Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labelling system

David Escudero[a], Lourdes Aguilar[b], Maria del Mar Vanrell[c,e], Pilar Prieto[d]

[a]*Dpt. of Computer Sciences, Universidad de Valladolid, Spain*
[b]*Dpt. of Spanish Philology, Universitat Autònoma de Barcelona, Spain*
[c]*Dpt. of Catalan Philology, Universitat Autònoma de Barcelona, Spain*
[d]*Dpt. of Translation and Language Sciences, ICREA-Universitat Pompeu Fabra, Spain*
[e]*Dpt. of Translation and Language Sciences, Universitat Pompeu Fabra, Spain*

Abstract

A set of tools to analyze inconsistencies observed in a Cat_ToBI labelling experiment are presented. We formalize and use the metrics that are commonly used in inconsistency tests. The metrics are systematically applied to analyze the robustness of every symbol and every pair of transcribers. The results reveal agreement rates for this study that are comparable to previous ToBI inter-reliability tests. The inter-transcriber confusion rates are transformed into distance matrices to use multidimensional scaling for visualizing the confusion between the different ToBI symbols and the disagreement between the raters. Potential different labelling criteria are identified and subsets of symbols that are candidates to be fused are proposed.

Keywords: Prosody, Prosodic labeling, Inter-transcriber consistency, ToBI

1. Introduction

The framework of intonational phonology, also known as the Autosegmental-Metrical (AM) model of intonation, has been applied to many languages, to thoroughly describe prosodic systems and develop methods of collecting intonation data [1]. This framework has also been applied in automatic speech processing and database annotation to yield ToBI (TOnes and Break Indices) a prosodic labelling standard for speech databases that is based on Pierrehumbert's thesis [2]. ToBI-based systems have been developed to label oral databases for many languages such as English [3], Spanish [4, 5, 6], German [7], Japanese [8], Greek [9], Korean [10] and Catalan [11, 12] among others.

It is important to make clear that, as the developers of ToBI explicitly state, ToBI is not an International Phonetic Alphabet for prosody. Because intonation and prosodic organization differ from language to language, and often from dialect to dialect within a language, there are many different ToBI systems, each one specific to a language variety and the community of researchers working on that language variety [13]. From this point of view, a full intonational and prosodic description of a given language is needed before a ToBI-based transcription system is accepted as a community-wide standard.

The ToBI system consists of annotations at several time-linked levels of analysis. The three obligatory tiers are: an orthographic tier, of time-aligned words; a break index tier, which indicates the degree of junction between words; and a tonal tier, where pitch accents, phrase accents

and boundary tones define intonational events. A fourth tier, the miscellaneous tier, is provided to annotate any additional phenomena, such as disfluencies.

One of the advantages of using the ToBI systems for prosodic annotation is its reliable inter-transcriber consistency (see the favourable inter-transcriber reliability scores for the different systems in section 5.1) due to the relatively simple labelling procedure proposed. Moreover, the ToBI systems presented for each language are generally based on and directly linked to fundamental research on prosody for each language. Yet despite the widespread use of the ToBI system, it also has its detractors [14, 15, 16, 17], in particular, because of the confusions have arisen either in the tagging process, when more than one transcriber must label the same utterances, or when labelling is done automatically, since in the automatic labelling process, some of the points where ToBI markers need to be placed are not easily identifiable from the acoustic signal [18, 19, 20].

In phonologically-oriented prosodic transcribing systems, like ToBI, intercoder inconsistencies appear because the labelling process depends on perceptual criteria that are mainly dependent on the subjective human judges. Our point of view is that inconsistencies are due to the non-uniform acoustic expression of prosody and are inevitable. However, they represent a challenge for the development of prosodic speech synthesis and recognition systems across languages, as well as automatic prosodic labelling systems.

This paper has two goals. First, to run an inter-transcriber consistency test for Catalan speech data annotated with the Catalan-adapted version of ToBI. Catalan has been intensively analyzed from a prosodic point of view and a full-fledged ToBI annotation proposal (Cat_ToBI) has been in place for some time now Prieto2012,Prieto2009,CatToBI. It is therefore of considerable interest to subject Cat_ToBI to an inter-rater consistency test at this point. To this end, ten transcribers labelled prosodic events independently on a Catalan corpus of twenty sentences from four different speech styles using the most recent version of the Cat_ToBI system. The twenty sentences were extracts from recordings of a variety of discourse types, including spontaneous speech. Though favourable inter-transcriber reliability results have been reported for ToBI-labelled corpora of mainly read speech produced in a laboratory setting, fewer inter-transcriber reliability studies have been carried out for spontaneous speech (e.g., [21]).

The second goal of this paper is to propose a low-cost procedure to automatically obtain three types of important information from an inter-transcriber consistency experimental test: (a) the most confusable symbols from experimental data; (b) the types of errors most commonly produced by labellers; (c) signs of insufficient pre-training in individual labellers. As is well known, the selection of skilled, experienced transcribers is crucial for producing a large database that is consistently and thus usefully labelled. The aim of the *Glissando Project*, which is one of the sponsors of this research (see section 8) is to do precisely that, i.e. to compile a Spanish/Catalan prosodic corpus enriched with ToBI labels, and it was regarded as essential to be able to carry out thes three tests before starting such a large-scale labelling process. It was assumed that the labels introduced by an unskilled labeller would differ significantly from the labels introduced by a proficient labeller, and consequently the consistency of the final corpus would be poor. In this paper, we review and formalize the commonly used metrics for measuring inter-transcriber consistency, and we use multidimensional scaling to easily discriminate proficient transcribers from those that are not. Furthermore, we propose a procedure to diagnose the common mistakes of the inexpert labeller in order to advise him/her in a potential retraining process.

That said, when a transcribing system is still undergoing development, the withdrawal of unskilled labellers may not be enough to increase consistency rates. This is because, as we will see in this paper, even taggers who are regarded as experts can exhibit low inter-labeller consistency

rates when they label the same set of sentences. The reason for this is that they apparently use different tagging criteria for some of the ToBI symbols. We will present a procedure for analyzing inconsistencies that permits these situations to be pinpointed by identifying the problematic symbols that cause these conflicts. This analysis will have an impact on the evaluation of the ToBI system in itself.

Another source of inconsistencies is the existence of pairs of tags, or sequences of tags that are commonly confused by the labellers because of their high perceptual or acoustic similarity. In [17] a set of transcribers are questioned about the inter-similarity of the various ToBI labels. Their answers show that, for example, they find the pair *H** and *L+H** the most difficult pair of symbols to separate. The identification of other easily confused labels suggest that it might be advisable to build alternative reduced versions of the prosodic set of labels. In fact, a reduction in the number of ToBI symbols has already been shown to be effective for not only speeding up the manual labelling process [22] but also increasing the automatic classification rates [23, 24, 25].

Thus, the overarching aim of this article is to present a language-independent procedure that will allow the inter-transcriber inconsistency to be computed and visualized when while a prosodic corpus is being labelled in order to easily identify, on the one hand, misuses of the conventions by taggers, and on the other hand, the most confusable symbols.

The paper is organized as follows: the database is presented in section 2 including a review of the Cat_ToBI system; next the experimental procedure is described with the report of the metrics (section 3) and the visualization techniques (section 4) that have been used to present the results that are reported in section 5. We conclude with a discussion of the results and suggestions for future work in sections 6 and 7.

2. Methods

This section consists of a description of the speech database to which the analysis tools were applied. The prosodic events were annotated within the Cat_ToBI framework.

2.1. Corpus

Twenty Central Catalan target utterances were selected from different corpora so that they represented the following four different discourse types:

1. Spontaneous speech excerpted from the guided interview subcorpus of the *Atles interactiu de l'entonació del català* [26],
2. Spontaneous speech excerpted from the Map Task subcorpus of the *Atles interactiu de l'entonació del català* [26],
3. Radio news,
4. Text reading (from the Festcat database[27]).

The full set of sentences in Catalan, together with their English translation can be found in the Appendix A. Nine out of the twenty utterances are yes-no questions or wh-questions, four are narrow focus statements and the rest are broad focus statements. In total, the sentences contained 264 words. The duration of the 20 files is 89.8 seconds. The speech sources were 12 native speakers of Central Catalan (5 males and 7 females).

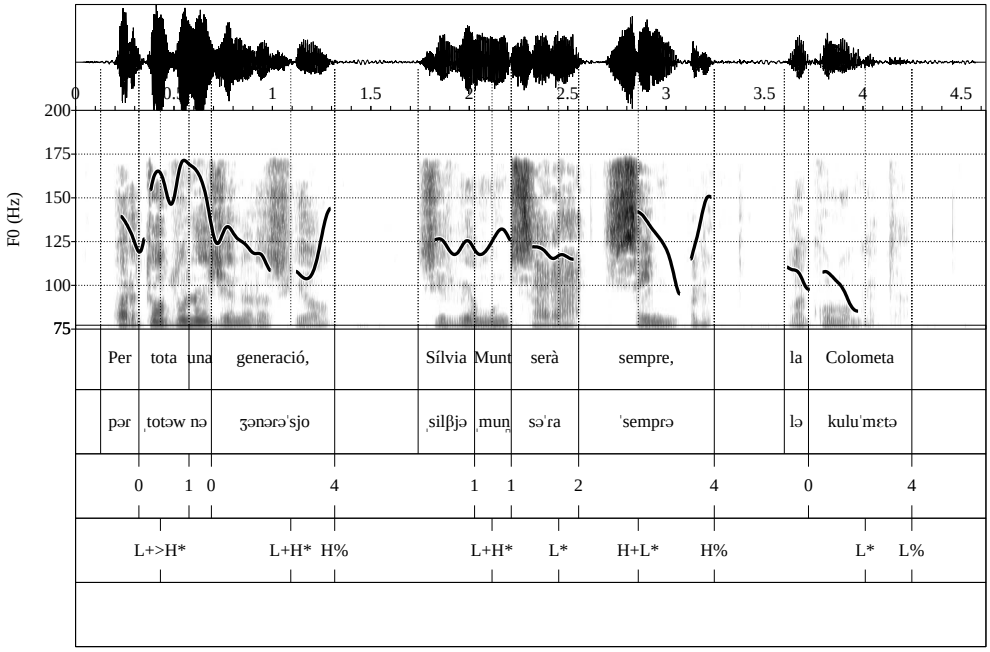


Figure 1: Example of an utterance in Catalan with audio signal analyzed using Praat to show formant frequency (top tier) and wave form (second highest tier), followed by orthographic and phonological transcriptions (middle two tiers), breaks and tones annotation (bottom tiers).

