Lang Resources & Evaluation
DOI 10.1007/s10579-015-9320-9

ORIGINAL PAPER

A tool for automatic transcription of intonation: Eti_ToBI a ToBI transcriber for Spanish and Catalan

Wendy Elvira-García[1,2] · **Paolo Roseano**[3] · **Ana María Fernández-Planas**[1] · **Eugenio Martínez-Celdrán**[1]

Abstract This article presents Eti_ToBI, a tool that automatically labels intonational events in Spanish and Catalan utterances according to the Sp_ToBI and Cat_ToBI current conventions. The system consists in a Praat script that assigns ToBI labels to pitch movements basing the assignments on lexical data introduced by the researcher and the acoustical data that it extracts from sound files. The first part of the article explains the methodological approach that has made possible the automatisation and describes the algorithms used by the script to perform the analysis. The second part presents the reliability results for both Catalan and Spanish corpora showing a level of agreement equal to the one shown by human transcribers among them in the literature.

Keywords Intonation · Automatic intonation recognition · Sp_ToBI · Cat_ToBI

1 Introduction

Interest in prosody has grown considerably in recent years, partly due to technical improvements, which have simplified its analysis, but mostly for its contribution to many different linguistics fields, knowledge about it is capital for linguistic fields

✉ Wendy Elvira-García
wendyelvira@ub.edu

[1] Laboratori of Phonetics, Universitat de Barcelona, Edifici Josep Carner, 5e pis, Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain

[2] Department of General Linguistics, Universitat de Barcelona, Barcelona, Spain

[3] Department of Spanish, Universitat de Barcelona, Barcelona, Spain

Published online: 24 October 2015

such as speech synthesis recognition, pragmatics, syntax, and neurolinguistics among others. Prosody is usually understood as the combination of suprasegmental features like duration, intensity, and intonation with intonation being the part that conveys more information. Intonation is created through the variation of pitch and provides information about a wide range of functions e.g., delimitative function, pragmatic function, expressive function, and modality. Its acoustic correlation is F0 (i.e., the frequency of the first harmonic). Consequently, intonational contours can be described through acoustic analysis and, thus, transcribed.

This notwithstanding, the transcription of intonation has often been a matter of controversy triggering the apparition of multiple systems of transcription. The multiplicity of systems for transcribing intonation arose from differences among the theories in which those systems are based. The systems for transcribing intonation can be classified, in general terms, according to three features: (1) approach, (2) goal and (3) method. As far as the approach is concerned, some systems are more phonetically based (e.g., IPO model, Aix-en-Provence), meanwhile other systems have a phonological approach in which differences between contours are relevant only if they convey a change in meaning (e.g., metrical autosegmental model). As far as the goal is concerned, some systems put emphasis on the theoretical approach (e.g., Aix-en-Provence) while others pursue a mostly applied goal, usually in speech synthesis and recognition (e.g., IPO model). Last but not least, systems can be classified based on the method they use to study the F0 contour: while some transcribe intonation 'by ear' -which is considered a poor method (Pierrehumbert 1980: 13)-, other systems have a protocol that relies on acoustic evidence. Among the latter, we can clearly distinguish those systems that perform 'a visual inspection of the contour' (which means that the researcher looks at the F0 contour and describes the curve as a series of F0 falls and rises), from those that base their transcriptions on numeric data (frequency in Hertz or semitones), like the IPO model, the Aix-en-Provence model, and some branches that belong to the Autosegmental-Metrical model (Dorta 2013; Fernández Planas and Martínez Celdrán 2003; Hart and Collier 1975; Hirst et al. 2000).

Nowadays, the most popular system of transcription of intonation is Tones and Break Indices (ToBI) (Hualde 2003, p. 155). It is not a single system applicable to all languages, but rather a group of systems which have been developed specifically for each language and are based on the Autosegmental-Metrical model (AM) (Pierrehumbert 1980). It offers the researcher a labelling system in order to transcribe speech corpora. The researches that have chosen the Autosegmental-Metrical model usually adopt a phonological approach (although there have been attempts of applying the theory to speech synthesis), pursue a theoretical end and their description is based on phonetic evidence (mostly derived from the visual inspection of the curve).

Every transcription system has its advantages and has been designed to perform as well as possible for its purposes, but transcription of intonation, like segmental phonetic transcription, is time consuming. Syrdal et al. (2001) estimate that labelling of prosody with the MAE_ToBI, i.e. the system for the transcription of the intonation of Mainstream American English, system takes from 100 to 200 in real time. For this reason, automating prosodic labelling has being a popular research topic since the first

works on speech recognition and synthesis (Lea 1980). In addition to this, automating the transcription also has another crucial advantage: it helps to prevent the possible mistakes due to subjectivity, distractions or typographical mistakes that a human transcriber can make (Tatham and Morton 2005, p. 372).

This article explains the functioning and presents the results of an automatic transcriber of intonation based on the Spanish and Catalan versions of the ToBI system (Sp_ToBI and Cat_ToBI). In the rest of Sect. 1, Sect. 1.1. presents the current automatic transcribers and 1.2 the general theoretical framework that is used in the paper. Section 2 deals with the implementation of the system and it is divided in Sect. 2.1 which presents the pitch extraction method, and Sect 2.2, an overview of the script structure. Section 3 is dedicated to explain the modified ToBI system that the script uses which uses three levels of analysis, Sect 3.1 explains the surface (narrow phonetic) and the deep (broad phonetic) transcription level and it is subdivided in Sect 3.1.1 for prenuclear accents, Sect 3.1.2 for nuclear pitch accents and 3.1.3. for boundary tones. 3.2 deals with the contents of the third level of transcription. Section 4 the presents the results of the reliability test that have been carried out on the labelling provided by the automatic transcriber. Section 5 contains the general conclusions of the paper.

1.1 The current transcribers

There have been several attempts to achieve computer-based automatic intonation transcribers. In fact, some intonation models and their automatic transcribers have been developed together. That is the case of the IPO and Aix-en-Provence's models, which consist of a number of phonetically based step-by-step rules for simplifying the F0 curve. During the first years, the stylization of the curve in these models was carried out manually, but soon computer programs for automatic stylization and transcription were created.

Although the attempts to achieve computer-based transcribers of intonation are numerous and the existing speech recognition systems are countless, major enterprises have not implemented the recognition of intonational phrases meaning in their software solutions. This is due to two main factors: (1) the difficulty in obtaining good pitch data and (2) the traditional distance between language engineers and linguists, which is magnified by the multiplicity of the theoretical models about prosody and its abstractness.

Automatic transcribers can be divided in two groups that we could call descriptive and applied. On one hand, there are the descriptive transcribers that make a transcription of a given utterance or text recorded in audio. The results of these kind of transcribers can be used for both theoretical research and automatic recognition of prosody. On the other hand, there are the transcribers that seek an application to speech synthesis. The transcriber we present in this article pertains to the first group insofar as the intonational transcription it offers is carried out automatically by means of a script based on a series of if–then conditions.

Automatic transcribers can also be divided according the data they use in order to perform the transcription, only linguistic knowledge (rule-based systems) and those which include statistical predictions. Some transcribers rely on the assignments made

by predictive statistical models, mostly regression and machine learning algorithms that base their assignment in both acoustic data and linguistic knowledge (parameters) and mathematical algorithms. In these cases a manual transcription of large and language specific oral corpora is used as a training corpora to create a model that has as its output the predicted transcription of the prosody of a written or oral text. The most developed systems in prosody recognition and synthesis nowadays (that is to say applied systems) use predictive statistical models in order to make prosody assignments. Between them there are models based on tree decision (Black and Hunt 1996; Lee et al. 2002) or Markov models (Wightman and Ostendorf 1994). Other examples based on the IPO model are the MOMEL algorithm which uses quadratic spline (Hirst and Espesser 1993), or the systems that use linear regression such as the one created by Rietveld (1984). The second technique for transcribing prosody uses only linguistic knowledge. They are rule-based systems where phoneticians learn how prosody works and design a series of condition (rules) that a utterance must match in order to be labelled in certain way. Most of them try to label events in a way that resembles human ear perception using knowledge about the intonational phonology of languages. In this group we can include the IPO approaches for Spanish (Garrido Almiñana 2008) and for French (Alessandro and Mertens 1995), which simulate the tonal perception. The tool we present in this paper belongs to the latter group.

Since the creation of the different ToBI systems (Silverman et al. 1992), many have tried to make easier and faster the work of the researchers who work with it by designing different tools that help with the tedious task of labelling every prosodic event of a sentence. In fact, the first AM based automatic transcriber was created before ToBI (Pierrehumbert 1983). The majority of such tools are programs that suggest to the human transcriber a set of labels, among which he/she has to choose. This way the task of transcribing is easier (Escudero-Mancebo et al. 2014; Syrdal et al. 2001), but the final decision of the label is still taken by the researcher. The other tools are fully-fledged automatic transcribers, in the sense that the program assigns the labels without human intervention. The first kind of transcribers can be called "prompters", while the second kind of transcribers can be called "labellers".

Due to the language specific character of ToBI systems, automatic transcribers designed up to now can be used only for the language they were conceived for. This has triggered the apparition of transcribers for many languages, such as those created for Korean (Kim et al. 2002), Japanese (Campbell 1996; Noguchi and Kiriyama 1999), Italian (Savino et al. 2002), and Swedish (Frid 1999). This notwithstanding, most of ToBI-based transcribers are meant for MAE_ToBI (Black and Hunt 1996; Ross and Ostendorf 1996; Sridhar 2008; Wagner 2008). Among them, probably the most complete transcriber is the one created by Rosenberg (2010), which does not predict only tones, but also stress and word boundaries. Out of the systems named only the one by Pierrehumbert (1983) is rule based, she designed a system where the system used the criteria she gave to it in order to classify tonal events.

