evaluation data were then gathered and statistically analyzed in order to measure the degree of consensus expressed by the evaluators towards the original annotation.[5] The cases of disagreement, where one or both evaluators corrected the original annotation, were compared with the total number of word boundaries and, more perspicuously, with the number of positions, which are reasonable candidates for a break (*base-line*).

A replicability statistics has been applied to compare the three annotations obtained by C-ORAL ROM and by the two evaluators. Following previous work on annotation coding for discourse and dialogue, the Kappa coefficient was calculated (Isard and Carletta, 1995).

### 3.1. Evaluation data

Word boundaries W (positions candidate 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 N); *terminal break* (tagged as T)**.** Each position in the evaluation file receives a tag expressing the agreement with the original annotation. Given that each mismatch between the original annotation and the evaluations is not equally critical, we have ordered disagreements according to their importance.

[1] The project reviewers J. Moore and L. Ten Bosh suggested the evaluation of prosodic tagging by independent users. Loquendo (Turin) accomplished the evaluation. Thanks to Marco Fabbri and Enrico Zovato for data sampling and computation.

[2] Evaluators were chosen, with medium cultural level and no specific expertise in phonetics and prosody.

[3] The sampling selection procedure, the criteria for the selection of evaluators, the training material, samples of the evaluation file and a detailed evaluation procedure are accessible on the net at http://lablita .dit.unifi.it/coralrom/loquendo

[4] Moneglia et alii (2002) for internal evaluation with experts.

[5] Precision and Recall indexes was discarded. In C-ORAL-ROM evaluation, neither the original annotation nor the evaluators' choices could be taken as a correct reference.

| Tag Semantics | Degree |
|---|---|
| agreement on non-break | Ok |
| agreement on non-terminal | Ok |
| agreement on terminal | Ok |
| non-terminal insertion | non critical |
| non-terminal misplacement | non critical |
| non-terminal deletion | non critical |
| non terminal substitution (N-T) | Critical |
| terminal substitution(T->N) | Critical |
| terminal insertion | Very critical |
| terminal deletion | Very critical |
| terminal misplacement | Very critical |

### 3.2. Binary and ternary comparison files

Starting from an evaluation file, which reports both the original tagging in the C-ORAL ROM corpus (C) and the evaluator choice (E), a first parser generates a comparison file (B) where each word boundary (candidate position for a break) is represented as a record containing information that allows to compare the original tagging and the evaluator's choice[6]. Second, starting from the two comparison files B-E1 and B-E2 for the two evaluators E1 and E2, a new file (T) is generated, where each word boundary is represented. This file allows to compare the original tagging, the choices of both evaluators and the choice of both evaluators between them.

## 4. Evaluation Results

The results are given separately for each language evaluation sub-corpus and for its relevant subsets, i.e. for its main bipartitions into Formal and Informal speech and for its two subsets of Dialogues and Monologues. In the following only the results on total sub-corpora are reported in a summary below. The results for each sub-nodes are detailed only for certain specific commentaries.[7]

The percentage of word boundaries that received a break tag in the C-ORAL-ROM annotation, is the following in the four corpora:

|  | French | Italian | Portuguese | Spanish |
|---|---|---|---|---|
| Word boundaries marked with a prosodic break | 19% | 35% | 32% | 31% |

The difference in marking of the French corpus is relevant, and is reflected by a higher mid-length of tone units and by a higher mid-length of utterances in the whole corpus (Moneglia, in this volume)

Looking at the Binary Comparison statistics on the evaluation data, the evaluators confirmed virtually all Terminal Breaks in C-ORAL-ROM. The percentage of T-breaks that were not deleted by the evaluators is 100% (with the single exception of the Formal section of the Spanish corpus, where it is around 98%). This means that where the original annotator perceived a terminal break, the evaluators perceived a break too, or at least a non-terminal one (**Generic Confirmation**).

In other words terminal breaks represent strong linguistic cues and few doubts can arise on the existence of a prosodic break when it is judged terminal. Also the absence of *misplacement* with respect to terminal position confirms this interpretation.

The evaluator's perception of a non-terminal break in 0 position is also rare, but the difference with non-terminal turns is evident. Even if non-terminal breaks are confirmed in most cases, the probability of a lower perceptual relevance with non-terminal is shown by the existence of both misplacement and non-terminal deletion, recorded by all evaluators, from around 0.5% till 7 %.