2.2. Labellers

A total of ten labellers participated in the labelling were ask to independently label audio files of the same twenty utterances. In terms of degree of prior experience with prosody and Cat_ToBI some of the labellers were absolute beginners while others had actually contributed to the development of Cat_ToBI and were fully comfortable with it. The labellers were divided into three groups: Group 1 (Experts), Group 2 (Familiar with prosodic annotation systems), and Group 3 (Beginners, completely new to any model of intonation or prosodic transcription). Group 1 comprised four labellers and Groups 2 and 3 had three labellers each. All labellers were native speakers of Catalan, with two dialects represented (Central Catalan and Balearic Catalan).

2.3. Transcription procedure

Following general ToBI conventions, transcribers had to perform the following tasks:

1. Mark any syllables which carry a clear prominence, that is, decide if there is a pitch accent.
2. If there is a pitch accent, decide the pitch accent type.
3. Mark different degrees of the strength of the boundary between two orthographic words, that is, decide the break index.

4. Decide the boundary tone type, according to the degree of prosodic breaking (intermediate phrase-ip vs. intonational phrase-IP).

Each transcriber was provided with a document describing the Cat_ToBI system [11] as well as Cat_ToBI training materials [28]. The training materials contain a tutorial explaining each of the labels used in Cat_ToBI, along with recorded examples of transcribed utterances. There are also exercises to practice assigning the labels described in the text. These materials are designed to be self-explanatory. Moreover, absolute beginners attended a course (three sessions of three hours each) on the basics of the AM model and the ToBI labelling systems taught by the last author of the article.

Manual annotation was performed using the Praat tool [29]. The starting point was a *TextGrid file* [29] for each sentence with its orthographic and phonetic transcription. The transcribers studied the visual display on a computer monitor of the audio signal (F0 curve and waveform) and then used that visual information to make labelling decisions about prosodic features. The key elements to be labelled were prominence, prosodic boundary strength and pitch accent and boundary tone types. An example showing audio signal information and orthographic and phonetic transcriptions with Cat_ToBI annotation added in the lowest tier is provided in Figure 1).

The utterances selected had not previously been labelled by any of the participating transcribers in the course of earlier research; each transcriber worked alone on the twenty utterances in the experimental data-set and they were not allowed to discuss these utterances with any other transcriber or researcher. After all ten transcribers had completed the transcription, their *Textgrid files* were collected and statistics for inter-labeller agreement were computed from the data, as will be explained in the following sections.

2.4. The Cat_ToBI system

The description of Catalan prosodic organization and intonation presented here is based on early work on Catalan within the framework of intonational phonology [11, 30]. The most updated description of the Cat_ToBI proposal may be found in [11] and on the *Cat_ToBI Training Materials* website [28]. As in other languages analyzed within the ToBI framework, Catalan intonational events are of two types, namely pitch accents (or pitch movements that are associated with metrically strong positions), and boundary tones (or tones that are anchored to phrase edges). The phrases that are marked by the placement of these boundary tones are an important component of the metrical structure in the language.

As far as prosodic organization is concerned, Cat_ToBI proposes to analyze the Catalan data as having four levels of phrasing: the prosodic word, the phonological phrase, the intermediate phrase (ip) and the intonational phrase (IP). Evidence in support of the prosodic word, the intermediate phrase and the intonational phrase are described in [11, 28], where it is also acknowledged that the existence in Catalan of the phonological phrase is an unresolved issue. According to this description, in Cat_ToBI, five levels are included in the break-index tier: Break 0, to mark cohesion between orthographic forms; *B1*, to mark boundaries between prosodic words; Break 2, to mark a level of phrasing below the intermediate phrase; Break 3, to mark the boundaries of intermediate phrases; and Break 4, to mark the boundaries of intonational phrases.

For the intonational analysis of Catalan utterances, in [11, 28] two types of tonal events are recognized: i) pitch accents, or local tonal events which are associated with metrically strong syllables and which confer accentual prominence to these syllables; and ii) boundary tones, or tonal events associated with the boundaries of prosodic domains, at both the right edge of intermediate phrases and the right edge of intonational phrases. It should be noted here that

some authors have argued that the phrase accent category can be dispensed with, and that only one type of boundary tone is needed.

According to this, Catalan has six basic pitch accents H*, L+H*, L+>H*, L*, L*+H and H+L*, with the following upstepped and downstepped pitch counterparts (i.e., scaled higher or lower than the previous pitch accent): !H*, H*, L+!H*, L+¡H* and !H+L*. With respect to the use of the symbol '>', the same convention used in MAE-ToBI [3] and in Gr_ToBI [9] is adopted: if the maximum F0 peak does not actually occur within the syllable nucleus, the late F0 event is marked by putting the symbol '>' before the H.

With respect to boundary tones, the following boundary tones and phrase accents have been attested (with the inventory of boundary tones differ as a function of its position in the prosodic hierarchy i.e., end of IP, end of ip, beginning of IP).

- 8 types of boundary tones at the end of IPs (marked with the % symbol after the tone): L%, M%, H%, HH%, HL%, LH%, LM%, LHL%
- 5 types of boundary tones at the end of ips (marked with the - symbol after the tone): L-, M-, H-, HH-, LH-
- 1 type of initial boundary tone (marked with the % symbol before the tone): %H.

For our analysis of inter-transcriber reliability, we distinguished a total of 7 distinct pitch accent categories. We decided to exclusively analyze the phonological identity of distinct pitch accent types, and upstep and downstep marks were disregarded in the analysis. Similarly, the distinction between ip and IP levels of phrasing was collapsed.

3. Measuring the inter-transcriber agreement

In a labelling process, inter-transcriber reliability quantifies the degree of agreement among labellers by giving a numerical score of how much consensus there is in the labels assigned by transcribers [**?**]. The measurements of inter-transcriber consistency used in this study will follow closely the ones used in previous prosodic tests to facilitate comparisons between studies. The ToBI labels are treated as categorical data so that the most commonly used metrics are joint agreement, kappa statistics and pairwise transcriber agreement, which are presented in the following sections.

3.1. Formulation

Let us refer to the prosodic event to be labeled by E_i, with $i = 1..e$. Likewise, let us refer to the transcribers or labellers that participate in the tagging process by T_j, with $j = 1..t$. Finally, let us refer to the categories into which assignments are placed by C_k, with $k = 1..c$, i.e.the number of marks that can potentially be used. $C_{i,j} \in C_k$ will be the category assigned by the labeller j to any event i.

3.2. Joint agreement

The joint agreement is the number of times each rating (i.e. the label C_k) is assigned by each labeller, divided by the total number of ratings [31]. Let n_{ik} represent the number of raters who assigned the i^{th} subject to the k^{th} category. By computing the n_{ik} values for every i and displaying this information for a given k the distribution of the quantity of agreement associated with each symbol, $f(n_k)$, can be visualized.

For a given k, the distribution of frequency $f(n_k)$ has a domain of values that goes from 1 to t (the 0 value is ignored as it represents events where none of the t raters assigned the symbol k). In the extreme case in which the t raters agree every time the symbol k appears, the mode of $f(n_k)$ would be t. Thus, a right mono-lobulated distribution indicates a high agreement as most raters agree when they label the category k. Thus the closer the mode of the distribution is to t, the greater the consensus.

On the other hand the closer the mode of the distribution $f(n_k)$ is to 1, the more problematic the symbol. In an extreme case, every time the symbol k appears, only one of the t raters would mark it. Thus a left mono-lobulated distribution is evidence of low agreement, since the labellers have used this symbol rarely and without consistency. Furthermore, bilobulated distributions and flat distributions indicate a high confusion with potential wrong tagging criteria.

We have found no reference to ToBI labelling consistency tests in which this metric had been used. Results in this paper show the usefulness of this joint frequency test to evaluate differing degrees of consensus in the assignation of different labels.

3.3. Kappa statistics

Fleiss' kappa [32] is a generalization of Scott's pi statistic[33], a statistical measure of inter-rater reliability. It is also related to Cohen's kappa statistic[34]. Whereas Scott's pi and Cohen's kappa work for only two raters, Fleiss' kappa works for any number of raters, giving categorical ratings to a fixed number of items. The kappa indices are referenced with the greek letter κ.

The κ index expresses the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely randomly. The κ index is computed by means of the formula:

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{1}$$

where P_o is the relative observed agreement among raters, and P_c is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$. The factor $1 - P_c$ gives the degree of agreement that is attainable above chance, and $P - P_c$ gives the degree of agreement actually achieved above chance. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.

Let n_{ik} represent the number of raters who assigned the i^{th} subject to the k^{th} category. First calculate p_c as the proportion of all assignations which were to the c^{th} category:

$$p_c = \frac{1}{e \cdot t} \sum_{i=1}^{e} n_{ik}, \tag{2}$$

The probability of change is then computed as: $P_c = \sum_{c=1}^{c} p_c$.
Now calculate P_i, the extent to which raters agree about the i^{th} event:

$$P_i = \frac{1}{t(t-1)} \sum_{k=1}^{c} n_{ik}(n_{ik} - 1) \tag{3}$$

Now compute P_o,to be entered into the formula for κ:

$\kappa value$	Meaning
< 0	No agreement
0.0 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost perfect agreement

Table 1: Interpretation of the kappa index value according to the Landis scale [36].

$$P_o = \frac{1}{e}\sum_{i=1}^{e} P_i \tag{4}$$

In this paper, the function `kappam.fleiss` of the package `irr` of the software R [35] has been used to compute the kappa index κ in the different scenarios.