As for Spanish and Catalan, little work has been previously done. The only Sp_ToBI transcriber is a "prompter" that was created with the purpose of labelling the statements (but not other types of sentences) contained in the closed corpus called Glissando (Escudero-Mancebo et al. 2014). Like many other ToBI transcribers (and speech system recognition systems in general), it relies mainly

on statistics for doing its transcriptions. It uses predictive modelling based on the fuzzy logic that predicts the best probability of a label to appear in a context given certain acoustic data and previous human assignments. Choosing the correct label is ultimately up to the researcher. So, nowadays for Spanish there is only a Sp_ToBI prompter, which only works wich declarative sentences, while for Catalan there are neither prompters nor labellers.

This article, thus, tries to fill the gap of ToBI automatic transcription systems in Spanish and Catalan. The transcriber we describe in this paper performs an intonational analysis based on linguistic features. The pitch assignments arise from acoustic data and the knowledge available of intonational phonology. The system foresees three levels of transcription of intonation: a surface or phonetic level, a deep or phonological level and a standardized phonological level, which will be detailed in following sections. It can label 13 prenuclear pitch accents, 15 nuclear pitch accents and 10 boundary tones at the surface level, which makes a total account of 150 nuclear configurations. At the deep level, it can label 9 prenuclear pitch accents, 8 nuclear pitch accents, and 10 boundary tones[1] (like the standardised tier), a total account of 80 nuclear configurations that are later standardised in 61 nuclear configurations. The system is able to recognise the patterns described in (Prieto and Roseano 2010), that means that it can recognise the patterns of all the following dialects: Castilian Spanish, Cantabrian Spanish, Canarian Spanish, Dominican Spanish, Puerto Rican Spanish, Venezuelan Andean Spanish, Ecuadorian Andean Spanish, Chilean Spanish, Argentinian Spanish, and Mexican Spanish. As for Catalan, the script recognises all the possible geographic varieties as described in (Prieto and Cabré 2013).

Furthermore, the script offers the researchers several options to customise their transcriptions for those pitch accents and nuclear configurations that are still being matter of controversy. In this way, the researcher can choose between, for example, the prenuclear accents L* + H and L + >H* for a rising pitch movement that begins in the stressed syllable and has its target in the post-tonic (Face and Prieto 2007; Roseano et al., accepted) or between L + ¡H* and L + H* for a pitch accent that rises over the extrahigh threshold (Borràs-Comes et al. 2014).

1.2 General theoretical framework

As it has been mentioned previously, the tool we present, which is called Eti_ToBI, is based on the Autosegmental-Metrical (AM) model (Pierrehumbert 1980), a theory that states that tones are phonological autosegments anchored to the prominent positions in the utterance (i.e., stressed syllables and boundaries). More specifically, Eti_ToBI is based on Sp_ToBI (Beckman et al. 2002; Face and Prieto 2007; Prieto and Roseano 2010) and Cat_ToBI (Prieto and Cabré 2013; Prieto 2014) conventions for tone transcription.

The ToBI notation systems are multi-tier transcription systems. General ToBI transcription guidelines (Beckman and Elam 1997) establish a first notation tier for

[1] Despite having the same number, the deep pitch accents for Catalan and Spanish differ. The tritonal accent L + H* + L never appears in Catalan, meanwhile ¡H + L* does not in Spanish.

the segmental transcription, a second one for tones, a third one for break indices and a fourth one for comments. The notation of the first tier can be automated with the traditional recognition of speech systems. As for the third notation tier, break indices, there are automatic segmenters and prosodic transcribers, which use lexical content and syntactic structure together with prosody for transcribing boundaries (Shriberg et al. 2000; Liu et al. 2000; Tür et al. 2001; Rosenberg 2010). In this paper, we focus on intonation transcription, which is, besides, the most complex level of analysis in terms of transcription.

The ToBI systems for Catalan and Spanish transcribe two types of intonational events: pitch accents and boundary tones. Pitch accents are anchored to the stressed syllable and a star (*) marks the tone that coincides with the stressed syllable. Boundary tones are related with the edge of the intermediate phrase (ip), which is marked with a dash (-) or with the end of intonational phrase (IP), which is marked by a percent sign (%). The intonational annotation in ToBI systems is a level-based approach, which is to say that it labels intonational events depending on the height of the pitch, high (H) or low (L). For Spanish and Catalan, besides the classical high and low levels, two more levels have been attested: an extrahigh level (¡H) for pitch accents and a mid level for boundary tones (!H). This way, the system can transcribe pitch movements like a succession of tones of different levels. The Spanish and Catalan systems include two kinds of pitch accents[2]: monotonal (T*) and bitonal (T* + T and T + T*). As for the boundary tones, they establish the existence of monotonal (T %), bitonal (TT %) and tritonal (TTT %) boundary tones. The combination of the last pitch accent of an utterance and the following boundary tone is usually called nuclear configuration.

It has been largely discussed whether the ToBI systems are really phonological and, if so, to what extent (Breen et al. 2012; Nolan and Grabe 1997; Siebenhaar and Leemann 2012, p. 33). It has been agreed, traditionally, that ToBI is conceived as a phonological system, but Pierrehumbert in her thesis (1980) already underlines that all phonological knowledge must be based on a phonetic reality and she will pick this idea again to apply it to the concept of Laboratory Phonology (Pierrehumbert 2000; Pierrehumbert et al. 2000).

Focus has been made that 'ToBI is not an International Phonetic Alphabet for prosody' (The Ohio State University Department of Linguistics 1999). The most significant consequence of that is that it varies from language to language. The same contour can be labeled differently in different languages depending on what is considered to be more relevant, i.e., what is considered phonological in a specific language (Hualde 2003, p. 180; Roseano and Fernández Planas 2013, p. 294). Yet, the acoustic phenomenon, the F0 contour, is the same independently from its relevance, so, at a strictly phonetic level, the same contour should be transcribed in a similar way, despite the phonological variation and this is something that does not happen in the current ToBIs, which are phonological intended. This contradiction is better understood with an example. Figure 1 contains two utterances, both showing the same F0 contour, consisting in a F0 rise till the mid of the stressed syllable, followed by a fall till the end of the utterance. However, these utterances belong to

[2] The tritonal pitch accent L + H* + L has been attested in Argentinian Spanish.

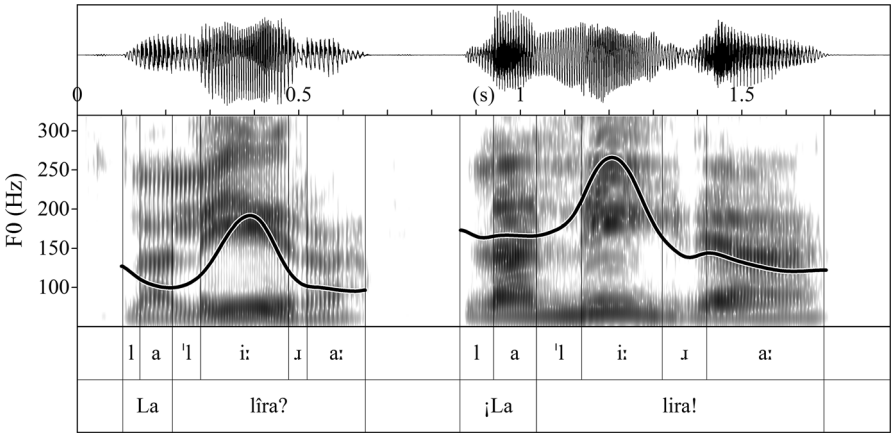

Fig. 1 F0 contour of the utterance "*La lira*?" 'The pound?' in Friulan and "*¡La lira!*" 'The lyre!' in Argentian Spanish. Both images show a rise and a fall in the stressed syllable

two different languages (Friulian and Argentinian Spanish) the Friulian utterance is a question and it is transcribed as L + ¡H* L %, meanwhile the Argentinian utterance is a narrow focus statement and transcribed as L + H* + L L %.

To solve the problem of the ambiguity of the phonetic versus. phonological transcriptions in ToBI, we propose 3 different levels of transcription instead of the traditional unique phonological level. Using different levels of transcription is not new to prosody studies: the models that have traditionally given more weight to the phonetic basis of intonational phonology, such as the IPO model and the Aix-en-Provence model (Hart and Collier 1975; Hirst et al. 2000), study intonation at two different levels (surface and deep). On the other hand, within an AM framework, despite the emphasis that has been traditionally made on the phonetic basis of intonational phonology and that Pierrehumbert herself talked about the "surface representation" and the "underlying form" (Pierrehumbert 1980:10), ToBI systems do not implement two levels of analysis since they are mainly interested phonological analysis. However, The Korean ToBI system (K-ToBI) includes the use of a phonetic tone tier and a phonological tone tier (Jun 2000) and distinguishing those two levels of analysis was what allowed Lee et al. (2002) to implement a speech synthesis method based on K-ToBI transcription. Nowadays a more phonetic transcription of intonation that permits to use the same labels for the same phonetic phenomena and, thus that can be used for different languages, is being discussed (Prieto and Hualde in press).

As far as the studies about the intonation of Spanish and Catalan are concerned, more phonetically based transcriptions have been used by different laboratories since a dual system—it consists of a surface and deep level of analysis- was proposed by Fernández Planas et al. (2002).