The percentage of **Specific Confirmations**, where the evaluators confirmed that the break was indeed a T-break, is in most cases above 95%. On the contrary, the substitution of a terminal break with a non-terminal one (around 5% of cases) is an event that may occur, even if the actual incidence is very low. In other words the quality of being terminal is highly prominent and easy to recognize. On the other hand, the evaluators seldom perceived T-breaks where the original annotator did not perceive any kind of break, as shown by the **Terminal Missing** percentages, which are close to 0%. This confirms the accuracy of the original annotation, at least with respect to the detection of terminal prosodic breaks.

For what concerns the reliability of terminal annotation, the evaluation focused on **Terminal Substitution** (when the evaluator replaces a terminal with a non-terminal) and on **Terminal Adding** (when an evaluator replaces a non-terminal with a terminal). Terminal substitution involves around 5% of cases, while values regarding terminal adding record a significant score only in the case of the French corpus, while for the other languages, the percentages are mostly below 1%. In some cases the French evaluators perceived a stronger break where the annotator marked a non-terminal break; the percentage of Added Terminals, i.e. N-breaks substituted with T-breaks, ranges from 1.29% to 6.51% of the original N-breaks. C-ORAL-ROM ensures the consistency of terminals tags.

The reliability of non-terminal is not the primary objective of the evaluation. However the operations performed by evaluators on non-terminal breaks are, as a whole, much more than the corresponding operations on terminal breaks. Non-expert evaluators show a great sensibility towards prosodic parsing even in the weakest positions, thus strengthening the reliability of their low reaction to original terminal breaks.

As for Ternary Comparisons, which give a measure of the inter-annotator agreement and of the reliability of the C-ORAL-ROM prosodic tagging, we can see that the original annotation is basically confirmed, especially for terminal breaks: the percentages of T-breaks specifically confirmed by both evaluators are above 94% for all languages (Total agreement on T)

In general, K coefficient is slightly lower in Formal than in Informal speech, and in Monologues vs. Dialogues

| Kappa Coefficient (realistic) | | | | |
|---|---|---|---|---|
|  | French | Italian | Portuguese | Spanish |
| **Total** | **0,766** | **0,807** | **0,920** | **0,827** |
| **Formal** | 0,765 | 0,785 | 0,893 | 0,772 |
| **Informal** | 0,767 | 0,826 | 0,946 | 0,885 |
| **Dialogues** | 0,790 | 0,839 | 0,921 | 0,880 |
| **Monologues** | 0,675 | 0,779 | 0,944 | 0.818 |

This tendency is systematic. In face-to-face dialogues the linguistic units of references are brief strings, each one

[6] A few percentage of positions where the speech material was not perceived by some evaluators has been excluded

[7] Data available at http://lablita.dit.unifi.it/coralrom/loquendo/

matching with a speech act, and always ending with a terminal break. Therefore, the judgment that each sequence is "concluded" is relevant at the semantic level, the action level and the prosodic level alike.

This may not be frequently the case when spoken language performances feature long textual strings, as in formal monologues. In this case, although the terminal break always ensures that what follows belongs to a different linguistic domain, the string may include various linguistic domains, gathered within the same prosodically concluded structure. Judgment about the terminal or non-terminal nature of the prosodic breaks may be less certain.

Taking as a reference the whole set of evaluated word boundaries, the most general measure of agreement is the Total Agreement Rate (percentage of boundaries on which both evaluators agree).The highest consensus is expressed on the Portuguese corpus, but the values are very close for all languages, ranging from 95% to 98.9%.

The percentage of totally agreed word boundaries may sound too optimistic, due to the disproportion between word boundaries and actual candidates for a break (around 30% of the total). The total agreement is however significant when compared with a "baseline" that may be considered the worst possible realistic result, obtained in case all N's and T's were deleted and a comparable number of N's and T's were inserted in different positions.

| | **French** | **Italian** | **Portuguese** | **Spanish** |
|---|---|---|---|---|
| **Total Agreement Rate** | 96,48% | 95,21% | 98,93% | 97,17% |
| **Worst possible Result (baseline)** | 62,17% | 30,84% | 37,28% | 37,93% |

**Consensus in disagreement**, when both evaluators disagreed with the original tagging, is a very marginal phenomenon in the evaluation. The percentages of **globally disagreed positions** (when at least one evaluators modified C-ORAL-ROM) that were actually strongly disagreed range from 9% to 23.5%.