Table 1 shows how to interpret the significance of the kappa index value according to the Landis scale [36]. This table is widely used, although it is not universally accepted. Some authors point out that these guidelines may be more harmful than helpful [37], as the number of categories and events will affect the magnitude of the value: the kappa will be higher when there are fewer categories [38]. In the context of prosodic labelling consistency tests, this is specially important since some of the labels occur very infrequently while other labels (or one of the labels) are very frequent.

This metric has been used in [21] to contrast the inter-transcriber reliability of prosodic events on a subset of the Switchboard [39] corpus using adapted ToBI for English. Cohen's kappa is also proposed in [40] to evaluate the reliability among transcribers using ToBI for American English under relatively optimal conditions.

3.4. Pairwise transcriber agreement

Another common procedure to measure interreliability in prosodic labelling experiments is to count the number of labelling agreements for all pairs of transcribers. Instead of comparing the labels assigned by individual transcribers against the group, this pairwise analysis compares the labels of each transcriber against the labels of every other transcriber for the particular event to be analyzed. That is, 4 transcribers (T1, T2, T3, T4) would produce 6 possible transcriber pairs (T1T2, T1T3, T1T4, T2T3, T2T4, T3T4), and the criterion is conservative: if 3 of 4 transcribers agree, only 3 of 6 pairs will match, making the agreement rate 50% (agreement = agree / (disagree + agree)). For example, if a particular pitch accent was labeled by the first transcriber as H*, by the second transcriber as LH*, and by transcribers 3 and 4, as H*, the number of transcriber pairs who agree with each other is three (T1T3, T1T4, T3T4) and the number of transcriber pairs who disagree with each other is also three (T1T2, T2T3, T2T4).

More formally, the set of pairs can be defined as:

$$Pairs = \{(C_{i,j_1}, C_{i,j_2}), \quad i = 1..e, \quad j_1, j_2 = 1..t, \quad j_1 < j_2\} \tag{5}$$

C_{i,j_1} and C_{i,j_2} being the categories assigned by the labellers j_1 and j_2 respectively to prosodic event i. Let us call $n^p_{m,n}$ (the superscript p refers to *Pair*) the number of times a labeller tagged a

given subject i with the category m and another different labeller judged the same event to be n, formally

$$n^p_{m,n} = Card\{Pairs \quad | \quad (C_{i,j_1}, C_{i,j_2} = m, n) \\ \vee \quad (C_{i,j_1}, C_{i,j_2} = n, m)\} \tag{6}$$

The number of pairs in agreement is $n^p_A = \sum_{c=1}^{k} n^p_{cc}$ and the disagreement is $n^p_D = \sum_{c=1}^{k} \sum_{d=c+1}^{k} n^p_{cd}$. The pairwise transcriber agreement index can be computed as:

$$pta = \frac{n^p_A}{n^p_A + n^p_D} \tag{7}$$

This index has been used to assess ToBI since the seminal ToBI papers [41] and [42], and it is considered a reference to test the consistency of other annotation systems before they can be considered standard (G_ToBI in [43], Gla_ToBI in [44], K_ToBI in [45], J_ToBI in [8]). Benefits obtained from the use of alternative tiers for ToBI have also been evaluated with this index [46].

The pairwise transcriber agreement index has the advantage of permitting the consistency of every class to be analyzed separately: $n^p_{e,e}$ represents the agreement of labellers when the class C_e is identified. $n^p_{e,d}$, or $n^p_{d,e}$ represents degree of the confusion of the symbol C_e with respect to the symbol C_d. This information can be displayed as a squared, triangular $c \times c$ contingency table or confusion matrix. To relate these indicators to the frequency of the symbol, we compute:

$$pta_{e,d} = \frac{n^p_{e,d}}{\sum_{a=1}^{k} n^p_{a,d} + \sum_{a=1}^{k} n^p_{a,e}} \quad e = 1..c \quad d = 1..c \tag{8}$$

Confusion matrices have been used by [17] and [21] to analyze the conceptual similarity of ToBI tones. [21] uses the confusion matrix in absolute terms while [17] introduces the equations above to compare tag assignments. [17] also presents separate tables for each pair of labellers.

4. Visualizing the inter-transcriber confusion with multidimensional scaling

The statistics described above have been commonly used to assess the degree of consistency in ToBI-framework systems, since high consistency is a requirement of the system before it can be considered a standard [13]. Nevertheless, the goal of this work is not to certify that Cat_ToBI has achieved the needed degree of consensus to be accepted as a standard system of prosodic annotation. As noted above, the speech database with prosodic annotations described in section 2 will be taken as a source of data to which a new procedure is applied in order to visualize intercoder agreement and identify those symbols that can introduce important biases in the annotations of projects like *Glissando* that involve working with large corpora. In this section we explain how Multidimensional Scaling can be useful in this regard.

4.1. Multimensional Scaling the basis

Multidimensional scaling (MDS) is a set of related statistical techniques often used in information visualization for exploring similarities or dissimilarities in data [47]. Generally, the data to be analyzed is a collection of I objects on which a distance function is defined, $\delta_{i,j}$ = the distance between i^{th} and j^{th} objects.

These distances constitute the entries in the dissimilarity matrix

$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix}. \tag{9}$$

such that $\delta_{i,i} = 0$, $\delta_{i,j} \geq 0$ and $\delta_{i,j} = \delta_{j,i}$. The goal of MDS is, given Δ, to find I vectors $x_1, \ldots, x_I \in \mathbb{R}^N$ such that

$$|x_i - x_j| \approx \delta_{i,j} \quad \forall i, j \in I, \tag{10}$$

Thus, MDS attempts to find a correspondence between the I objects and $\mathbb{R}^N$ such that distances are preserved. If the dimension N is chosen to be 2 or 3, we may plot the vectors x_i to obtain a visualization of the similarities between the I objects.

There are various approaches to determining the vectors x_i as they are not unique. MDS is formulated as an optimization problem to be solved numerically, where $(x_1, \ldots, x_I)$ is a minimizer of the cost function:

$$\min_{x_1,\ldots,x_I} \sum_{i<j} (|x_i - x_j| - \delta_{i,j})^2. \tag{11}$$

The obtained eigenvector and eigenvalues are used for displaying the plots [48] so that the distances in the Δ matrix are projected into the distances between I representative points. In this work, the command `cmdscale` of the software R [35] has been used. This is an implementation of the classical principal coordinates analysis for obtaining the eigenvalues from the data matrix.

4.2. Multimensional scaling for inter-rater consistency evaluation

We propose the use of MDS to visualize the consistency of judgements by making $\delta_{i,j}$ relative to the inter-transcriber metrics. There are two situations in which this technique will be used: visualization of the inter-rater consistency, and visualization of the inter-symbol confusion.

Visualizing the distance between the judgements of every pair of labellers can be useful to identify badly trained taggers or different tagging criteria. The κ index can be obtained for every pair of labellers where $\kappa_{i,j}$ is the κ index computed with the samples of labeller i and labeller j in isolation. By making $\delta_{i,j} = max(0, 1 - \kappa_{i,j})$ we obtain a measurement of the distance between the pair of taggers such that the higher its value, the greater the inter-rate confusion. The computation of $\delta_{i,j}$ $\forall i, j = 1..t$, $i \neq j$ permits a distance matrix Δ to be defined. MDS techniques allow a set of vectors x_i with $i = 1..t$ to be obtained so that each x_i represents a labeller and the distance between the vectors is assumed to be proportional to the confusion between the labellers. Dimension two is selected to easily display the distances between the judgements of the labellers on a 2D plot.

The second situation where we expect to obtain benefits from the application of MDS techniques is in visualizing the distances between the symbols that represent prosodic events. The index $pta_{i,j}$ can be interpreted as the confusion between the pair of symbols i and j as explained in section 3.4. The higher $pta_{i,j}$ the greater the confusion between the pair of symbols. By making $\delta_{i,j} = max(0, n^p_{i,i} + n^p_{j,j} - n^p_{i,j})$ $\forall i, j = 1..c$, with $n^p_{i,j}$ as described in section 3.4, the Δ matrix can be obtained to be displayed by using MDS techniques. By entering the terms $n^p_{i,i}$ and $n^p_{j,j}$, we guarantee that the more consistent symbols will be separated in the plot. As the term $n^p_{i,j}$ increases, the symbols get closer. The distances between the symbols on the MDS plot are representative of the confusion between them so that two symbols are close to each other in the

Multiclass decision

CORPUS	**L**	**W**	**S**	**Pitch Accents**	**Boundary Tones**	**Breaks**
Cat-ToBI	10	264	4	0.462 / 61.17 %	0.69 / 86.10 %	0.68 / 77.14 %
Am_ToBI (fe)[21]	4	644	2	0.69 / 71%	0.84 / 86%	0.65 / 74%
Am_ToBI (ma)[21]	4	644	2	0.67 / 72%	0.76 / 82%	0.62 / 74%
E_ToBI[44]	26	489	4	na / 68%	na / 85%	na / 67%
E_ToBI[18]	2	1594	1	0.51 / 86.57%	0.79/ 89.33%	na / na
G_ToBI[20]	13	733	5	na / 71%	na / 86%	na / na
K_ToBI[19]	21	153	5	na / 52.2%	na / 81.6%	na / 65.5%

Binary decision

CORPUS	**L**	**W**	**S**	**Pitch Accents**	**Boundary Tones**	**Breaks**
Cat_ToBI	10	264	4	0.706 / 85.56 %	0.802 / 92.15 %	0.75 / 88.38 %
Am_ToBI (fe)[21]	4	644	2	na / 92%	na / 93%	na / na
Am_ToBI (ma)[21]	4	644	2	na / 91%	na / 91%	na / na
E_ToBI[44]	26	489	4	na / 90%	na / 81%	na / na
E_ToBI[18]	2	1594	1	0.75 / 89.14%	0.58/ 90.9%	na / na
G_ToBI[20]	13	733	5	na / 87%	na / na	na / na

Table 2: Global inter-transcriber agreement results for Cat_ToBI contrasted with results reported for other ToBI systems. Columns labelled *PitchAccents*, *BoundaryTones* and *Breaks* separate results according to the respective ToBI events that have been considered. The figure in the cells are the κ index and the pairwise inter-transcriber rate (as a percentage). In the *Multiclass decision* Table all symbols are considered while the *Binary decision* one only contrasts the presence or absence of the corresponding event. **L** is the number of labellers, **W** is the size of the corpus in words and **S** is the number of styles. *(fe)* is female, *(ma)* is male and *(na)* means the information is not available.