For the current purpose of developing an automatic transcriber for Spanish and Catalan, a distinction between three levels (represented in separate tiers) has proved itself indispensable in order to reach an accurate transcription. The labels we use in

the three levels are based on the AM model and the only difference between them is the level of abstraction in the transcription: the most detailed transcription is made at the phonetic surface level and the most simplified in the standardised phonological level. The tree level partition is entirely new to prosodic transcription but has proved itself necessary for an automatic transcription based on Sp_ToBI and Cat_ToBI systems.

2 Implementation of the system

In this section, we give some details about how Eti_ToBI works. Section 2.1. provides a description of the methods for determining F0 values. Section 2.2. gives an overview on the script code.

2.1 Pitch extraction method

One of the problems for the automatic recognition of intonation from acoustic data has been the correct recognition of the F0 (Wasserblat et al. 2008), which was usually hazy due to the low quality of the signal. The truth is that, although we do not yet have any reliable mechanism to extract the acoustic information of F0 contours, we have different techniques that permit making an F0 prediction (Boersma 1993; Cohen et al. 1995; Hermes 1988).

This script works in a Praat (Boersma and Weenink 2015) environment, that is to say that it uses the autocorrelation pitch extraction method proposed by (Boersma 1993). This technique is based on frequency analysis and works measuring the distance between the harmonics in the spectrum. Despite being one of the most recognised and used techniques for the F0 analysis, the autocorrelation method is not free from error. The omission or duplication of periods and the consideration of voiced frames unvoiced or unvoiced frames voiced are not uncommon (Jeng et al. 2011; Kotnik et al. 2009).

In order to minimise these possible mistakes the script defines accurately the pitch range of the speaker, which is usual in intonation analysis. The innovation contained in Eti_ToBI consists in the fact that the pitch floor and ceiling are set automatically for each utterance. This is performed following the two step technique exposed in Hirst (2011). The first step consists of extracting the pitch of the sentence with a wide range and looking for the minimum and maximum pitch of the sound. The second step consists of extracting a new pitch object in which the floor pitch is established by multiplying by 0.75 the first quartile of the found range and the ceiling, multiplying by 1.5 the 3rd quartile (De Looze 2010).

2.2 Script structure

In order to analyse prosody the script needs two files. The first one is the wave sound of the IP that is being analysed. The second is a Praat TextGrid with an interval tier where the boundaries are aligned with the start and end point of each syllable and where there is a mark in lexically stressed syllables. The tonicity mark

can be any Unicode symbol chosen by the researcher. Optionally, the researcher can include a with Break Indices information, if this information is available the script will use it to assign tones to intermediate phrases. The script will add to this TextGrid three more tiers with the prosodic transcription and save it.

The core script is applied to all files in a folder. It contains a loop for all the syllables of the IP, the IP is determined manually, being either the end of the sound file or a 4 break index (BI) in the TextGrid. When a stressed syllable is reached a series of if–then conditions run. When a condition is accomplished, either the label can be assigned directly or the pitch tone disambiguated. In order to disambiguate a movement the script applies a series of sub-conditions usually looking for the alignment of the targets, range, and range within the stressed syllable. After that, the labels in the first and second level of analysis are assigned simultaneously. In addition, when the whole sentence is analyzed the "forbidden" pitch accent-boundary tone combinations are substituted in the third level of analysis.

3 The three level partition

The three levels proposed to give an accurate transcription of intonation are the ones that follow: a surface level of analysis, which gives an account of the acoustic details of the F0 movements, a deep level, which reaches a phonetic-phonological transcription, and a standardised level, which contains the intonational units foreseen by the Sp_ToBI and Cat_ToBI conventions.

Since we use three tiers for intonational transcription instead of one tier as in classical ToBI systems, our labelling of an utterance will include 5 tiers. The researcher needs to introduce the information of the first two tiers, which contain the phonetic transcription and break indices, whereas the script fills in the other three, an example can be seen in Fig. 2.

In the first level of transcription (tier 2 or 3), that is to say the surface tier, the script uses the transcription conventions proposed in Roseano and Fernández Planas (2013). The labels used give an acoustic account of every F0 movement that can be perceived by human ear in a linguistic context. The result is a transcription that we could compare with a narrow phonetic transcription in segmental phonetics. It is, thus, not language related, since every movement is labelled using symbols (i.e., labels) that correspond to an acoustic description of the actual curve. We could say that the surface transcription is a kind of transliteration of the F0 contour using the AM framework[3] and it is what makes possible an automatic analysis.

The second level (deep tier) is the suprasegmental equivalent of the kind of labelling that, in segmental phonetics, would be a broad phonetic transcription. The main difference with the surface/phonetic level is that the acoustic data are interpreted according to the intonational phonology of the language analysed. This labelling is, thus, language related because the rules that govern the phonetic implementation of phonological tones vary from one language to another; in other

[3] Despite being theoretically universal, the script is optimised for the recognition of the movements of Spanish and Catalan systems since the alignment of pitch target varies from language to language.

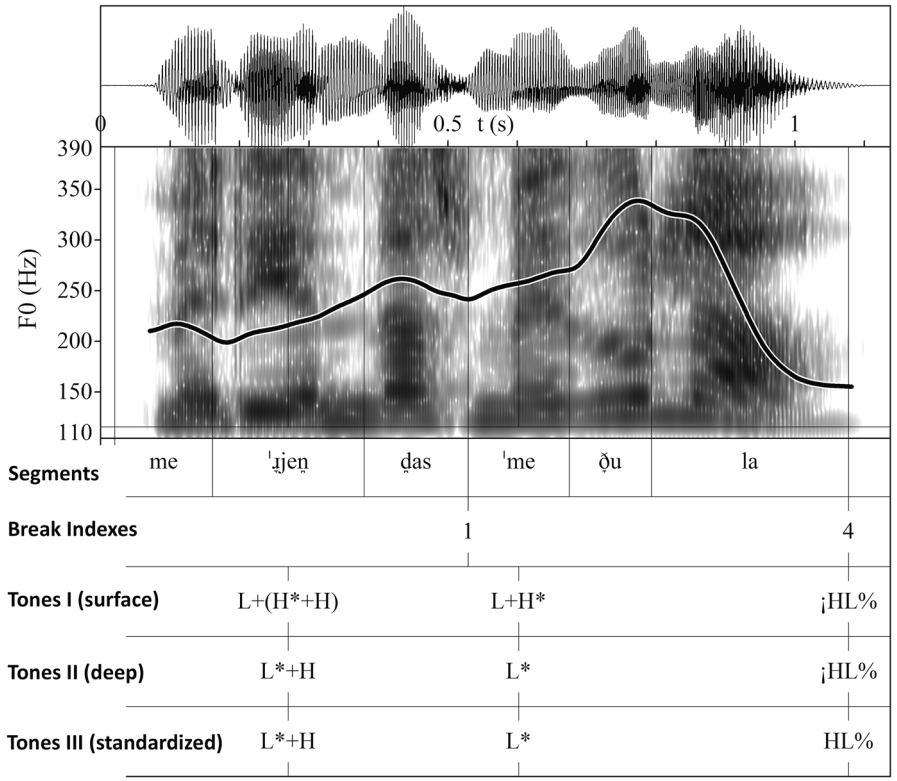

Fig. 2 Example of the output of the script, for the yes/no question '¿Meriendas médula?' 'Do you have medulla for your afternoon snack?' where the tiers have been labelled

words, the same phonetic fact has different phonological interpretations, depending on the language (as seen for Fig. 1. For this reason, when running the script, the researcher has to specify whether the language to be labelled is Catalan or Spanish. Since the number of labels used at this level is smaller than the number of labels used at the phonetic level, this level of transcription represents a further simplification of the representation of the data.

The third tier, which is also language related, standardises the transcription in order to make it coincide with the repertory of the current ToBI systems. It contains what can be considered as the best possible approximation to the phonological transcription. The basic difference between tier 2 and tier 3 lies in the nuclear configurations only: while the deep tier can contain all possible combinations of pitch accents and boundary tones, the standardized tier only contains the combinations of pitch accent and boundary tone (nuclear configurations) that are currently foreseen in the Sp_ToBI or Cat_ToBI systems, while prenuclear accents are the same that can appear in tier 2.

The progression from level 1 to level 3 represents a gradual approximation to the phonological representation, but one should bear in mind that Eti_ToBI is not

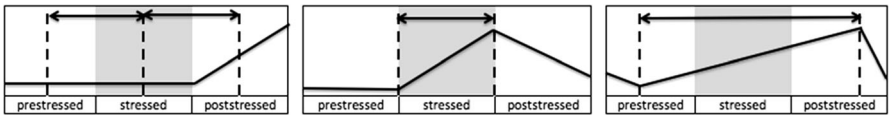

Fig. 3 Time points where the script looks for F0 differences in prenuclear accents. From *left* to *right*, in the first option, *values* are taken in the midpoints of prestressed, stressed and poststresssed syllable. In the *second*, at the *start* and the end point of the stressed syllable. In the *third box*, at the valley and the peak of the movement

always able to provide a fully phonological transcription. In fact, in some cases only a human being can go beyond acoustic data and label phonologically. The impossibility for Eti_ToBI to provide a phonological labelling appears, for example, in the cases of tonal truncation. In both Catalan and Spanish, in fact, boundary tones do not surface in some contexts (mostly when a word with stress on the ultimate syllable has no coda) (Colina 2009). If a phonological boundary tone does not surface acoustically, Eti_ToBI is not able to reconstruct it exactly because it works with acoustic data. More information about this phenomenon and its consequences for automatic transcription of prosody can be found in Roseano and Fernández Planas (2013).

3.1 The first and second tiers: from F0 values to labels

3.1.1 Prenuclear pitch accents

Once the F0 values have been determined according to the techniques mentioned above, Eti_ToBI proceeds with the labelling, in this order: surface tier, deep tier, standardized tier. The algorithms for prenuclear and nuclear pitch accents are slightly different.