Finally, K coefficient measures the reliability of the annotation scheme, that is the probability to obtain the same annotation by different evaluators. Two K coefficients have been calculated: a general one, comparing the three categories of boundaries (T, N, O); and a more realistic coefficient, limiting the analysis to the two break tags T and N, in order to avoid the positive effect of the high agreement rate on no-break boundaries. Both coefficients are largely above the 0.6 minimal threshold, confirming the reliability of C-ORAL-ROM.

| | French | Italian | Portuguese | Spanish |
|---|---|---|---|---|
| Total w. boundaries | **12893** | **10925** | **12958** | **11512** |
| ***Binary Comparisons*** | | | | |
| | French | Italian | Portuguese | Spanish |
| Spe.Confirmation T E 1 | 96,12% | 98,8% | 98,7% | 94,1% |
| Spe.Confirmation T E 2 | 100% | 97,12% | 99,4% | 99% |
| Gen. Confirmation T E1 | 100% | 99,9% | 100% | 99,8% |
| Gen. Confirmation T E 2 | 100% | 100% | 100% | 99,7% |
| T Missing E 1 | 0,01% | 0% | 0,03% | 0% |
| T Missing E2 | 0,02% | 0% | 0,04% | 0,02% |
| N Missing E 1 | 1,59% | 1,05% | 0,62% | 1% |
| N Missing E 2 | 0,46% | 2,75% | 0,2% | 0,7% |
| Added Terminal E 1 | 2,95% | 0,2% | 0,5% | 1,57% |
| Added Terminal E2 | 5,01% | 0.16% | 0,4% | 0,12% |
| Misplacement E1 | 0,19% | 5% | 0% | 1,19% |
| Misplacement E 2 | 0,14% | 0,98% | 0% | 0,65% |
| ***Ternary Comparisons*** | | | | |
| | French | Italian | Portuguese | Spanish |
| Partial Consensus on T 0T vs. 3d or 2ts | 4,36% | 2,42% | 1,51% | 5,16% |
| Partial Consensus on T 0T vs. 3d | 0% | 0% | 0% | 0% |
| Partial Consensus on W | 3,18% | 3,65% | 0,93% | 2,56% |
| Total Agreement on T | 95,05% | 97,14% | 98,12% | 94,84% |
| Total Agreement on N | 86,56% | 93,15% | 98,38% | 94,62% |
| Total Agreement on O | 97,54% | 95,1% | 99,22% | 98,28% |
| Global Disagreement | 3,52% | 4,78% | 1,57% | 2,83% |
| Consensus Disagreement | 8,97% | 23,50% | 12,12% | 9,26% |
| K Index (General) | 0,952 | 0,928 | 0,980 | 0,946 |
| K Index (Realistic) | 0,776 | 0,807 | 0,920 | 0,827 |

## 5. References

Austin, L.J. (1962). How to do things with words. Oxford: Oxford University Press.

Biber D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (Eds.) (1999). The Longman grammar of spoken and written English. London: Longman.

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.

Crystal, D. (1975). The English tone of voice. London: Edward Arnold.

Cresti, E. (2000). Corpus di italiano parlato, vol. I- II, CD-Rom, Firenze: Accademia della Crusca.

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.

't Hart, H., Collier, R., Cohen, A. (1990). A perceptual study on intonation. An experimental approach to speech melody. Cambridge: CUP.

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 University (pp. 60--66). Stanford, Ca.

Karcevsky, S. (1931). Sur la phonologie de la phrase. In Travaux du Cercle Linguistique de Prague, IV.

Moneglia, M. (in this volume). Measurements of spoken language variability in a multilingual corpus. Predictable aspects.

Moneglia, M., Scarano, A, Spinu, M. (2002) Validation by expert transcribers of the C-ORAL-ROM prosodic tagging criteria on Italian, Spanish, Portuguese corpora of spontaneous speech. http://lablita.dit.unifi.it/coralrom

Pitrelli, J., Beckman, M., Hirschberg, J. (1994). Evaluation of Prosodic transcription Labelling Reliability in the ToBI framework. In Proceedings of Int. Conf. On Spoken Language Processing, Yokohama, Vol, 2 (pp. 123-126).

Syrdal, A., Mc Gorg, J. (2000). Inter-Transcriber Reliability of ToBI prosodic labelling. In Proceedings of Int. Conf. On Spoken Language Processing, Beijing, China, vol. 3 (pp.235-238).

Winpitch http://www.winpitch.com