MDS plot when different labellers have frequently assigned these symbols to the same event in the transcription procedure.

MDS techniques allow a set of vectors x_i with $i = 1..x$ to be obtained such that each x_i represents a class of symbols and the distance between the vectors is assumed to be proportional to the confusion between the symbols. Again, we select dimension two to easily display the distances between the ToBI symbols on a 2D plot.

MDS has been already used in the context of ToBI labeling as an inter-transcriber reliability measure in [17]. In [17], MDS is used to convert into distances a categorical index named the *Conceptual Similarity Index*. These distances are assumed to be representative of the difference in criteria between taggers and displayed in a set of 2D plots, one for every pair of labellers. Our approach differs in that we use MDS to project on a 2D plot the confusion matrices for helping on interpreting inter-rater information indices.

The next section reports the results obtained when these three tests were applied to the Cat_ToBI annotations made by the ten participants labellers on the twenty utterances taken from the Catalan corpus.

5. Results

5.1. Global inter-transcriber agreement

Table 2 presents the inter-rate agreement matrix according to the type of ToBI events (Pitch Accents, Boundary Tones, and Breaks – upper table) and according to the distribution of the presence or absence of the same ToBI events (lower table). The measures correspond to the two

numbers in each cell of the three right-most columns are the kappa index and the pairwise inter-transcriber rate given as a percentage: in the upper *Multiclass decision* table, all symbols are considered while the lower *Binary decision* table contrasts only the presence or absence of the corresponding event. In both cases, the first row shows the global inter-rate agreement obtained in the prosodic annotation of the Catalan corpus using Cat_ToBI, while the rows below show the results reported for other ToBI systems, namely American English ToBI Am_ToBI [40], English E_ToBI [42, 21], German G_ToBI [43] and Korean K_ToBI [45].

In general, the agreement results obtained in this study are comparable to the agreement results obtained in other ToBI studies. First, as in previous studies, the results for the binary decision task are higher than for the multiclass decision task (i.e., choice of Pitch Accent/Boundary Tone/Break type). In terms of the presence or absence of Pitch Accent, regardless of its type, agreement is 85.56%, while agreement on the presence or absence of Boundary Tones is 92.15%. These figures are also in the range reported by previous studies. The kappa coefficients for Pitch Accents, Boundary Tones and Breaks are of over 0.7, which indicates that those categories have been reliably labelled. In the case of binary decision, results increase to *Almost Perfect Agreement* on the Landis scale (see Table 1) for Boundary Tones and Breaks. As expected, the correlation between the pairwise inter-transcriber agreement and the kappa index is high: the higher the kappa the higher the inter-rater agreement.

The upper *Multiclass decision* Table shows that, as in other studies, the agreement on which label is assigned within a Pitch Accent, Boundary Tone and Break index category is lower than in the binary decision task, as shown by the relatively smaller inter-rate agreement results and kappa coefficients for these measures. The agreement on the choice of Pitch Accent is 61.17%, agreement on the choice of Boundary Tone is 86.10% and agreement on the choice of Break Index is 77.14%. These agreement results are comparable to previous ToBI studies, which are in the interval of [52.2%, 86.57%] for Pitch Accents, [81.6%, 89.93%] for Boundary Tones and [65.5%, 74%] for Breaks. According to the Landis scale, we have thus obtained *Substantial Agreement* in most cases. Only Pitch Accents shows *Moderate-Fair Agreement*.

In general, the task of labelling Boundary Tones and Breaks gives more consistent results than the task of labelling Pitch Accents: the consistency in the labelling of Boundary Tones is 86.10%, the consistency for Breaks falls to 77.14% and the consistency for Pitch Accents is lower still at 61.17%.

Despite the a priori importance of the number or classes in the value of the metrics, results are better for Boundary Tones than for Pitch Accents. Transcribers had a choice of 9 Boundary Tone types and 7 Pitch Accent types. This result is representative of the degree of difficulty of the Pitch Accent labelling task, which we will take up in the Discussion (section 6).

Despite the high inter-transcriber reliability results, Table 3 shows examples of certain types of inter-transcriber labelling inconsistencies, which may be significant. For instance, there is no complete agreement in the identification of the presence of Pitch Accents: in the selected example *No m'has dit que anava a comprar roba?* (*You told me that he/she went to buy clothes, didnt you?*), raters differed in their labelling, detecting the presence of two, three or four accented syllables. Another very frequent inconsistency is the selection of rising Pitch Accents, which were labelled as *L+H** by some transcribers and as *H** by others (see example *Què li duries?* (*What would you bring him/her?*) in Table 3). Another type of inconsistency found in the data has to do with the levels of prosodic break (e.g. in the sentence *Empassant saliva amb esforç vaig abraçar-lo tendrament, tement que esclatés a plorar i jo ja no pogués aguantar més* (*Swallowing hard, I hugged him tenderly, fearing that he would break into tears and I could not take it any more*) coder I2 discriminates level 3 and 4, whereas coder I1 interprets all the Breaks as intonational

Sentence	Labels	Rater
Presence / Absence of accents		
No m'has dit que anava a	[no], [na:] [bra:] [ro]	E4
comprar roba?	[no], [na:] [ro]	I1
(Didn't you tell me to go shopping?)	[dik] [ro]	E2
Type of pitch accents		
Què li duries?	H* H*	E2
(What would you bring him/her?)	H* L+H*	I1
	L+H* H*	E4
Breaks		
Empassant saliva amb esforç	Empassant ... esforç 4 ... tendrament 2	E4
vaig abraçar-lo tendrament,	... plorar 4 ... més 4	
tement que esclatés a plorar	Empassant ... esforç 3 ... tendrament 3	I1
i jo ja no pogués aguantar més	... plorar 3 ... més 4	
(Swallowing with efford, I embraced him tenderly, fearing that he	Empassant ... esforç 4 ... tendrament 4	I2
would break into tears and I would not be able to take it any more)	... plorar 4 ... mé 4	
Boundary Tones		
Eren les sis de la matinada	Eren ... matinada M%	E2
i tota aquella gent	... gent H% ... prou L%	
semblava no tenir-ne mai prou.	... dormir HH% ... conmpanys? HH%	
Que no voleu anar a dormir,	Eren ... matinada M%	E3
conmpanys?	... gent H% ... prou L%	
(It was six a.m. and these people never seemed	... dormir HH% conmpanys? HH%	
to get enough. Don't you want to go to sleep, folks?)	Eren ... matinada M- ...	E4
	... gent H- ... prou L%	
	... dormir H- ... conmpanys? HH%	

Table 3: Examples of inter-transcriber labelling inconsistencies

phrases) and the implementation of Boundary Tones. Finally, in the fourth example, labellers have variously labelled the Break Index category after the words *matinada*, *gent* or *dormir*.

In the following sections, we put forward the use of a set of global inter-transcriber metrics to show that the analysis of inconsistencies can shed light on the reasons behind the observed confusions.

5.2. Joint agreement

In order allow us to go into the consistency analysis in greater depth, Table 4 depicts the joint agreement results, taking into account each of the categories considered in the prosodic annotation of the Catalan corpus. The *Count* columns show the number of labellers that assigned a given symbol and the $Statistics$ columns report the grouping (mean, median, and mode values) and dispersion statistics (i.e., *Asymmetry coefficient* (AC) and *kurtosis coefficient* [49]) of the distribution function $f(n_k)$. The use of the joint agreement distribution is new in the field of prosodic labelling and allows us to identify the problematic categories, that is, categories showing a high degree of disagreement among raters.

The interpretation of the results in the Table should proceed as follows:

1. The closer the *mean*, *median* and *mode* values are to the maximum, the higher the consensus (the maximum is 10 as the number of labellers is 10).
2. The *asymmetry coefficient* measures how close the rates are to the minimum value (positive *AC*) or to the maximum value (negative *AC*). The *Kurtosis coefficient* is higher when data are grouped around a given value.

With respect to the information offered by the *mean*, *median* and *mode* values, two observations may be made:

Pitch Accents

	Count										Statistics				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**mean**	**median**	**mode**	**AC**	**CK**
0	15	14	3	8	10	5	4	5	14	49	6.7	8	10	-0.5	1.6
H*	35	14	11	6	3	1	1	1	1		2.3	2	1	1.8	6.2
H+L*	15	8	8	2	5	2		1			2.6	2	1	1.0	3.3
L*	40	10	15	8	9	5	3	3	2	1	3.0	2	1	1.1	3.3
L*+H	5		1	1	1	1					2.6	1	1	0.5	1.4
L+>H*	30	7	2	1							1.4	1	1	2.1	6.9
L+H*	25	5	11	13	9	8	12	13	4	4	4.6	4	1	0.2	1.8

Boundary Tones

	Count										Statistics				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**mean**	**median**	**mode**	**AC**	**CK**
0	10	4	2	4	5	3	2	5	9	121	8.7	10	10	-1.9	5.1
H%	11	7	5	2	4	5	1	1	4		3.7	3	1	0.7	2.1
HH%	2	2	2	1		1	1	1	3	2	5.6	6	9	-0.0	1.2
HL%	4		1		1	1	1		1		3.8	3	1	0.4	1.4
L%	10	3	1	3	2	3	2	2	2	11	5.6	6	10	-0.0	1.3
LH%	7		1								1.2	1	1	1.9	4.7
LHL%	2			1							2.0	1	1	0.4	0.7
LM%	2				1						2.3	1	1	0.4	0.7
M%	12	2							1		1.7	1	1	3.0	10.7

Breaks

	Count										Statistics				
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**mean**	**median**	**mode**	**AC**	**CK**
B0	23	16	6	3	4	3	3	3	9	47	6.1	8	10	-0.2	1.2
B1	12	6	9	6	5	6	7	24	11	18	6.3	8	8	-0.5	1.8
B2	23	4	2	1	2						1.6	1	1	1.9	5.4
B3	8	4	2	5	9	9	1	6	5	2	5.1	5	6	-0.0	1.9
B4	3	1		5	3	1	1			22	7.6	10	10	-0.8	2.0

Table 4: Joint agreement table for Cat_ToBI results. The three tables refer to the different Cat_ToBI prosodic events that have been considered within the categories of *PitchAccents*, *Boundary Tones* and *Breaks*. Each row refers to a different prosodic category. *Count* columns show the number of labellers that assigned the corresponding symbol to a given prosodic event. *Statistics* columns report *mean*, *median*, and *mode* values the *asymmetrix* (AC) and *kurtosis coefficients* (KC).