For prenuclear pitch accents, the script calculates the differences in semitones between the midpoint of the pre-stressed, stressed and post-stressed syllable (see Fig. 3), as well the differences between the start and end points of the stressed syllable and the difference between the F0 valley and the F0 peak. Differences between the frequencies in two different points of the contour are calculated using the following logarithmic formula $(12/\log10(2))*\log10(f0_2/f0_1)$. The prenuclear pitch accents that the script can label are detailed in the Appendix, Fig. 9.

If there is no significant[4] movement between any of those points, a monotonal pitch accent will be assigned. The label chosen for the monotonal pitch accent (i.e., either L* or H*) will depend on the pitch value of the centre of the stressed syllable.

[4] In order to accomplish the objective of labelling the perceivable F0 movements, we follow the convention of a threshold of 1.5 semitones for a movement to be considered significant. The 1.5 semitones threshold has proved as the effective operative threshold for the perception of intonation for Spanish (Pamies et al. 2002) and also to other intonational languages (Rietveld and Gussenhoven 1985). The 1.5 threshold is also adopted in Roseano and Fernández Planas (2013).

For being labelled as high, the value needs to be above the high third of the range of the utterance.[5]

If there is movement, the script follows a series of rules that go from the least restrictive to the most restrictive to determine which label has to be assigned. For example: if from the F0 valley to the F0 peak, there is a difference greater than 1.5 semitones, a L* + H label applies but, if there is also a difference greater than 1.5 between the start point and the end point of the stressed syllable, a L + H* label is written. If the previous two are true, but the target of the movement is in the post-tonic syllable, the label will be L + >H*.[6]

If Eti_ToBI detects the presence of significant F0 movements around the stressed syllable, there are two possibilities: either there is only one significant movement (i.e., one rise or one fall) or there are two significant movements (i.e., fall-rise, rise-fall, two consecutive rises, two consecutive falls). If there is only one significant movement, a bitonal pitch accent is labelled. If there are two significant F0 movements, a tritonal accent is written in the first tier (i.e., the tier containing the phonetic transcription). It is important to note that tritonal pitch accents and the corresponding labels are not present in the standard phonological ToBI labelling systems, but have been introduced by phoneticians in order to give more acoustic details about tonal movements (Fernández Planas and Martínez Celdrán 2003; Martínez Celdrán and Fernández Planas 2003). In the case of the assignment of a tritonal accent, the label will include information about which movement is greater: the smaller F0 movement being indicated between parentheses. Thus, for example, the L + (H* + L) label means that there is a rising-falling movement where the rise is greater than the fall.

Eti_ToBI also makes use of some diacritic symbols on the labels that appear in the first tier, basically ¡ and !. The ¡ symbol is used to indicate a rise over 6 semitones[7] as superhigh (L + ¡H*/L¡H %[8]). In addition to this, as usual in ToBI systems, the (¡) symbol also is used to describe upstepping, which means that the ¡H is used to indicate that a pitch accent is scaled higher than the previous ones. Within Eti_ToBI this last usage of the symbol will be used only in the cases where there is a rise from a previous high point, i.e., a high plateau (¡H* and ¡H + L*/¡HL %).

The second tier (deep level) contains the labels for the same tonal events that have been labelled in the first tier. The difference, as mentioned above, lies in the fact that some implementation rules are taken into account, the most important being those related with tritonal accents and non-phonological falling movements.

[5] In the case that the pitch accent is not the first in the IP, and it has been a high target before, the script will label that accent as high if there has not been declination (i.e., a falling greater than 1,5 semitones) since the last target.

[6] The label will be L* + H if the researcher has specified in the form that he/she does not want that label in the transcription.

[7] The 6 semitone threshold has been recently suggested for Spanish and Catalan basing on the perception of phonological contrasts between a high and an extra high levels (Borràs-Comes et al. 2014; Vanrell 2011).

[8] The boundary tone L¡H % is not included in current ToBI systems. However, the script is able to contrast between two nuclear configurations only by having this label. The extrahigh tone contrasts in the script with a LH % that has been usually identified with LM % (Vanrell 2011). Of course, in the standardized tier L + H* LH % becomes L + H* L!H % and L + H* L¡H % becomes L + H* LH % as the conventions state.

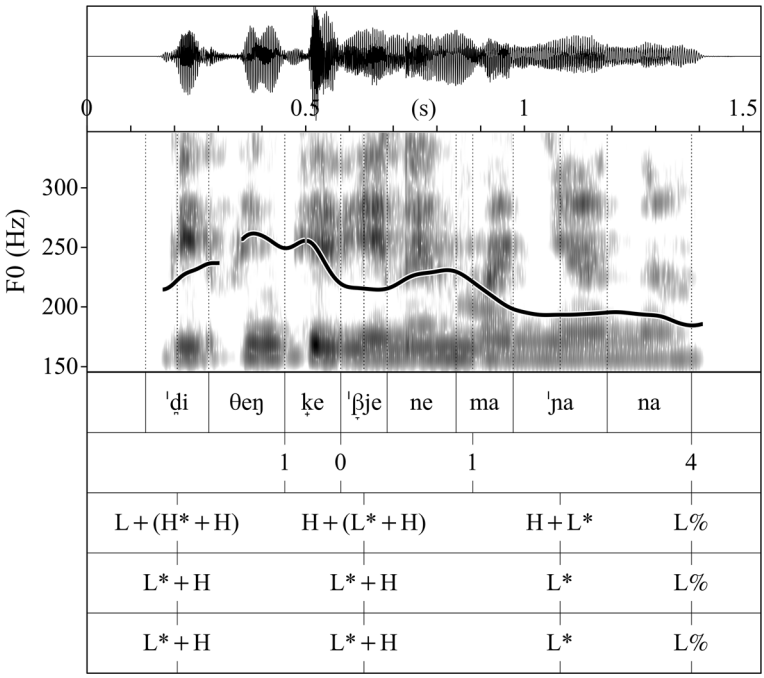

Fig. 4 F0 contour of the utterance "Dicen que vienen mañana" 'They say they are coming tomorrow' where the last pitch accent is labelled as $H + L^*$ in the first tier (a falling greater than 1.5 semitones from the midpoint of the presstresed syllable to the midpoint to the stressed syllable) and L^* in the second

Because of the implementation of these rules, only the pitch accent in the right columns of Fig. 9 appears in the second tier.

As far as tritonal accents are concerned, all phonetically tritonal pitch accents are transformed in phonologically bitonal pitch accents, except for Argentinian Spanish, whose phonological inventory includes tritonal accents.[9]

Another important set of rules has to do with non-phonological falling movements. One major problem with automatic labelling is that many falling movements are not real falling pitch accents but rather movements caused by declination.[10] Eti ToBI includes a set of algorithms aiming at avoiding the presence of such "false" pitch accents in the deep labelling. The situations where Eti_ToBI has been programmed to distinguish between real falling pitch accents and incidental F0 falls are: 1) deaccented pitch accents in the prenucleus (Ø), and 2)

[9] The script gives the possibility, in the initial form, to choose if the tritonal accent of Argentinian Spanish has to be used in the second tier or not.

[10] Declination is pitch natural tendency to decline from the beginning of an intonational phrase to the end.

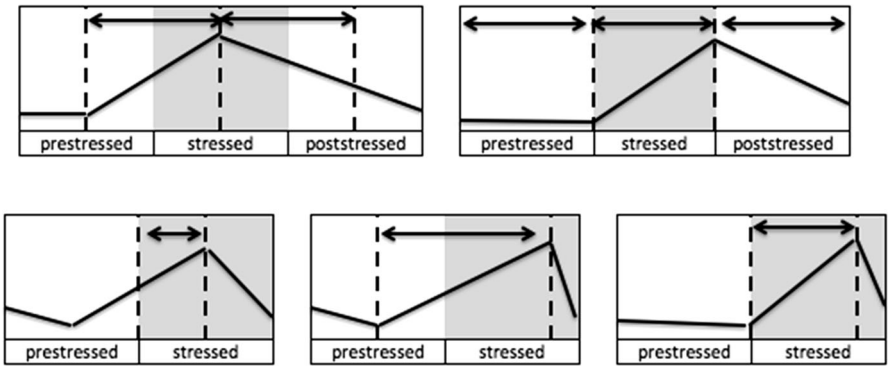

Fig. 5 Time points where the script looks for F0 differences in nuclear accents. From *left* to *right*, in the first option, *values* are taken in the midpoints of prestressed, stressed and poststresssed syllable. In the second, at the *start* and the *end point* of the prestressed, stressed and poststressed syllable. In the *third box*, between the midpoint of the prestressed syllable and the maximum pitch value of the stressed syllables and in the *third box*, between the *start point* of the stressed syllable and the maximum pitch value of that same syllable

"false falling" pitch accents in the nucleus (H + L* L %), which are actually low pitch accents (L* L %)[11](Fig. 4).

3.1.2 Nuclear pitch accents

The description made up to now is applicable to every prenuclear pitch accent, no matter its accentual position (oxytone, paroxytone, preparoxytione). Nevertheless, when it comes to nuclear pitch accents and boundary tones, the script goes through slightly different formulas depending on the accentual type, since if the last word is oxytone, special actions have to be taken.

For nuclear pitch accents, the time points where the script measures pitch values are either predefined taking into account the literature published on tonal alignment, both in the nucleus and prenucleus, for Catalan and Spanish (Prieto 2009; Prieto et al. 1995), or defined depending on the pitch peak (see Fig. 5).

