Procedure for assessing the reliability of prosodic judgements using Sp-TOBI labeling system

David Escudero-Mancebo[1], *Lourdes Aguilar* [2]

[1]Department of Computer Science, Universidad de Valladolid
[2]Department of Hispanic Studies, Universidad Autonoma de Barcelona
descuder@infor.uva.es, Lourdes.Aguilar@uab.cat

Abstract

This paper reports on the results of a pilot study that was run to assess the labeling consistency of the proposed approach in Sp-ToBI before starting a large-scale production of annotations in the project Glissando. This test should serve to refine the model and to maintain consistently the annotation conventions across transcription sites. The Spanish ToBI labeling system has been proved as an effective system to annotate intonation for Spanish, although the annotation conventions across transcribers require a broader consensus. This is specially needed in the following pitch accents: high pitch accent (H*) vs rising pitch accent (L+H*), downstepped pitch accents versus non-downstepped counterparts, and mid tones. A related issue is the difficulty to decide in a very low pitch range if a tone is present or if the syllable has been unaccented. Moreover, the statistical procedures will shed light on the most confusable tones suggesting new approaches for the automatic prediction of ToBI labels in a Spanish Spoken corpus.

1. Introduction

The development of speech processing techniques is permitting the treatment of increasingly huge amounts of data, but the creation of speech corpora is extremely costly in terms of both time and resources, and requires manual intervention by experts who transcribe or review the collected material. This is especially true of prosodic annotation within an autosegmental-metrical framework ([1]).

Among the existing corpora of Spanish with prosodic information are those collected for the MULTEXT project, about 50 minutes of speech that has been annotated within the MOMEL/INTSINT theoretical framework ([2]), and for the C-ORAL-ROM project –about one million words for Spanish, taken from different speech situations that includes prosodic segmentation in tonal units, aligned with the speech signal. On the other hand, the main goal of the AMPER initiative is to study the prosody of the Romance languages and to reflect the results in maps that will be accessible, visually and perceptually, in the website http://www.ub.es/labfon/amper/.

As far as the use of ToBI-framework system in Romance languages is concerned, a considerable amount of data of spoken Italian varieties is collected in the AVIP (Archivio di Varietà di Italiano Parlato) corpus of spontaneous Map Task dialogues, whose intonation phenomena have been labelled according to a ToBI-based approach, and the project "Atles Interactiu de l'entonació del catalá" (http://prosodia.uab.cat/atlesentonacio/cat-tobi/) has as its goal the gathering of speech samples in Catalan across dialectal varieties and various speech situations in order to provide an intonational analysis using the Cat-ToBI system. The same methodology is being applied in the project "Atlas interactivo de la entonación del español" (http://prosodia.uab.cat/atlasentonacion/index-english.html) to systematically collect audio and video materials for the study of prosody and intonation in the various dialects of Spanish, as a first step in a comprehensive study of the great dialectal diversity present in this language.

Since the study of spoken Spanish prosody can benefit from the existence of large prosodically annotated corpora, it was decided that a subset of the Glissando corpus (Project coordinator: U. Pompeu Fabra in collaboration with UAB and U. Valladolid) will be labeled using the the Spanish ToBI labeling scheme, or Sp-ToBI ([3]). The Glissando corpus is going to be a compilation of about 16 hours of speech of Spanish as it is spoken in the standard variety of Central Peninsular Spanish. It is being developed for a multi-disciplinary user group, and it is going to contain speech from various situational settings, namely, news reading, conversational speech and task-oriented speech. Once created, the corpus will be of enormous value as a mine of empirical data and a tool with to test hypotheses related to the intonational phonology of Spanish within the autosegmental-metrical framework and, in particular, to answer questions about accent categories, degrees of prominence and prosodic segmentation. Given the size of the Glissando corpus, it became clear that we would have to rely on various transcribers, and related to this, since Sp-ToBI is a fine-grained labeling which requires well-trained transcribers, it soon became transparent that high-quality prosodic annotation would be impossible to achieve without checking the intertranscriber coherence.

This paper reports on the results of a pilot study that was run to assess the reliability of the prosodic judgements using Sp-ToBI before starting any large-scale production of annotations. The goal of the evaluation was to estimate the attainable degree of consistency between transcribers, but most importantly, it was intended to identify the most confusable tones in order to make recommendations for the refinement of the Sp-ToBi system. Moreover, the statistical procedures will shed light on the most problematic issues suggesting new approaches to better predict ToBI labels for Spanish.