- For Pitch Accents (Table 4), only the symbol 0 (absence of accent) seems to achieve an acceptable degree of consensus (*mode* = 10). For the remaining Pitch Accents, only the symbol *L+H** has a mean value higher than 4. The symbols *L*+H* and *L+>H** are problematic because they have been identified very rarely (low total count) and whenever they have been assigned by any of the raters, the remaining raters do not agree (*median* = 1).

- For Boundary Tones (Table 4), symbols 0 (absence of Boundary Tone) and *L*% obtain the highest agreement rate, with *mode* = 10. *HH%* and *L%* seem to be the easiest boundaries to label (*mean* = 6). On the other hand, the symbols *LH%*, *LHL%*, LM% and *M%* are problematic. *H%* and *HL%* symbols achieve high number of isolated occurrences (*mode* = 1) but they also have a significant number of occurrences with a high agreement (*mean* > 3.7).

- For Breaks (Table 4), the highest agreement is obtained for Break 0 and 4 (*mode* = 10). Break 1 and Break 3 have a significant agreement (*median* = 8 and 5 respectively), but Break 2 is clearly problematic (*median* = 1).

Concerning the dispersion statistics, in terms of the *asymmetry coefficient* measurement, only the symbols 0 (*Pitch Accents* and *Boundary Tones*) and Breaks 0, 1 and 4 obtain satisfactory results for this indicator ($AC < -0.2$ in Table 4). The *kurtosis coefficient* is higher when data are grouped around a given value. For Cat_ToBI (Table 4), the highest values are obtained with the symbols *H**, *L+>H**, *Boundary Tone=0*, *M%*, *LH%* and *Break 2*. Only *Boundary Tone=0* has a negative *AC* value. The remaining symbol have a grouped distribution which mean is close to one, indicating a problematic situation.

It is inferred from the results that the joint agreement table is useful to identify problematic symbols, when different symbols have been used to label the same prosodic event. Nevertheless, the information about the one or the other category to which each of the symbols is inconsistently assigned is missing. This is the reason why contingency tables in combination with multidimensional scaling have been applied to the data, as explained in the following subsection.

5.3. Pairwise inter-transcriber agreement and MDS plots

Table 5 reports results for pairwise inter-transcriber agreement measured for the different ToBI categories. Again, the results are organized for the data corresponding to *Pitch Accents*, *Boundary Tones* and *Breaks*. At left are shown contingency tables (two tables per type of ToBI event, both representing the number of pairs, in absolute and relative terms, respectively), while at right are shown the corresponding 2D plots that depict the inter-symbol distance obtained by applying the procedure explained in section 4.2.

Contingency tables are difficult to interpret due to the high number of pairs taken into account. The transformation of these tables into a 2D plot by using multidimensional scaling is a useful tool that helps in the interpretation of results. Briefly, the shorter the distance in the 2D plot, the higher the inter symbol confusion.

These tables provide that some pairs of symbols are more easily to be confused than others, in the following terms:

- For *Pitch Accents*, the symbols 0 (2594-36.2%) –the number in parenthesis here represent the value of the corresponding cell from upper and lower sub-tables in Table 5; i.e., absolute and relative agreement rates, respectively – and *L+H** (1162-23.3%) are the least confused ones, showing the highest rates both in absolute and relative terms, followed by *L** (397-13.6%). Very low rates have been obtained for the symbol *L+>H**: 19-2.4%. Symbols *H**, *H+L** and *L*+H* are problematic because they obtain low relative rates (10%, 9.7% and 12.7% respectively).

  The most frequent inter-class confusions can be visualized in the MDS 2D plot of Table 5. It presents four clusters of labels: (the first cluster) no accent, (the second cluster) *L+H**, (the third cluster) *L*+H*, *H+L** and *L** and (the fourth cluster) *H** and *L+>H**. The third cluster is composed of the low accent tones (*L*) and the fourth cluster the high accent tones (*H*) except *L+H**. The closer the symbols the easier it is to confuse them so that most of the confusions seem to appear among conceptually similar symbols.

- For *Boundary Tones*, results are also coherent with the results obtained in Table 4: labels *0*, *L%* and *HH%* seem to be the easiest symbols to tag; the symbol *H%* is also quite easily identified. The rest of the symbols are very frequently confused among themselves, forming a common cluster in the 2D MDS plot.

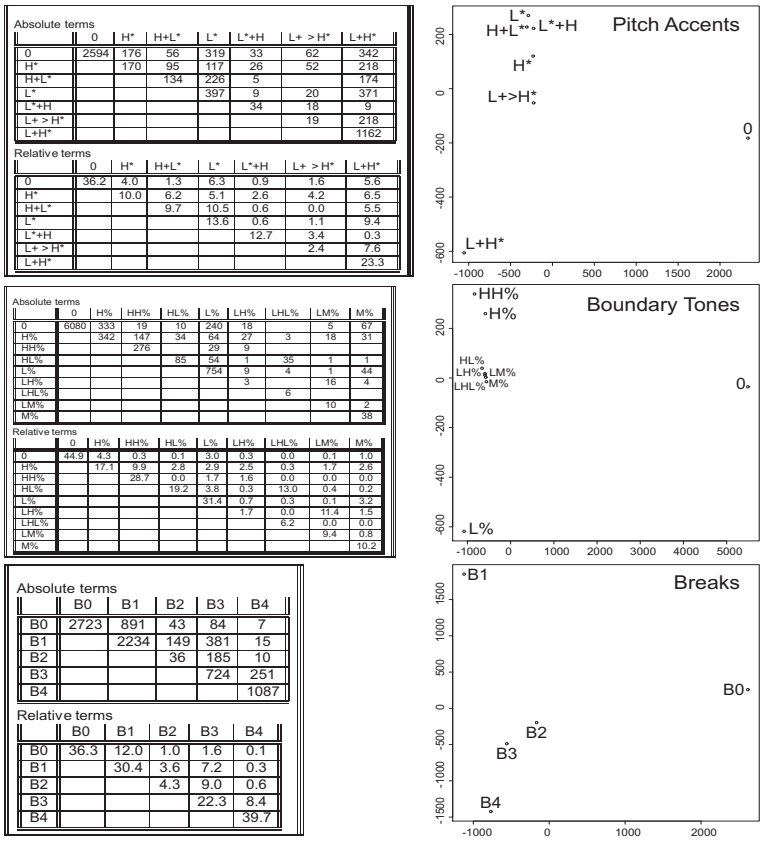
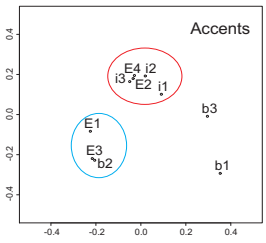


Table 5: Contingency table for Pairwise inter-transcriber agreement in Cat_ToBI transcriptions. From top to bottom, there are results for *Pitch Accents*, *Boundary Tones*, *Breaks* respectively (from top to bottom). Contingency tables are at left and the corresponding 2D plot depicting the inter-symbol distance are at right. There are two contingency tables per type of ToBI event, the upper of the two showing counts in absolute terms and the lower table showing counts in relative terms (in percentage). BI is Break I, with I=0..4.

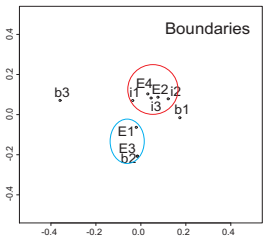
- For *Breaks*, the contingency table shows a fairly good percentage of pairwise inter-transcriber agreement in the case of *Break 0, 1*, and *4*: 36.3%, 30.4% and 39.7% respectively. Consequently, the MDS 2D plot shows a triangle formed by these symbols. By contrast, the transcribers disagree with respect to the use of the symbol *Break 2*, which is frequently labelled as *Break 3*: this behaviour can be observed as a cluster in the MDS 2D plot. This symbol is close to *Break 4* because *Break 3* is most often mislabelled as *Break 4* (8.4%).

Although these results are consistent with the ones obtained when applying the joint agreement measures, we will explore our results further in order to find the reasons behind the reported disagreements. The next sections focus on the labellers' behavior in order to identify whether the disagreements detected are due to lack of training or rather to difficulty in the application of different labelling criteria.

	E1	E2	E3	E4	i1	i2	i3	b1	b2	b3
E1		0.60	0.73	0.47	0.46	0.48	0.56	0.32	0.72	0.34
E2			0.45	0.39	0.51	0.56	0.83	0.31	0.44	0.40
E3				0.37	0.46	0.46	0.51	0.34	0.98	0.35
E4					0.44	0.48	0.40	0.25	0.37	0.28
i1						0.58	0.54	0.37	0.47	0.41
i2							0.59	0.32	0.46	0.37
i3								0.31	0.50	0.39
b1									0.35	0.31
b2										0.36
b3										



	E1	E2	E3	E4	i1	i2	i3	b1	b2	b3
E1		0.78	0.82	0.75	0.76	0.70	0.81	0.64	0.82	0.59
E2			0.67	0.78	0.77	0.74	0.90	0.74	0.67	0.55
E3				0.65	0.68	0.64	0.69	0.65	0.99	0.55
E4					0.71	0.70	0.78	0.67	0.65	0.56
i1						0.69	0.80	0.64	0.68	0.61
i2							0.76	0.63	0.64	0.47
i3								0.69	0.69	0.57
b1									0.65	0.44
b2										0.55
b3										



	E1	E2	E3	E4	i1	i2	i3	b1	b2	b3
E1		0.73	0.83	0.71	0.52	0.62	0.68	0.65	0.83	0.66
E2			0.69	0.72	0.55	0.73	0.95	0.79	0.69	0.70
E3				0.78	0.54	0.63	0.68	0.67	1.00	0.70
E4					0.56	0.66	0.72	0.71	0.78	0.62
i1						0.56	0.52	0.60	0.54	0.50
i2							0.73	0.67	0.63	0.57
i3								0.76	0.68	0.69
b1									0.67	0.65
b2										0.70
b3										

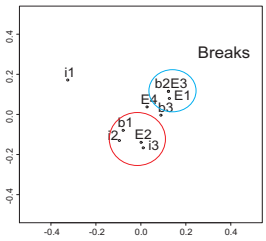


Table 6: Kappa index for each pair of Cat_ToBI transcribers. The rows correspond with the prosodic categories *Pitch Accents*, *Boundary Tones* and *Breaks* respectively (from top to bottom). Tables at left display the κ coefficient of agreement between the labellers indexed in the respective row and column (*b* stands for beginner, *E* for expert and *i* for intermediate skill level). At right are displayed the corresponding matrices in an MDS 2D plot that interprets the κ coefficient as a distance.