The rules applied are different depending on the accentual type of the last word: for non-oxytone words the script will measure the pitch at 3 points of the pre-stressed syllable, 5 points in the stressed syllable and 5 points in the post stressed syllable; meanwhile for oxytone words the labels for the pitch accent and boundary tone are assigned together (see section below). Figure 10 shows the possible nuclear pitch accents.

[11] This conversion from the first tier H + L* to the second L* (Fig. 2) is possible in Spanish and Catalan because real H + L* tones consist of a fall within the accented syllable (Estebas-Vilaplana and Prieto 2010; Prieto 2014). For the script, this implies that a phonological H + L* must have a fall greater than 1.5 semitones within the start point and the end point of the stressed syllable.

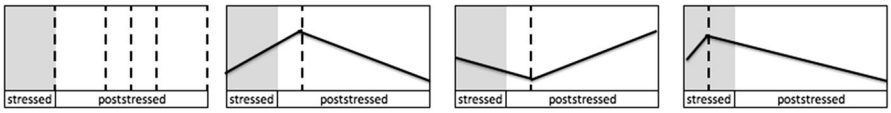

Fig. 6 Time points where the script takes F0 measures for calculating boundary tones. From *left* to *right*, the fixed points, 1/6, 2/6, 3/6, 4/6 and 6/6 of the poststressed syllables. The maximum *pitch value* of the first half of the poststressed, the minimum *pitch value* of the first half of the poststressed and the maximum *pitch value* of the stressed syllable

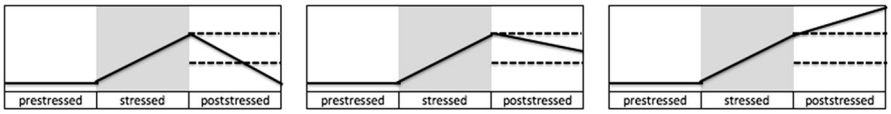

Fig. 7 Schematic representation of different tonal levels in boundary tones. From *left* to *right* *L + H* L %, L + H* !H %* and *L + H* H %*

3.1.3 Boundary tones

After transcribing the nuclear pitch accent, the script transcribes the boundary tone. In order to do so, the time range of the post-tonics is split into 6 equal parts and those are the six points where the F0 is read, the script also reads the F0 in the maximum pitch value of the stressed syllable, in the minimum pitch value of the post-tonics and in the maximum pitch value of the post-tonics (see Fig. 6). The script calculates the differences in semitones among those points and labels the boundary consequently. Thus, for example, a difference greater than 1.5 semitones between the first point of the post-tonics and the last point is transcribed as H %, if there are not other movements.

However, the boundary tones of Spanish and Catalan (like in other languages) do not consist only of two levels, but three: high, mid, and low level (Fig. 7). In order to distinguish between these three tonal levels, the range of the whole utterance is split into thirds. The rising movements that are lower than the 2nd third, and the falling movements that are higher than the 1st third are labelled as mid. Thus, if a significant final rise falls within the low/medium range of the utterance, the boundary is labelled as !H %.[12]

As we said before, the analysis goes slightly different for oxytone words. When the script detects that the last syllable of the IP is stressed, the nuclear configuration and the boundary tone are labelled together. Not having post-stressed syllables, the stressed syllable time range is split into 12 equal parts (see Fig. 8) and 1 value of the pre-stressed syllable and 8 values of the stressed syllable are measured. Some of

[12] In Spanish and Catalan intonational phonology the mid-level is not very productive, actually, it is reduced to the transcription of nuclear configurations of the vocative chant (L + H* !H %) and the emphatic obviousness statement (L + H* L!H %). These configurations have specific durations, and, in the case of the vocative chant, many other characteristics that make it recognisable (Ladd 2008: 135), that made that some other parameters such as duration were included in order to help to recognise the contours.

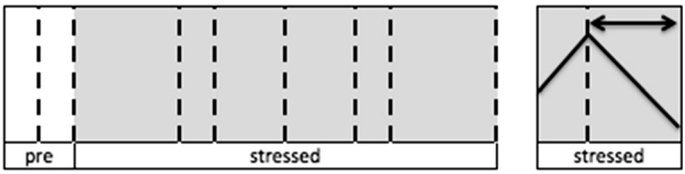

Fig. 8 Time points where the script takes F0 measures for calculating the nuclear configuration of oxytone tonemes. In the *left*, fixed time points: *midpoint* of the prestressed syllable, 3/12, 4/12, 6/12, 8/12, 9/12 and 12/12 of the stressed syllable. In the right, the maximum pitch in the stressed syllable

those values are considered part of the nuclear pitch accent and part of them are considered part of the boundary tone. Out from this data the formulas consider the movement that is taking place in both parts and they assign directly its nuclear configuration.

This solves the cases of compression and multiple alignments that are possible in oxytone words when they appear in nuclear position. It does not solve, though, the cases where the configuration is truncated, as mentioned before. In these cases, the transcription will reflect what is visible in the acoustic data and would be up to the researcher to reinstate the information that is not in the curve and neither predictable from it.

3.2 The standardised tier

Following the steps mentioned in the previous sections, the script is able to label the pitch movements throughout 60 independent formulas and 25 subformulas that are concatenated in a specific order. The final step of the script is the standardisation of labels.

The standardised tier differs from the second tier only in what concerns nuclear configurations. As mentioned in a previous section, this tier is a copy of the deep labelling, with the difference that it only contains the labels foreseen in the standard versions of Cat_ToBI and Sp_ToBI. This standardisation is necessary due to notation conventions and it has two main reasons. The first one is the existence in the current Cat_ToBI systems of pitch accents that are contrastive only in certain combinations. For instance L + H* contrasts phonologically with L + ¡H* when it is followed by a low boundary tone, but this contrast does not exist in the current ToBI between L + H* LH % and L + ¡H* LH %.[13] The second reason is merely conventional: Spanish and Catalan systems do not allow trailing tones in nuclear positions,[14] so they are moved to the boundary edge. This way, a rising tone with the target in the post-tonic syllable is transcribed in the prenucleus as L* + H, but when it comes to the nuclear position and it is followed by a low tone, the L* + H L % configuration needs to become L* HL %.

[13] A recent work proves that they are phonologically different but the change has not been integrated yet (Roseano, Fernández Planas, Elvira-García, and Martínez Celdrán 2015).

[14] Algherese Catalan has a H* + L L % nuclear configuration.

4 Reliability results

In this section, we present the results performed by the script compared with the answers given by human transcribers for the same samples. Different experiments were driven for Catalan and Spanish.

4.1 Catalan data results

An ad hoc corpus was created to test the correct functioning of the script. The corpus had 100 sentences pronounced by 20 different speakers, aged between 20 and 30 years, who had Catalan as their dominant language and spoke its central dialect, which is the most widely spoken and the basis for the standard variety, for the rest of variaties reliability tests had not been conducted. The utterances of the corpus show 9 different nuclear configurations and were elicited by DCT (Blum-Kulka 1982). The data were recorded by means of a Marantz PMD620 recorder and a Shure SM58 microphone. All the recordings were done at the Laboratory of Phonetics of the University of Barcelona.

Four different transcribers (among which the first author of this paper) transcribed phonologically the intonation of the utterances. The transcribers were experts who received a 30-minute training. They performed their labelling task without knowing that the results were being compared with an automatic transcriber. Their transcriptions were later compared with those obtained by the automatic transcriber.

The labels obtained in the third tier of the script have been compared with the ones written by each one of the human transcribers. Since different formulas run for different tonal events, the results compared are classified in nuclear pitch accents, and boundary tones for each transcriber. The statistical analysis has been carried out with the technique of Cohen's kappa which is suitable for measuring the level of agreement between two individuals (J. Cohen 1968; Jacob Cohen 1960) and calculated with an on-line calculator (GraphPad 2014). The values range from -1 (total disagreement) to 1 (total agreement) and 0 will be the chance level. Generally speaking, a kappa .60 is considered good.

The reliability test shows (Table 1) different level of agreement depending on the transcriber. However, all of them are beyond 75 % of agreement in the case of nuclear pitch accent results (NPA) and beyond the 90 % for boundary tones (BT). The kappa for these values of agreement are considered either good or very good.

The mean level of agreement (in percentage) is similar to the one performed for multiclass pitch accent and boundary tone labelling by human transcribers between them (Escudero et al. 2012; Jun et al. 2010; Pitrelli et al. 1994; Ann K. Syrdal and McGory 2000). Thus, Eti_ToBI labelling results are comparable with the results obtained by human transcribers.

Table 1 Level of agreement and reliability scores between Eti_ToBI and four different human transcribers

Tonal event	n	Agreement (%)	Kappa	Assessment
NPA 1	98	85.71	0.772	Good
NPA 2	98	85.71	0.770	Good
NPA 3	98	82.65	0.722	Good
NPA 4	98	78.79	0.657	Good
BT 1	98	92.86	0.884	Very good
BT 2	98	92.86	0.885	Very good
BT 3	98	93.88	0.900	Very good
BT 4	98	90.82	0.851	Good

4.2 Spanish data results

For Spanish the corpus was based on the emission of 1186 sentences produced by 4 speakers from 4 different varieties of peninsular Spanish. All of them had a range of age between 21 and 28 years old and were representative subjects of the dialect they spoke. The varieties chosen were Cantabrian Spanish (Ruiloba) with 286 utterances, Castilian Spanish (Madrid) 300 utterances, Spanish spoken in Catalonia (Barcelona) 300 utterances and Andalusian Spanish (Seville) 300 utterances. The corpus was semispontaneous speech elicited with a variation of DCT (Blum-Kulka 1982). The data were recorded by means of a Marantz PMD620 recorder and a Shure SM58 microphone.