2. Overview of Sp-ToBI

The ToBI-framework system is a broadly accepted framework for the transcription of prosodic phenomena. It was originally developed for English, based on Pierrehumbert's autosegmental

Partially founded by the Ministerio de Ciencia e Innovación, Spanish Government Glissando project FFI2008-04982-C003-02

model, but since then applied to a large number of languages, among them Spanish [4]. It is important to make clear, however, 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 langauge, there are many different ToBI systems, each one specific to a language variety and the community of researchers working on that language variety (http://www.ling.ohio-state.edu/tobi)

Despite some attempts of automatisation [5, 6] , the prosodic annotation of speech using ToBI-framework system is a process which involves a large phase of manual work by several experts simultaneously in order to obtain an agreed transcription. This is the reason why a very important step before accepting as community-wide standard a specific language-ToBI system is to check the consistency of labels across transcribers, especially in the course of its development. At present, the Spanish ToBI labeling scheme provides a tool for the prosodic annotation of Spoken Spanish within the ToBI framework, although we must take into account the ongoing development in many varieties. The original Sp-ToBI was presented by [7], and since that time, some workshops have been organized in order to improve the system (http://prosodia.uab.cat/sp-tobi/en/references/Sp-ToBI-workshops.html) and more insight into the Spanish intonational phonology has been achieved ([3]). Due to this body of research, some modifications were proposed in [3] and a website presents a reviewed version of Sp-ToBI including those modifications [8]. The data presented on the website "Sp-ToBI training materials" (http://prosodia.uab.cat/sp-tobi/en/) are based on the analysis of Northern Peninsular Spanish, meaning that the tonal units and tonal contrasts proposed have been attested at least in this dialect.

3. Experimental procedure and results

In order to have the most objective data to evaluate the stability of the system, a test of labeling consistency was conducted to measure inter-transcriber agreement in prosodic annotation using Sp-ToBI system.

3.1. Corpus

A set of twenty sentences excerpted from a single-speaker text-to-speech corpus of read speech (http://www.talp.cat/ttsdemo/) was independently labeled by five transcribers using the Sp-ToBI system. The sentences had not previously been annotated by any of the transcribers, and each transcriber worked alone on the samples, without any prior training, since all of them are experts in intonational phonology: two of the transcribers had participated in the building of the Sp-ToBI proposal, and all of them have investigated prosodic phenomena from an intonational phonology perspective. All subjects are native speakers of Spanish, with two dialects represented (Northern Peninsular Spanish, Southern Spanish). The total amount of tonal units to be annotated in the sentences is 133 resulting 1330 comparisons of pairs of labels. Half of the sentences are declarative and half are questions. They are short sentences with at most two boundary tones.

3.2. Transcription procedure

The manual annotation was performed using the Praat tool. The transcribers were looking at a computer screen with a display of the waveform and the F0 curve, but they relied on their perception to take prosodic decisions when the phonetic implementation details were unclear. The key elements to be labeled were (1) prominence, (2) prosodic boundary strength and (3) pitch accent and boundary tone types.

For the intonational analysis of Spanish, two types of tonal events (pitch accents and boundary tones), and two levels of phrasing (the intermediate phrase-ip and the intonational phrase-IP) are recognized. According to the Sp-ToBI reviewed version, the inventory includes:

1. Six basic pitch accents : 2 monotonal (H* and L*) and 4 bitonal (L+H*, L+>H*, L*+H and H+L*).
2. Seven IP-final boundary tones (L%, M%, HH%, LH%, LM%, HL%, LHL%) and seven ip-final boundary tones (H-, L-, M-, HH-, LH-, HL-, LHL-).

Differently from other ToBI-framework systems, phrase accents are not used in Sp-ToBI, since there has been found no evidence for their need to account for the tonal movement on unaccented preboundary syllables. On the other hand, due to the size of the corpus, the annotators do not need the full list to complete their task.

3.3. Reliability measurements

In the ToBI framework system, standard procedures have been developed to evaluate with the most objective criteria the annotation carried out by more than one subject [9, 10]

3.3.1. Pairwise transcriber agreement

Agreement was measured by counting the number of labeling agreement for all pairs of transcribers. 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).