5.4. Inter-rater disagreement

Table 6 represents the kappa fleiss inter-transcriber agreement for the three types of prosodic events that were annotated in the Catalan utterances. The tables at left display κ coefficient of agreement between the labellers, while the grahs at right displays the respective matrices in an MDS 2D plot that interprets the κ coefficient as a distance (Section 4.2 explains the procedure applied to obtain the distances). The advantage of the 2D plot is that it permits the pair of taggers that show the highest degree of agreement to be detected easily since the greater the agreement, the closer the labellers appear on the plot.

Results reveal particular behaviours in the labelling tasks since some of the labellers are plotted at quite a distance from the other labellers in the 2D Plot. This behaviour is exhibited by *b*1 in the *Pitch Accent* plot, *b*3 in the *Boundary Tones* plot, and *i*1 in the *Breaks* plot in Table 6.

On the other hand, some of the coders are grouped together in clusters. In Cat_ToBI, for *Pitch Accents* the first cluster (red oval) is *E1, E3, b2* (inter-transcriber κ from 0.72 to 0.98), and second cluster (light blue oval) is *i2, i3, E2* (inter-transcriber κ from 0.56 to 0.83) (Table 6); for *Boundary Tones* the first cluster (red oval) is *E1, E3, b2* (inter-transcriber κ from 0.82 to

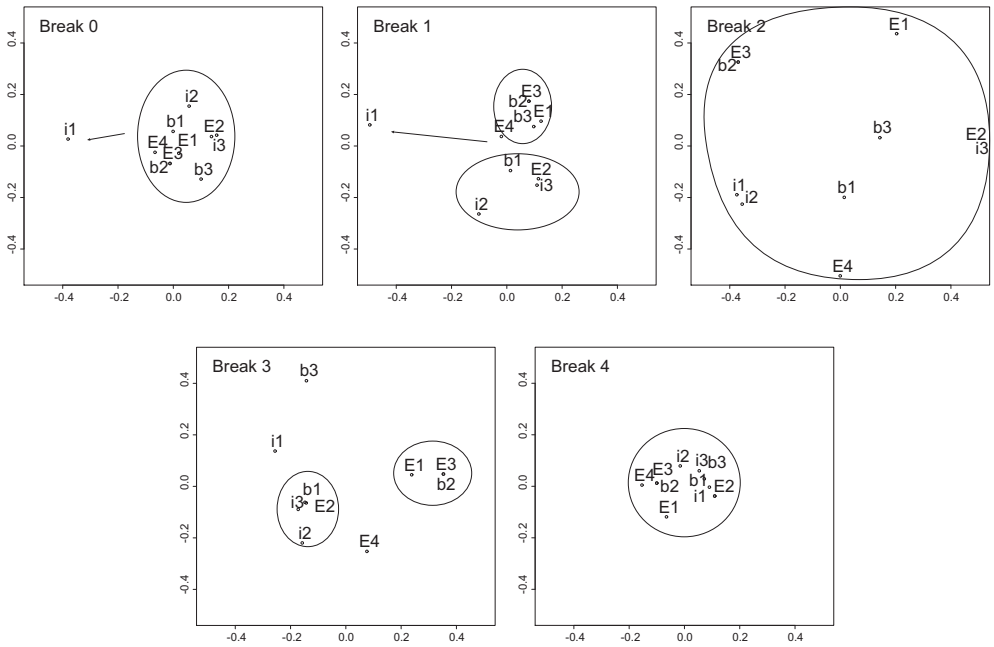


Figure 2: Intertranscriber discrepancy for Breaks. The 2D MDS plots represent the distance between taggers for a given ToBI symbol.

0.99) and the second luster (light blue oval) is *E2, E4, i1, i2, i3* (inter-transcriber κ from 0.69 to 0.90) (Table 6) and for Breaks the first cluster (red oval) is *E2, b1, i2, i3* (inter-transcriber κ from 0.67 to 0.95) and the second cluster (light blue oval) is *E1, E3, b2* (inter-transcriber κ from 0.83 to 1.00) (Table 6). This tendency could indicate that the different groups of taggers are using alternate annotation criteria.

Figure 2 mines the available data further in order to gain more information about the reasons for the observed inter-transcriber grouping. The five graphs in Figure 2 show the inter-transcriber discrepancy for the Break Index category (with a graph for each of the five levels in this category). The results show that while Break 0 and Break 4 are quite consistent among labellers, the rest of the Breaks are more problematic. While Break 1 and Break 3 show two groups of labellers, Break 2 shows a greater dispersion, which indicates that the presence of this category generates clear uncertainty in labellers. The main point of the discussion (section 6) is to assess the behaviour of certain Cat_ToBI categories, which might be at the source of these inconsistencies transcription.

6. Discussion

There are a number of statistics which can be used to determine the degree of agreement among raters (inter-rater reliability, inter-coder agreement, or concordance), and they are more or less appropriate depending on the different types of measurement. For categorical data, the most popular ones (used to evaluate the consensus regarding the ToBI systems [41, 42, 44, 21] but also to quantify the agreement in other annotation tasks, either phonetic [50] or prosodic [51]) are the joint agreement, the kappa statistics and the pairwise transcriber agreement, which

is why we have applied them in this study. Nonetheless, we have also shown their limitations, and therefore proposed new procedures to refine the processing of the data.

The difficulty that human labellers face when it come to annotating a corpus of spontaneous unread speech is well known (to start with, it is hard to decide where a clause begins and ends, due to changes in communicative strategies, unfinished sentences, etc.) and these difficulties increase when the criteria are mainly perceptual, as in the prosodic labelling task used by a ToBI system. To address this problem, the tests for evaluating the degree of confidence of the manually obtained measurements have been refined by incorporating new procedures to visualize inconsistencies and to identify the sources of different annotation criteria. The MDS techniques are especially useful for this.

This section will assess the results of this methodology whose goal is to address problems involved in the perceptually-based transcription of levels of prosodic organization, namely, the identification of problematic symbols, the identification of problematic labellers and the identification of potentially different tagging criteria. Although the paper presents results for Catalan at all levels of a ToBI-framework system, we suggest that in order to demonstrate the benefits of the application of the proposed methodology it is not necessary for us to be exhaustive in our analysis of all the cases we have identified.

In this section, therefore, we will concentrate on an analysis of the Break Indices, which are the cues for prosodic organization in ToBI systems. The results, which reveal a high degree of coincidence across languages, show that these cues can be considered stable, except when a difference appears in the annotation criteria due to the different degrees of proficiency of the transcribers insufficient study of certain properties of some of the prosodic levels.

6.1. Identification of problematic symbols

Since the ToBI system is grounded in the current state of knowledge of the prosodic and intonational phonology of a given language, it is unsurprising that different annotation criteria correspond to different stages of this knowledge. This is clear when we analyze the behaviour of the transcribers with respect to the break indices Break 2 and Break 3. Table 2 shows that the inter-transcriber agreement for Cat_ToBI Breaks is substantial with $\kappa = 0.68$ and pairwise inter-transcriber index *pta* = 77.14%. Nevertheless, results in Table 4 demonstrate that there is no consensus for some of the levels. As a whole, the symbols Break 0 and Break 4 have high agreement rates, but this is not true for the symbol Break 2, which is highly problematic because it is very infrequent and when it appears, few of the labellers agree on how and when to use it. The *pta* index results (Table 5) offer objective results on the proximity of Break 2 with respect to other symbols. The pairwise inter-transcriber agreement shows us that Break 2 is often confused with Break 3.

In practical situations, such as tagging a corpus, it might be decided to dispense with the symbol Break 2. If we merge the symbols Break 2 and Break 3 in order to build a new category, the new computation of the kappa fleiss metric increases from 0.68 to 0.71 and the *pta* goes from 77.14% to 79.24%. Even though the rate does not improve dramatically, the complexity of the task performed by the transcribers can decrease significantly if the number of symbols is reduced, and as a consequence, the time required to complete the labelling task will be shorter. Regardless of the theoretical implications, what we want to demonstrate is that the proposed methodology can make such decisions more objective and informed.

6.2. Identification of undertrained labellers

The joint agreement results (Table 4) and pairwise inter-transcriber agreement results (Table 5) point out a high confusion concerning the Break 3 and Break 1 labels. In order to better know the reasons that can explain the discrepancies, we analyze the inter-rater disagreement results of Table 6, where one of the labellers (in particular, *I1*) is emitting judgements that are clearly divergent from the rest of the labellers. Since the Break 1 label is well defined in the conventions of the Cat_ToBI system (to mark boundaries between prosodic words), the interpretation of this behaviour is that the transcriber is mis-assigning both Break 3 and Break 1.

In practical applications, such as the selection of the transcribers to work in the processing of the *Glissando* corpus mentioned in section 1, Table 6 and the respective MDS plot could be used to identify badly trained transcribers taking into account objective criteria. In particular, the labeller *I1*, should be discarded due to his/her divergences with respect to the rest of the labellers.

Thus, we offer a tool that can be used to select and evaluate the potential subjects that will participate in a given labelling task particularly any research project in which a high degree of consistency among labellers is needed in order to build a reliable prosodic corpus.