One human transcriber (the first author of this paper, and first transcriber of the previous test) carried out the intonational transcription of all the sentences of the corpus. The decision to make use only of one human transcriber instead of four was due to the size of the Spanish corpus. The fact of using only one transcriber does not represent a methodological problem, for the transcribers had shown a high level of agreement among them in the previous tests with Catalan data. In fact, the intertranscriber agreement (Fleiss 1971) was tested for the Catalan results between the four human transcribers and they scored 85 % of agreement and a *Free-marginal kappa* of .84 (Randolph 2008). Thus, we considered that all four transcribers were valid transcribers and any of those transcribers could carry out the transcription of the Spanish corpus.

The transcription of the labeller was compared with the one made by Eti_ToBI. Since different formulas run for different tonal events, the results compared are classified in (1) prenuclear pitch accents, (2) nuclear pitch accents, and (3) boundary

Table 2 Level of agreement and reliability scores between Eti_ToBI and the human transcriber

Tonal event	n	Agreement (%)	Kappa	Assessment
PPA	1660	94.94	.907	Very good
NPA	1186	88.11	.831	Very good
BT	1186	81.28	.756	Good

tones. All of them have an agreement of >80 % and kappa values beyond .7 (Table 2).

These results are again similar to the ones performed by expert human transcribers (Pitrelli et al. 1994; Syrdal and McGory 2000; Jun et al. 2010; Escudero et al. 2012). Thus, Eti_ToBI proves itself an effective tool for transcribing Catalan and Spanish intonation, whose performance equals the results obtained by trained human transcribers.

5 Conclusions

Eti_ToBI is an automatic tool for the analysis if intonation. It is the first automatic intonation transcriber based on ToBI developed for Catalan, and the only one in Spanish that is a transcriber, and not only a prompter. In both cases, the script is ready to work with utterances belonging to different geographic varieties and with utterances with a wide range of contours.

Eti_ToBI is a Praat script under GNU license, which makes it usable for any phonetician on any operating system and provides the confidence of working with a well-known pitch detection algorithm. Furthermore, the only prerequisite to use the script is having a TextGrid with a tier that indicates the boundaries of the syllables and the stressed syllable (a process that can also be made automatically with other scripts).

The first level of labelling provides an objective, transparent and universal transcription based only on acoustic data. This proves that a standard (non-language related) ToBI method for analysing prosody is possible and therefore applicable to automatic recognition systems.

As far as the third level of labelling is concerned (the phonological and language-specific one), Eti_ToBI has proved to be as reliable as an expert human transcriber, for both Catalan and Spanish data.

Acknowledgments This work has been funded by a grant awarded by the Spanish government FFI2012-35998 for the AMPER-CAT project and the predoctoral grant APIF-2012 of the University of Barcelona.

Appendix

See Figs. 9, 10 and 11.

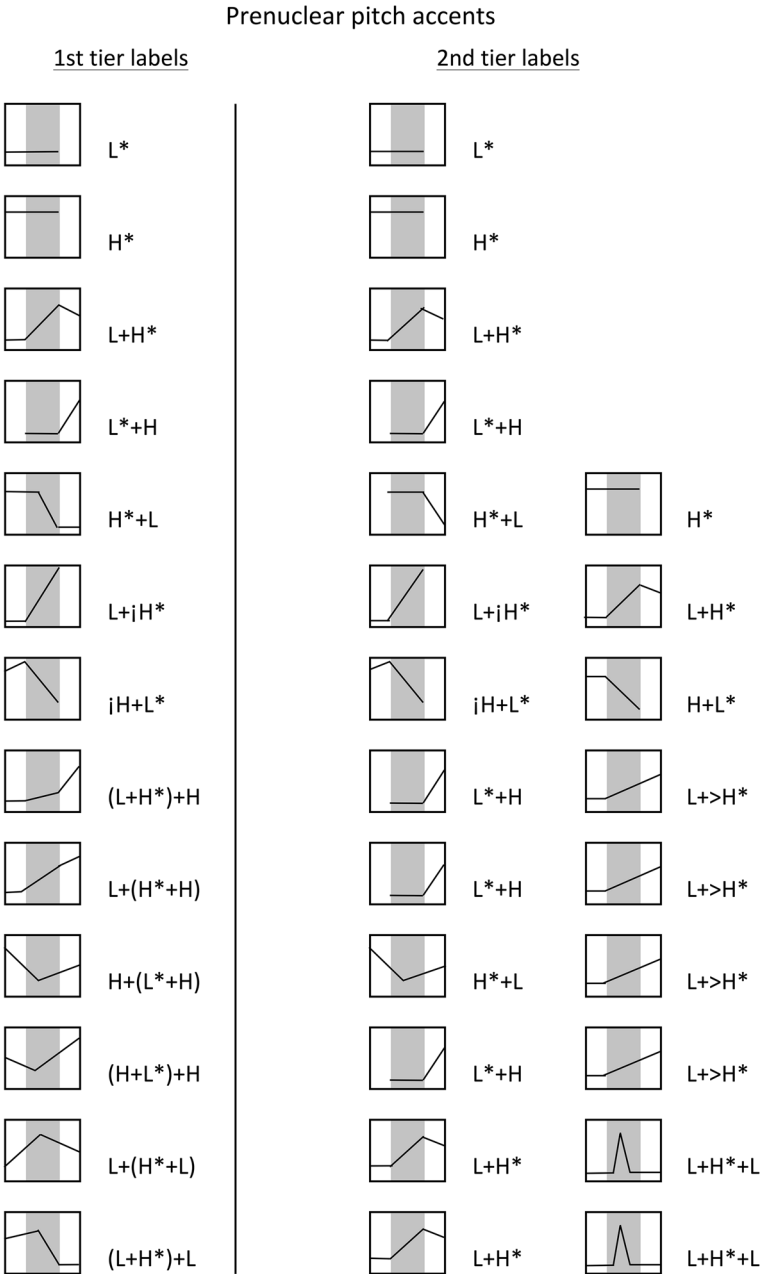

Fig. 9 Prenuclear pitch accents detected by the script in the first tier (*left*) and its equivalencies in the second tier (*right*). When two choices are given in the second column it indicates that both transcriptions are possible depending on what options and language the researcher has chosen in the form of the script

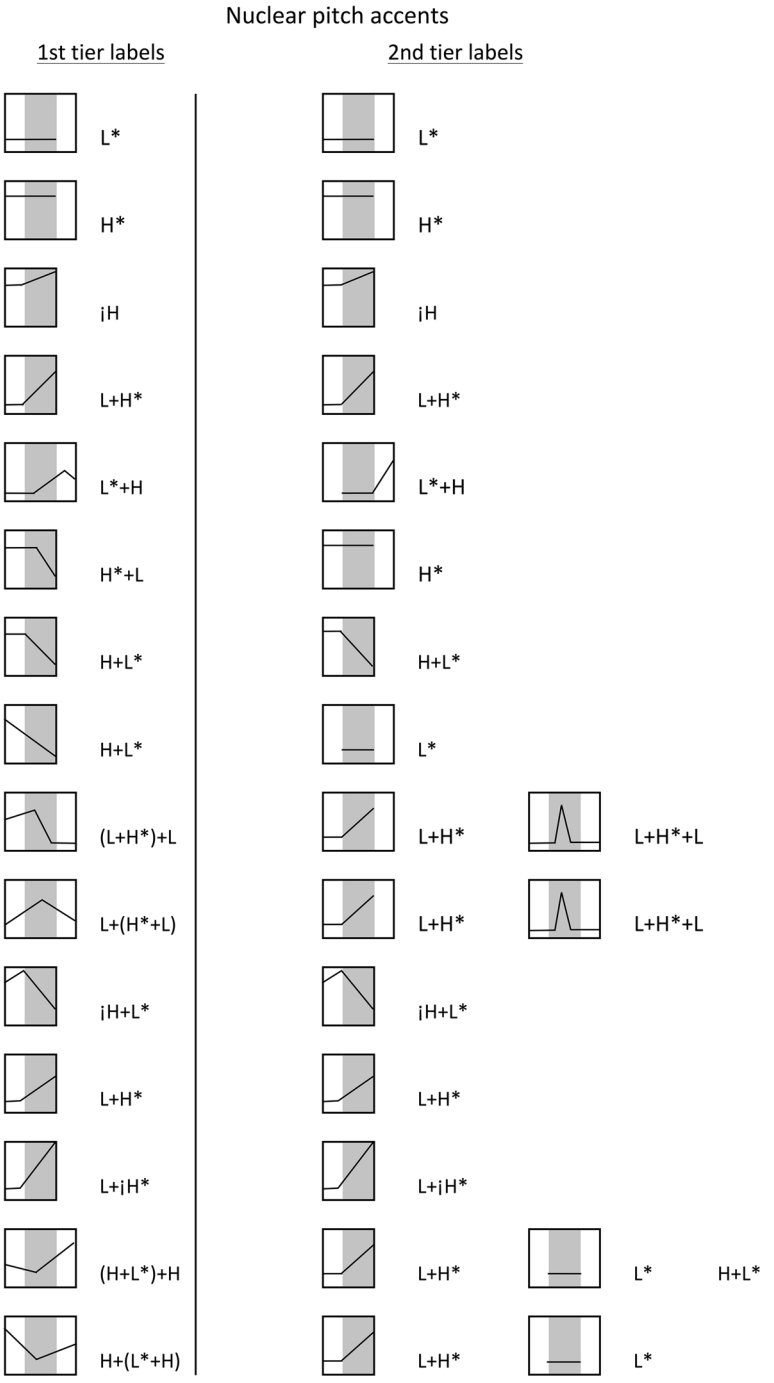

Fig. 10 Schematic representation and labelling of the possible nuclear pitch accents detectable by the script and its equivalence in the second tier