3.3.2. Kappa coefficient

The inter-transcriber consistency for prominence, break strength and pitch accent and boundary tone inventory was quantified by means of kappa coefficient. A kappa between 0.61 and 0.80 is considered to point at a substantial consistency. [11] considered a good level of agreement when the value obtained from the kappa statistic is greater than 0,7.

We compared the transcriber decisions in pairs for all types of ToBI labels (pitch accents and boundary tones) and for the decisions presence versus absence of pitch accent and presence versus absence of prosodic boundary. Given a ToBI symbol, we binarize the decisions as the identification of such symbol or a different one. The kappa coefficient is computed for all pairs of transcribers and for all ToBI labels and decisions.

3.4. Results

Table 1 and table 2 depicts the obtained reliability results computed according to the pairwise transcriber agreement and the Kappa coefficient.

Overall, the data show a low rate of agreement. With the exception of the judgement about the existence of a prosodic boundary (symbolised by * in the boundary class, and with a

	Kappa coefficient				
	max		min		
Accent	Value	Pair	Value	Pair	Mean
*	0.89	(T1,T2)	0.04	(T1,T4)	0.28
H*	0.49	(T1,T2)	-0.03	(T2,T5)	0.14
H+L*	0.66	(T2,T3)	-0.03	(T2,T5)	0.07
L*	0.89	(T1,T2)	0.19	(T3,T4)	0.39
L*+H	0.66	(T3,T5)	-0.02	(T1,T4)	0.06
L+>H*	0.68	(T1,T2)	0.32	(T1,T4)	0.52
L+H*	0.40	(T4,T5)	0.07	(T2,T4)	0.26

	Kappa coefficient				
	max		min		
Boundary	Value	Pair	Value	Pair	Mean
*	0.93	(T1,T3)	0.77	(T1,T4)	0.84
H-	0.85	(T1,T3)	0.53	(T1,T5)	0.68
H%	1.00	(T3,T5)	0.00	(T1,T3)	0.32
HH%	1.00	(T1,T2)	0.00	(T1,T3)	0.14
L-	0.56	(T2,T3)	-0.01	(T2,T5)	0.29
L%	1.00	(T1,T2)	0.96	(T1,T3)	0.97
LH%	0.00	(T1,T3)	0.00	(T1,T3)	0.00
M-	1.00	(T1,T2)	0.00	(T1,T3)	0.14

Table 1: Kappa coefficient among the different pair of transcribers: min is the minimum agreement, max is the maximum one and $mean$ the mean value of the agreements

"very good agreement" in the Kappa scale), it can be observed that the disagreement is high, as revealed by the relative low percentage in the diagonals of table 2 and the great distance between the elements in column min and max of table 1. Nevertheless, a more detailed analysis of the behaviour of individuals suggests that the disagreement could be caused by the use of different annotation criteria. For instance, the best values of agreement are for the pair T1-T2, and the worst values are for the pairs in which T1 and T2 disagree with T4 and T5.

As far as the identification of types of pitch is concerned, the most confusable labels are H*, H+L*, L+H*. Only L+>H* is close to 'good agreement' in the Kappa scale, reaching a coefficient of 0.68 for the pair of transcribers T1-T2. In the case of L*, the best values are again found for the pair T1-T2 (max value=0.89). But if we look at the confusion matrix, we found that L* has been identified as "unaccented" (symbolised by * in the pich accent class) 29% of the times, and this suggests that syllables carrying low tones can be phonetically realised in a range of frequencies so low that any accent is perceived. In line with this observation, falling tones (H+L*) are also identified as "absence of accent" 38% of cases. More surprisingly, H* is confused with * (absence of pitch accent). But if we look at the frequency with which each transcriber has used the label, we realize that T1 and T2 have used H* only twice, while T3 have used it 39 times. Another source of disagreement is the use of downstep. Table 3 quantifies the difficulty for transcribers to decide if a pitch accent is downstepped or not, with figures of agreement below 10%.