Moreover, this tool can have applications in the field of teaching the system to new transcribers. In our set of transcribers, if is seen as dirirable to improve the proficiency of the labeller *I1*, the precise visualization of the prosodic judgements of the rest of the labellers is a valuable source of information about how to correct the labeller's misjudgements.

Plots in Figure 2 have been obtained by computing the kappa fleiss index for each pair of labellers, as in Table 6, but isolating the subjects that have assigned the given symbol at least once. The κ index has been obtained for every pair of taggers and this index has been transformed into a distance by applying the procedure described in section 4.2. As a consequence of the procedure, we have obtained one plot per symbol where the distances between the points on the graph representing raters are proportional to the inter-rater agreement. In our particular case, we have evidence that the dispersion of labeller $I1$ is due to faulty interpretation of Breaks 1 and 0. A more detailed explanation of the differences between the levels of prosodic organization should be enough to improve the proficiency of this labeller.

6.3. Identification of differences among labelling criteria

Inter-rater disagreement results depicted in Table 6 allow two different groups of labellers to be identified as far as the prosodic transcription of break levels is concerned: group 1 consists of labellers *E1, E3* and *b2* and group 2 is made up of the labellers *E2, i3* and *b1*. The clustering cannot be explained by the training or proficiency of the transcribers, since in both groups experts and beginners are found. Another possible explanation for these discrepancies is the annotation criteria. If the kappa fleiss index is computed with the cluster of labellers *E1 E3 b2*, the kappa fleiss goes from 0.68 to 0.89 . If additionally, as suggested in the previous section, we merge Break 2 and Break 3, the kappa fleiss goes up to 0.90 which constitutes *Almost Perfect Agreement* according to the Landis scale (see Table 1).

When the MDS plot of Table 6 referring to Breaks is split into the plots corresponding to the different breaks in Figure 2, we observe that the grouping is evident in the plots corresponding to Break 1 and Break 3 but the grouping disappears when the symbols Break 0, Break 2 and Break 4 are taken into account. We can conclude that these two groups seem to use different criteria with regard to the symbols Break 1 and Break 3 and that these different criteria are responsible for the problematic results observed in terms of joint agreement (see Table 4) for Break 1 and Break 3.

The lack of consensus or the use of alternate criteria in the detection of Break 3 and Break 4 is clear from the data obtained in the different reliability scores. As can be observed in the Boundary Tone example in Table 3 (and in many other similar examples in the corpus), the different labelling obtained for Break 3 and Break 4 can be explained by the use of two different criteria in the identification of the two breaks. In the Cat_ToBI documents, including the *Cat_ToBI Training Materials* which the labellers used as an online reference, there is a description of the two criteria that must be used to identify the intermediate phrase boundaries, or Break 3, namely (1) the presence of a weaker disjuncture from a perceptual point of view, which is instantiated generally by the absence of pauses; and (2) the idea that the intermediate phrase is typically marked by the presence of H- boundary tones, also called continuation rises. The fact is, however, that these two identifying criteria are partially non-overlapping, and one can find continuation rises that are followed by clear pauses. Depending on whether specific labellers attach more importance to one or the other of these two criteria, they will transcribe the boundary as either Break 3 or Break 4. It is clear that the revised version of the *Cat_ToBI Training Materials* must establish a priority ranking in the criteria for identifying intermediate phrase boundaries.

In this discussion, we have shown that the tools presented here represent a useful starting point for an inter-expert discussion about the points of discrepancy observed in the sentences of the corpus, a process which will be taken up shortly by the Cat_ToBI developers group.

7. Conclusions

For the preparation of an oral corpus for research purposes, the availability of tools that can help human subjects in the sometimes difficult task of prosodic annotation is undoubtedly of great interest. Thus the development of a tool that can estimate with objective measures the attainable degree of agreement between transcribers constitutes an important towards achieving of homogeneity and consistency in the data contained in the oral corpus.

In this paper we have systematically compared the performance of several transcribers carrying out Cat_ToBI prosodic labelling experience on various examples of Catalan utterances and evaluated inter-rater consistency of their transcriptions. In general, the results demonstrate that there is a high degree of coincidence in the transcriptions, and therefore that the audio and visual cues to prosodic and intonational organization can be considered relatively stable. Comparison of the present results with those of previous ToBI reliability studies for other languages (namely G_ToBI in [43], Gla_ToBI in [44], K_ToBI in [45] and J_ToBI in [8]) reveals comparable agreement rates for this study. The global inter-transcriber results are 86.10% for Boundary Tone choices, 77.14% for Break Index choices, and 61.17% for Pitch Accent choices. These results lie in the range of previous interreliability results in the cited ToBI studies, which are in the interval [81.6%, 89.93%] for Boundary Tones, [65.5%, 74%] for Breaks and [52.2%, 86.57%] for Pitch Accents (see Table 2). Based on the results of the present inter-transcriber consistency tests, we feel that there is ample evidence to regard the Cat_ToBI system as a standard reference for prosodic labelling.

Although our reliability results for Catalan are of the same order of magnitude as previous studies, the slightly lower scores we obtained in the choice of Pitch Accent, Boundary Tone and Break Index types have deserved further investigation. While it is possible that the inconsistencies detected might be related to the type of speech transcribed (given that the Catalan speech corpus contained four ddifferent speech styles) or the relatively brief training given to some participants, the tools presented here have allowed us to identify a set of issues related to the difficulties involved in transcribing some specific categories.

In this paper, inter-rate reliability has been assessed by means of a set of metrics (joint agreement, pairwise agreement and kappa coefficient) and a visualizing tool (multidimensional scaling) under a common framework. The use of the joint agreement distribution is innovative in the field of prosodic labelling and has been demostrated to be useful for identifying categories with a high disagreement among raters. The combined use of the pairwise inter-transcriber agreement with multidimensional scaling has permitted us to visualize the pairs of symbols that are frequently confused and those pairs that tend to yield greater consensus. The kappa index has allowed us to visualize the existing coincidence among every pair of labellers with the goal of identifying under-trained raters and differences in tagging criteria among different groups of labellers.

On the one hand, our analysis of the confusion clusters has revealed a number of issues that lead to the presence of problematic categories. For example, in section 6, the common confusion between Break 3 or Break 4 has been traced back to partially overlapping identification criteria, which will need to be clarified in a revised version *Cat_ToBI Training Materials* through a more precise description and more clearly constrating examples.

On the other hand, the high number of categories available to the transcribers for both Pitch Accent and the Boundary Tone categories has proven to be one of the serious sources of transcription confusions. Careful evaluation of the data has revealed that, for example, the inventory of rising pitch accents (*L+H**, *H** and *L+>H**) is highly confusable. In the next periodic review of the Cat_ToBI system, this issue will have to be taken up. As noted above, the *Cat_ToBI Training Materials* are a web-based manual for teaching the system to new transcribers, with many recorded examples of transcribed utterances. The conventions are used, maintained and updated consistently from this site, and periodic rechecks are being performed on the data. As a result of the analysis offered in this paper, a simplified Cat_ToBI proposal is going to be put forward as a possible improvement of the system.

In sum, we have presented a low cost procedure that has proved useful for assessing two aspects of a consistency test in particular. First, the identification of the most frequently confused symbols provides evidence that their definitions deserve fresh consideration, and their fusion with more agreed symbols might be a one plausible option. In the specific case of Cat_ToBI, a set of suggestions have been put forward for fewer labelling distinctions both for the transcription of pitch accent events and for boundary tone events. Second, the results of this analysis can help guarantee the necessary level of proficiency of labellers prior to their undertaking the labelling of bigger corpora. Likewise, labellers whose output is seen to deviate from the general consensus must be retrained.

Finally, the proposed procedure can contribute to an efficient and reliable method for evaluating prosodic transcription of speech across languages, something which is needed for linguistic research on prosody in general, and for the development of prosody-dependent labelling and speech recognition systems in particular.

8. Acknowledgements

The authors are indebted to other researchers of the Grup d'Estudis de Prosòdia GrEP (Departament de Traducció i Ciències del Llenguatge Universitat Pompeu Fabra) who contributed constructively to the discussion of this work while it was underway. Particular thanks are due to the subjects who performed the annotations of Catalan utterances (J. Borràs-Comes, V. Crespo-Sendra, R. Sichel-Bazin, E. Estebas-Vilaplana and the postgraduate students CA, CR, EP and GV) for their valuable comments and information.

This research has been funded by three research grants awarded by the Spanish Ministerio de Ciencia e Innovación, namely the *Glissando project* FFI2008-04982-C003-02, FFI2009-07648/FILO and CONSOLIDER-INGENIO 2010 Programme CSD2007-00012, and by grants awarded by the Generalitat de Catalunya to the Grup d'Estudis de Prosòdia (2009SGR-701)

Appendix A. Contents of the corpus

Spontaneous speech extracted from the Map Task subcorpus of the Map Task dialogue corpus *Atles interactiu de l'entonació del català* [26]:

1. Un cop deixes la paret lateral a la teva dreta, la hi deixes? *Once you have left the wall on your right... Have you left it?*
2. No, o sigui, és com si anessis cap al jardí Menor, per abans d'arribar-hi tires cap amunt i cap al jardí Major. *No, in other words, it's like going to the Small Garden but before you get there, go up and towards the Main Garden.*
3. Hi ha un arbre, no?, suposo, a l'esquerra de l'acadèmia? *There is a tree, right?, To the left of the academy?*
4. O sigui que tu vas en direcció cap al final de la paraula Bàrbara? *In other words, you go towards the end of the word Barbara?*
5. No m'has dit que anava a comprar roba? *Didn't you tell me to go shopping?*

Radio news subset from the Festcat database[27]:

1. Però no és molt esclau, això? *But isn't it a very slave occupation?*
2. Per tota una generació, Sílvia Munt serà SEMPRE, la Colometa. *For an entire generation, Sílvia Munt will ALWAYS be the Colometa (nickname, 'little pigeon').*
3. El Bernabéu està completament desesperat! *The whole Bernabeu stadium is utter despair!*
4. Però això no és res! *But its nothing!*
5. Què hi fa, als camps de refugiats? *What is s/he doing in the refugee camp?*