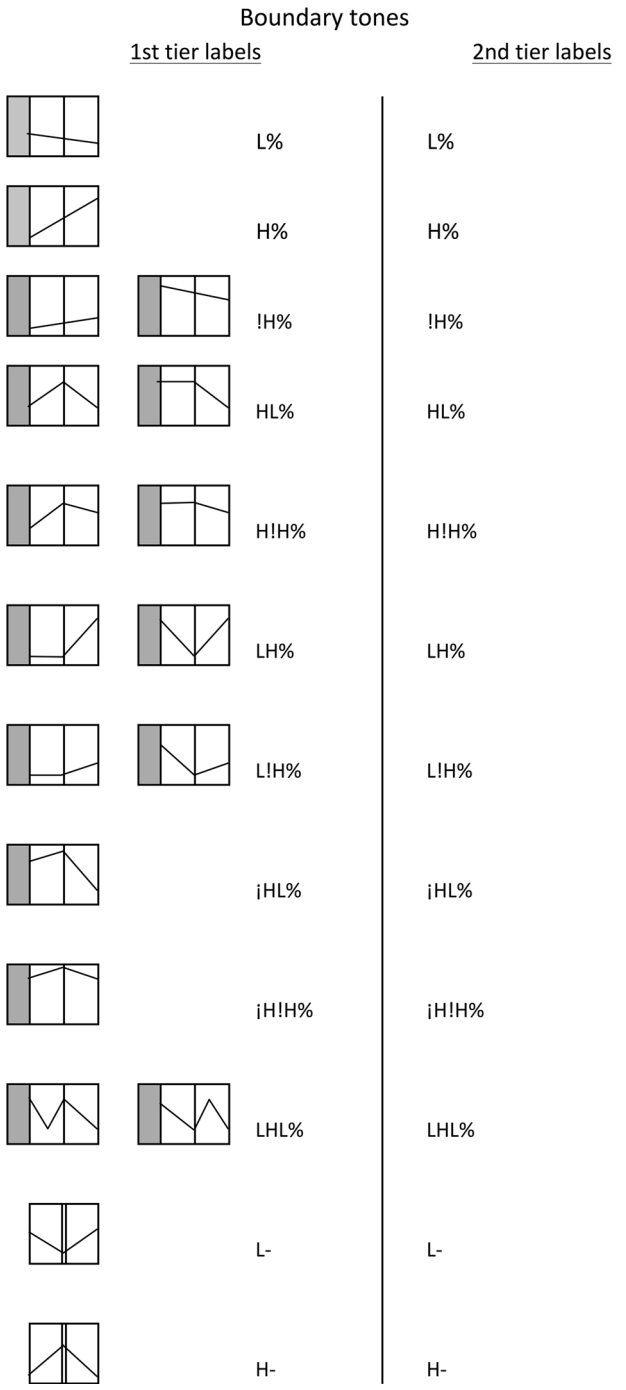

Fig. 11 Schematic representation and labelling of the possible boundary tones detectable by the script and its equivalence in the second tier

References

Alessandro, C., & Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language, 9*(3), 257–288.

Beckman, M., Díaz-Campos, M., McGory, J. T., & Morgan, T. A. (2002). Intonation across Spanish, in the tones and break indices framework. *Probus, 14*, 9–36. doi:10.1515/prbs.2002.008.

Beckman, M., & Elam, G. A. (1997). Guidelines for ToBI Labelling. The Ohio State University Research Foundation.

Black, A. W., & Hunt, A. J. (1996). Generating F 0 contours from ToBI labels using linear regression. In *ICSLP 96. Fourth International Conference on Spoken Language Proceedings* (pp. 1385–1388). Philadelphia: IEEE. doi:10.1109/ICSLP.1996.607872.

Blum-Kulka, S. (1982). Learning to Say What You Mean in a Second Language: A Study of the Speech Act Performance of Learners of Hebrew as a Second Language1. *Applied Linguistics*, *3*(1), 29–59. http://applij.oxfordjournals.org/content/III/1/29.short. Accessed January 21 2015.

Boersma, P. (1993). Acurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA Proceedings 17* (pp. 97–110). http://www.fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer. http://www.praat.org/.

Borràs-Comes, J., Vanrell, M. del M., & Prieto, P. (2014). The role of pitch range in establishing intonational contrasts. *Journal of the International Phonetic Association*, *44*(01), 1–20. http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=9212002&fileId=S0025100313000303. Accessed April 7 2014.

Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. (2012). Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory*, *8*(2), 277–312. http://www.isca-speech.org/archive_open/int_97/inta_259.html. Accessed November 17 2014.

Campbell, N. (1996). Autolabelling Japanese ToBI. In *ICSLP 96. Fourth International Congress on Conference on Language Processing Proceedings* (Vol. 4, pp. 2399 – 2402). Philadelphia: IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=607292. Accessed September 3 2014.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. http://psycnet.apa.org/journals/bul/70/4/213/. Accessed July 18 2014.

Cohen, M. A., Grossberg, S., & Wyse, L. L. (1995). A spectral network model of pitch perception. *The Journal of the Acoustical Society of America*, *98*(2 Pt 1), 862–79. http://www.ncbi.nlm.nih.gov/pubmed/7642825. Accessed July 1 2015.

De Looze, C. (2010). *Analyse et interprétation de l'empan temporel des variations prosodiques en français et en anglais*. Aix-en-Provence. Retrieved from http://halshs.archives-ouvertes.fr/tel-00470641/.

Dorta, J. (Ed.). (2013). *Estudio comparativo preliminar de la entonación de Canarias, Cuba y Venezuela*. Madrid-Sta Cruz de Tenerife: La Página ediciones.

Escudero, D., Aguilar, L., Vanrell, M. del M., & Prieto, P. (2012). Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labeling system. *Speech Communication*, *54*(4), 566–582. http://www.sciencedirect.com/science/article/pii/S0167639311001749. Accessed April 7 2014.

Escudero-Mancebo, D., González-Ferreras, C., Vivaracho-Pascual, C., & Cardeñoso-Payo, V. (2014). A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling. In *Computer Speech & Language* (Vol. 28, pp. 326–341). doi:10.1016/j.csl.2013.08.001.

Estebas-Vilaplana, E., & Prieto, P. (2010). *Castilian Spanish intonation* (pp. 17–48). Lincom Europa, München: Transcription of Intonation of the Spanish Language.

Face, T., & Prieto, P. (2007). Rising accents in Castilian Spanish: a revision of Sp-ToBI. *Journal of Portuguese Linguistics, 6*(1), 117.

Fernández Planas, A. M., & Martínez Celdrán, E. (2003). El tono fundamental y la duración: dos aspectos de la taxonomía prosódica en dos modalidades de habla (enunciativa e interrogativa) del español. *Estudios de fonética experimental*, *12*, 166–200. http://www.raco.cat/index.php/EFE/article/viewArticle/140007/0. Accessed April 7 2014.

Fernández Planas, A. M., Martínez Celdrán, E., Salcioli Guidi, V., Toledo, G., & Castellví Vives, J. (2002). Taxonomía autosegmental en la entonación del español peninsular. In *Actas del II Congreso de Fonética Experimental* (pp. 180–186). Sevilla.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. doi:10.1037/h0031619.

Frid, J. (1999). An environment for testing prosodic and phonetic transcriptions. In *Proceedings of ICPhS 99* (pp. 2319–2322). San Francisco. http://lup.lub.lu.se/record/529087/file/1624474.pdf. Accessed September 3 2014.

Garrido Almiñana, J. M. (2008, April 28). Modelling Spanish Intonation for Text-to-Speech Applications. Universitat Autònoma de Barcelona. http://www.tdx.cat/handle/10803/4885. Accessed July 3 2014.

GraphPad. (2014). QuickCalcs. http://graphpad.com/quickcalcs/kappa1/. Accessed January 6 2014.

Hart, J. t', & Collier, R. (1975). Integrating Different Levels of Intonation Analysis. *Journal of Phonetics*, *3*(4), 235–255. http://eric.ed.gov/?id=EJ127873. Accessed September 2 2014.

Hermes, D. (1988). Measurement of pitch by subharmonic summation. *The journal of the acoustical society of America*, *83*(1), 257–264. http://scitation.aip.org/content/asa/journal/jasa/83/1/10.1121/1.396427. Accessed July 16 2015.

Hirst, D. (2011). The analysis by synthesis of speech melody: from data to models. *Journal of Speech Sciences*, *1*(1), 55–83. http://www.journalofspeechsciences.org/index.php/journalofspeechsciences/article/viewArticle/21.

Hirst, D., Di Cristo, A., & Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. *Prosody: theory and experiment* (pp. 51–88). Dordrecht: Kluwer.

Hirst, D., & Espesser, R. (1993). Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function. *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, 75–85.

Hualde, J. I. (2003). El modelo métrico y autosegmental. In P. Prieto (Ed.), *Teorías de la entonación* (pp. 155–181). Barcelona: Ariel.

Jeng, F., Hu, J., Dickman, B., & Lin, C. (2011). Evaluation of two algorithms for detecting human frequency-following responses to voice pitch. *International Journal of audiology*, *50*(1), 14–26. http://www.tandfonline.com/doi/abs/10.3109/14992027.2010.515620. Accessed September 16 2015.

Jun, S.-A., Lee, S., Kim, K., & Lee, Y. (2010). Labeler agreement in transcribing korean intonation with K-ToBI. In *Interspeech'10* (pp. 211–214). http://www.linguistics.ucla.edu/people/jun/ICSLP-KtobiAgree.pdf. Accessed December 6 2014.