Read text subset from the Festcat database[27]:

1. Des de sempre Hollywood ha produït pel.lícules desaconsellables per a homes sensibles amb serps, llops, aranyes o, fins i tot, extraterrestres. *As long as I can remember Hollywood has produced inadvisable movies for sensitive men with snakes, wolves, spiders or even aliens.*
2. He pensat que la olor havia de ser una de les primeres diferències noticeable. *I thought that the smell should be one of the first notable differences.*
3. Empassant saliva amb esforç vaig abraçar-lo tendrament tement que esclatés a plorar jo ja no pogués aguantar més. *Swallowing with effort, I embraced him tenderly fearing that he started to cry and it was unbearable.*
4. Anem a Eivissa? A Eivissa? A la platja d'Eivissa. *Shall we go to Ibiza? To Ibiza? To the Ibiza beach!*
5. Eren les sis de la matinada i tota aquella gent semblava no tenir-ne mai prou. Que no voleu anar a dormir, companys? *It was six in the morning and these people never seemed to get enough. Don't you want to go to sleep, folks?*

Spontaneous speech extracted from the guided interview subcorpus of the *Atles interactiu de l'entonació del català* [26]:

1. Què li duries? *What would you bring him/her?*
2. Teniu mandarines? *Do you have tangerines?*
3. Home, és d'en Jaume! *It's Jaume's (obviously)!*
4. Va vine..! *Aw, come on...!*
5. És la MARIA la que vol venir? *Is it MARIA, who wants to come?*

References

[1] D. R. Ladd, Intonational Phonology, Cambridge University Press, 1996.
[2] J. B. Pierrehumbert, The phonology and phonetics of English intonation, Ph.D. thesis, MIT (1980).
[3] M. Beckman, J. Hirschberg, S. Shattuck-Hufnagel, The original ToBI system and the evolution of the ToBI framework, in: S.-A. Jun (Ed.), Prosodic Typology: The Phonology of Intonation and Phrasing, Oxford University Press, 2005, pp. 9–54.
[4] M. E. Beckman, M. D. Campos, J. T. McGregory, T. A. Morgan, Intonation across Spanish, in the tones and break indices framework, Tech. Rep. http://www.ling.ohio-state.edu/ tobi/sp-tobi/, University of Ohio (2000).
[5] M. Beckman, Intonation across Spanish in the Tones and Break Indices framework, Probus 14 (2002) 9–36.
[6] E. Estebas, P. Prieto, La notación prosódica del español. Una revisión del Sp_ToBI, Estudios de Fonética Experimental XVIII (2009) 263–283.
[7] M. Grice, R. Benzmueller, Transcription of German using ToBI Tones- The Saarbrücken System, Phonus 1 (1995) 33–51.
[8] J. Venditti, J. Jennifer, The J_ToBI model of Japanese intonation, in: S.-A. Jun (Ed.), Prosodic Typology: The Phonology of Intonation and Phrasing, Oxford University Press, 2005, pp. 172–200.
[9] A. Arvaniti, M. Baltazani, Intonational analysis and prosodic annotation of Greek spoken corpora, in: S.-A. Jun (Ed.), Prosodic Typology: The Phonology of Intonation and Phrasing, Oxford University Press, 2005, pp. 84–117.
[10] M. Beckman, S. Jun, K-ToBI (Lorean ToBI) labeling conventions: version 3, The Korean Journal of Speech Science 7 (1) (2000) 143–169.
[11] P. Prieto, The intonational phonology of Catalan, in: S. Jun (Ed.), Prosodic Typology 2, Oxford University Press: Oxford, to appear in 2012.
[12] E. Estebas Vilaplana, P. Prieto, Castilian Spanish Intonation, in: P. Prieto, P. Roseano (Eds.), Transcription of Intonation of the Spanish Language, Lincom Europa: München, 2010, pp. 17–48.
[13] Ohio-State-University, What is ToBI? (2006).
URL `http://www.ling.ohio-state.edu/ tobi/`
[14] S. Prom-on, Y. Xu, B. Thipakorn, Modeling tone and intonation in Mandarin and English as a process of target approximation, Journal of the Acoustic Society of America 125 (2009) 405–424.
[15] D. Hirst, Form and function in the representation of speech prosody, Speech Communication 46 (2005) 334–347.
[16] C. Wightman, ToBI or not ToBI, in: Proceedings of Prosody, 2002.
[17] R. Herman, J. McGory, The conceptual similarity of intonational tones and its effects on intertranscriber reliability, Language and Speech 45 (2002) 1–36.
[18] C. González, C. Vivaracho, D. Escudero, V. C. noso, On the automatic ToBI accent type identification from data, in: Proceedings of Interspeech 2010, 2010, pp. 142–145.
[19] A. Rosenberg, Automatic Detection and Classification of Prosodic Events, Ph.D. thesis, University of Columbia, USA (2009).
[20] A. Rosenberg, Classification of Prosodic Events using Quantized Contour Modeling, in: HLT-NAACL, 2010, pp. 721–724.
[21] T. Yoon, S. Chavarría, J. Cole, M. Hasegawa, Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI, in: Proceedings of Interspeech, 2004, pp. 2729–2732.
[22] A. K. Syrdal, J. Hirshberg, J. McGory, M. Beckman, Automatic ToBI prediction and alignment to speech manual labeling of prosody, Speech Communication (2001) 135–151, number 3.
[23] S. Ananthakrishnan, S. Narayanan, Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence, Audio, Speech, and Language Processing, IEEE Transactions on 16 (1) (2008) 216–228.
[24] V. Rangarajan Sridhar, S. Bangalore, S. Narayanan, Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework, IEEE Transactions on Audio, Speech, and Language Processing 16 (4) (2008) 797–811.
[25] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, T. Yoon, S. Chavarría, Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus, Speech Communication 46 (2005) 418–439.

[26] P. Prieto, T. Cabro, (COORDS)The Interactive Atlas of Catalan Intonation (2007-2010).
URL http://prosodia.upf.edu/atlesentonacio/index-english.html/
[27] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, J. Perez, Corpus and Voices for Catalan Speech Synthesis, in: Proceedings of LREC 2008, 2008.
[28] L. Aguilar, C. de la Mota, P. Prieto, (coords.) Cat_ToBI Training Materials (2009-2011).
URL http://prosodia.upf.edu/cat_tobi/
[29] P. Boersma, D. W. at the Institute of Phonetic Sciences, Praat, doing phonetics by computer, http://www.praat.org, 2010.
[30] P. Prieto, L. Aguilar, I. Mascaró, F. Torres-Tamarit, M. Vanrell, L'etiquetatge prosòdic Cat_ToBI, Estudios de Fonética Experimental XVIII (2009) 287–309.
[31] J. Uebersax, Diversity of decision making models and the measurement of interrater agreement, Psychological Bulletin 101 (1987) 140–146.
[32] J. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (5) (1971) 378–382.
[33] W. Scott, Reliability of content analysis: The case of nominal scale coding, Public Opinion Quarterly 17 (1955) 321–325.
[34] J. Cohen, A coefficient for agreement for nominal scales, Education and Psychological Measurement 20 (1960) 37–46.
[35] R. Ihaka, R. Gentleman, R: A language for data analysis and graphics, Journal of Computational and Graphical Statistics 5 (3) (1996) 299–314.
[36] J. Landis, G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[37] K. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, British Journal of Mathematical and Statistical Psychology (2005) 26–48.
[38] J. Sim, C. Wright, The kappa statistic in reliability studies: use, interpretation, and sample size requirements, Physical Therapy 85 (3) (2005) 257–268.
[39] J. J. Godfrey, E. C. Holliman, J. Mcdaniel, SWITCHBOARD: telephone speech corpus for research and development, in: Proc. ICASSP, 1992, pp. 517–520 vol.1.
[40] A. Syrdal, J. McGory, Inter-transcriber reliability of ToBI prosodic labeling, in: Proceedings of ICSLP, 2000, pp. 235–238.
[41] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg, ToBI: A standard for labelling English prosody, in: Proceedings of ICSLP-1992, 1992, pp. 867–870.
[42] J. Pitrelli, M. Beckman, J. Hirschberg, Evaluation of prosodic transcription labeling reliability in the ToBI framework, in: Proceedings of ICSLP, 1994, pp. 123–126.
[43] M. Grice, M. Reyelt, R. Benzmuller, J. Mayer, A. Batliner, Consistency in Transcription and Labelling of German Intonation with GToBI, in: in Proc. ICSLP, 1996, pp. 1716–1719.
[44] C. Mayo, M. Aylett, D. Ladd, Prosodic Transcription of Glasgow English: An Evaluation Study of GlaToBI, in: Proceedings of ESCA Workshop: Theory, Models and Applications, 1997.
[45] S. Jun, S. Lee, K. Kim, Y. Lee, Labeler agreement in transcribing Korean intonation with K-ToBI, in: Proceedings of ICSLP, Vol. 3, 2000, pp. 211–214.
[46] A. Brugos, N. Veilleux, M. Breen, S. Shattuck-Hufnagel, The Alternatives (Alt) Tier for ToBI: Advantages of Capturing Prosodic Ambiguity, in: Proceedings of Speech Prosody 2008, 2008, pp. 273–276.
[47] J. Kruskal, M. Wish, Multidimensonal Scaling, Sage Univerity Paper series on Quantitative Application in the Social Sciences, 1978.
[48] I. Borg, P. Groenen, Modern Multimensional Scaling: theory and applications, Springer-Verlag New York, 2005.
[49] D. P. na, Estadística. Modelos y Métodos, Alianza, Madrid, 1999.
[50] M. Pitt, K. Johnson, E. Hume, S. Kiesling, W. Raymon, The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability, Speech Communication 45 (2005) 89–95.
[51] J. Buhmann, J. Caspers, V. van Heuven, H. Hoekstra, J. Martens, M. Swerts, Annotation of prominent words, prosodic boundaries and segmental lenghening by non-expert transcribers in the spoken dutch corpus, in: Proceedings of LREC, 2002, pp. 779–785.