Kim, B., Lee, J., & Lee, G. (2002). Corpus-based Pitch Prediction based on K-ToBI Representation. In *ACM Transactions on Asian Language Information Processing (TALIP)* (Vol. 1, pp. 207–224). ACM New York, NY, USA. doi:10.1145/772755.772757.

Kotnik, B., Höge, H., & Kačič, Z. (2009). Noise robust F0 determination and epoch-marking algorithms. *Signal Processing, 89*(12), 2555–2569. doi:10.1016/j.sigpro.2009.04.017.

Ladd, D. R. (2008). *Intonational phonology Cambridge* (2nd ed., Vol. 2). New York: Cambridge University Press.

Lea, W. (1980). Prosodic aids to speech recognition. In W. Lea (Ed.), *Trends in Speech Recognition* (pp. 166–205). Englewood: Prentice-Hall.

Lee, J., Kim, B., & Lee, G. (2002). Automatic corpus-based tone and break-index prediction using K-ToBI representation. *ACM Transactions on Asian Language Information Processing (TALIP), 1*(3), 207–224. doi:10.1145/772755.772757.

Liu, M., Xu, B., Hunng, T., Deng, Y., & Li, C. (2000). Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling. In *Proceedings of Acoustics, Speech, and Signal Processing, ICASSP 2000* (pp. 1025–1028). Washington, DC.

Martínez Celdrán, E., & Fernández Planas, A. M. (2003). Taxonomía de las estructuras entonativas de las modalidades declarativa e interrogativa del español estándar peninsular según el modelo AM en habla de laboratorio. In E. Herrera & P. Martín (Eds.), *La tonía: dimensiones fonéticas y fonológicas* (pp. 267–294). México D.F.: El Colegio de México.

Noguchi, H., & Kiriyama, K. (1999). Automatic labeling of Japanese prosody using J-ToBI style description. In *EUROSPEECH'99. Sixth European Conference on Speech Communication and Technology* (pp. 2259–2262). http://20.210-193-52.unknown.qala.com.sg/archive/archive_papers/eurospeech_1999/e99_2259.pdf. Accessed September 3 2014.

Nolan, F., & Grabe, E. (1997). Can "ToBI" Transcribe Intonational Variation in British English? In *Intonation: Theory, Models and Applications* (pp. 259–262). Athens, Greece. http://www.isca-speech.org/archive_open/int_97/inta_259.html. Accessed November 17 2014.

Pamies, A., Fernández Planas, A. M., Martínez Celdrán, E., Ortega-Escandell, A., & Amorós Cespedes, M. C. (2002). Umbrales tonales en español peninsular. In *Actas del II Congreso de Fonética Experimental* (Vol. Sevilla, pp. 272–278).

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. Cambridge, Massachusetts: MIT.

Pierrehumbert, J. (1983). Automatic recognition of intonation patterns. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics* (pp. 85–90). http://dl.acm.org/citation.cfm?id=981328. Accessed December 1 2014.

Pierrehumbert, J. (2000). The phonetic grounding of phonology. *Bulletin de la communication parlée, 5*, 7–23.

Pierrehumbert, J., Beckman, M. E., & Ladd, D. R. (2000). *Conceptual foundations of phonology as a laboratory science* (pp. 273–304). Phonological knowledge: Conceptual and empirical issues.

Pitrelli, J. F., Beckman, M. E., & Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the tobi framework. *ICSLP*. http://20.210-193-52.unknown.qala.com.sg/archive/archive_papers/icslp_1994/i94_0123.pdf. Accessed July 13 2014.

Prieto, P. (2009). Tonal alignment patterns in Catalan nuclear falls. *Lingua, 119*(6), 865–880.

Prieto, P. (2014). The intonational phonology of Catalan. In S.-A. Jun (Ed.), *Prosodic typology* (Vol. 2, pp. 43–80). Oxford: Oxford University Press. http://www.elebilab.com/documentos/archivos/publicaciones/3_GGT-08-04.pdf. Accessed August 26 2014.

Prieto, P., & Cabré, T. (Eds.). (2013). *L'entonació dels dialectes catalans*. Rubí: Publicacions de l'Abadia de Montserrat.

Prieto, P., & Hualde, J. I. (n.d.). Towards an international phonetic alphabet. *Laboratory Phonology*. (in press)

Prieto, P., & Roseano, P. (Eds.). (2010). *Transcription of Intonation of the Spanish Language*. München: Lincom Europa.

Prieto, P., van Santen, J., & Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics, 23*(4), 429–451.

Randolph, J. J. (2008). Online Kappa Calculator. http://justus.randolph.name/kappa.

Rietveld, A. C. M. (1984). *Syllaben, klemtonen en de automatische detectie van beklemtoonde syllaben in het Nederlands*. Université de Nijmegen.

Rietveld, T., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics, 13*, 299–308.

Roseano, P., & Fernández Planas, A. M. (2013). Transcripció fonètica i fonològica de l'entonació: una proposta d'etiquetatge automàtic. *Estudios de fonética experimental*, *XXII*, 275–332. http://www.raco.cat/index.php/EFE/article/view/275413. Accessed July 18 2014.

Roseano, P., Fernández Planas, A. M., Elvira-García, W., & Martínez Celdrán, E. (2015). *Els tons de continuació en parla espontània: Descripció i transcripció*. Barcelona: VII Workshop sobre la prosòdia del català.

Rosenberg, A. (2010). AuToBI - a tool for automatic ToBI annotation. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association* (pp. 146–149). Mihama, Japan. http://eniac.cs.qc.cuny.edu/andrew/papers/autobi-is10.pdf. Accessed August 26 2014.

Roseano, P., Fernández Planas, A. M., Elvira-García, W., Cerdà Massó, R., & Martínez Celdrán, E. (accepted). Caracterització acústica dels accents prenuclears de les interrogatives absolutes i les declaratives neutres en català central. *Estudios de Fonética Experimental*, XXV.

Ross, K., & Ostendorf, M. (1996). Prediction of abstract prosodic labels for speech synthesis. *Computer Speech & Language*, *10*(3), 155–185. http://www.sciencedirect.com/science/article/pii/S0885230896900108. Accessed October 29 2014.

Savino, M., Refice, M., & Daleno, D. (2002). Methods and Tools for Prosodic Analysis of a Spoken Italian Corpus. In *Proceedings of the I International Conference on Language Resources and Evaluation* (pp. 307–312). http://lrec-conf.org/proceedings/lrec2002/pdf/101.pdf. Accessed September 8 2014.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., & Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, *32*(1), 127–154.

Siebenhaar, B., & Leemann, A. (2012). Methodological reflections on the phonetic-phonological continuum, illustrated on the prosody of Swiss German dialects. In A. Ender, A. Leemann, & B. Wälchli (Eds.), *Methods in Contemporary Linguistics* (Vol. 247, pp. 21–44). Berlin: Walter de Gruyter. http://books.google.es/books?hl=es&lr=&id=cf8YDeYvBuQC&oi=fnd&pg=PA21&dq=This+system+has+been+formalized+in+the+ToBI+transcription+sys-+tem.+…+phonetic-+phonological+continuum,+illustrated+on+the+prosody+of+Swiss+German+dialects&ots=cIfe-1AYbo&sig=M9W96TM_PcPLCC49gwaKEGURcg0. Accessed November 17 2014.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., et al. (1992). ToBI: A Standard for Labeling English Prosody. In M. M. H. and G. E. W. J. J. Ohala, T. M. Nearey, B. L. Derwing (Ed.), *ICSLP 92 Proceedings 1992 International Conference on Spoken Language Processing. Volume 2* (pp. 867–870.). Department of Linguistics, University of Alberta.

Sridhar, V. (2008). Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(4), 797–811. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4453862. Accessed April 7 2014.

Syrdal, A. K., Hirschberg, J., McGory, J., & Beckman, M. (2001). Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*, *33*(1), 135–151. http://www.sciencedirect.com/science/article/pii/S016763930000073X. Accessed April 7 2014.

Syrdal, A. K., & McGory, J. T. (2000). Inter-transcriber reliability of toBI prosodic labeling. *INTERSPEECH, 2000*, 235–238.

Tatham, M., & Morton, K. (2005). *Developments in Speech Synthesis*. John Wiley & Sons. http://books.google.com/books?id=6mPk1Dkt_V0C&pgis=1. Accessed November 17 2014.

The Ohio State University Department of Linguistics. (1999). ToBI. http://www.ling.ohio-state.edu/∼tobi/. Accessed August 9 2014.

Tür, G., Hakkani-Tür, D., Stolcke, A., & Shriberg, E. (2001). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, *27*(1), 31–57.

Vanrell, M. del M. (2011). *The phonological relevance of tonal scaling in the intonational grammar of Catalan*. Universitat Autònoma de Barcelona.

Wagner, A. (2008). Automatic labeling of prosody. In *Proceedings of the 2nd ISCA Workshop on Experimental Linguistics, ExLing 2008* (pp. 25–27). Athens, Greece. http://isca-speech.org/archive_open/archive_papers/exling2008/exl8_221.pdf. Accessed September 3 2014.

Wasserblat, M.., Gainza, M.., Dorran, D.., & Domb, Y.. (2008). Pitch tracking and voiced/unvoiced detection in noisy environment using optim at sequence estimation. In *IET Irish Signals and Systems Conference* (pp. 43–48). Galway, Ireland.

Wightman, C., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. In *IEEE Transactions on Audio, Speech, and Language Processing* (Vol. 2, pp. 469–481). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=326607. Accessed November 17 2014.