Data referred to prosodic boundaries and the type of boundary tones are clearly more consistent than the pitch accent one: according to table 1, the decision of annotating a prosodic boundary is shared by almost all pairs of labelers, with values in between 0.93 and 0.77. As expected, the boundary tone L% has 'very good agreement' in the Kappa scale (mean value between 0.96 and 1). The case of discrimination between H% and HH% is specially relevant for the issue of annotation conventions. According to the reviewed version of Sp-ToBI, at the break-index 4, only HH% is used, and this is the option adopted by T1 and T2, while the other pair of transcribers use H% in the same positions. For boundary tones at the break-index 3, a

Accent	H*	!H*	L+H*	L+!H*	L+>H*	L+>!H*
H*	**21**	08	27	09	44	00
!H*	08	**08**	12	12	05	02
L+H*	27	12	**40**	22	55	06
L+!H*	09	12	22	**07**	15	09
L+>H*	44	05	55	15	**126**	11
L+>!H*	00	02	06	09	11	**05**

Table 3: Pairwise transcriber agreement for rising pitch accents and their downstepped counterparts

'good agreement' is found for H-. according to the figures in table 2, while L- are easily confounded with M- (there are no cases of M% identified in the corpus).

4. Discussion and conclusions

Previous works of intertranscriber reliability of ToBI-framework systems have certified between 81% and 92% of agreement in determining pitch accents for English [11], overall mean scores of 88.9% of agreement for German [9], and agreement percentages of between 59% and 91% (depending on accent categories) for Korean ([12]). The results obtained here are far from those. Nevertheless, it should be pointed out that measuring the levels of agreement in terms of success of the system was not our goal. Neither the delimitation of the corpora (from differents speech sources in other experiments) nor the selection of the transcribers (usually experts versus beginners) meet the conditions established within ToBI framework to evaluate the reliability, learnability and comprehensiveness of a given system. The main objective of this study was to compare the prosodic judgements of expert annotators in order to identify the most confusable labels.

A great majority of the divergences can be explained by the similarity of some types of pitch accents and the contexts in which they are phonetically implemented. Related to this, results replicate the findings in previous studies that have used the MAE-ToBI categories. For instance, [13] reports a procedure to derive a conceptual similarity index indicating the distance between tone categories. It is found that tones which are conceptually similar result in higher disagreement among labelers, while tones which are conceptually dissimilar result in lower disagreement. This is the case for the pair of pitch accents in which the only difference between the two tones is whether there is a leading tone present or not: H* versus L+H*. This is because many H* accents have an apparent L target at the start of their rise and because the distinction can be held to involve peak height (with the H* in L+H* being lower). The phonetic dimensions of these intonational categories (H* and L+H*) have been investigated in [14] and the results show that the pitch variation and segmental alignment are the result of the interaction of three distinc levels of intonational effects (global extrinsic, local extrinsic and intrinsic effects).

There are two pairs also noted as confusable in our test in which the only difference is the use of downstep: L+H* versus L+!H* and H* versus !H*. This disagrement is in line with the results reported by [15] and with the categorization as similar in [13] due to the confusability of whether there is pitch range reduction present or not.

With respect to the confusability L- vs M-, or alternatively L% vs M%, the mid tone is a controversial tone considered to be necessary for the description of the Spanish intonation since the first proposal of Sp-ToBI in [7]. The web Sp-ToBI Training Materials and [3] offers the following definition: "M- and M%

Accent	*	L*	L*+H	H+L*	H*	L+H*	L+>H*
*	**132 (29%)**	123 (36%)	05 (14%)	18 (38%)	97 (32%)	56 (18%)	21 (07%)
L*	123 (27%)	**117 (34%)**	15 (43%)	09 (19%)	45 (15%)	27 (09%)	07 (02%)
L*+H	05 (01%)	15 (04%)	**01 (03%)**	01 (02%)	01 (00%)	02 (01%)	10 (03%)
H+L*	18 (04%)	09 (03%)	01 (03%)	**01 (02%)**	08 (03%)	08 (03%)	02 (01%)
H*	97 (21%)	45 (13%)	01 (03%)	08 (17%)	**37 (12%)**	60 (20%)	51 (16%)
L+H*	56 (12%)	27 (08%)	02 (06%)	08 (17%)	60 (20%)	**69 (22%)**	85 (27%)
L+>H*	21 (05%)	07 (02%)	10 (29%)	02 (04%)	51 (17%)	85 (28%)	**142 (45%)**
Total	452	343	35	47	299	307	318

Boundary	*	L-	M-	H-	L%	LH%	HH%	H%
*	**781 (91%)**	24 (56%)	00 (00%)	50 (38%)	00 (00%)	00 (00%)	00 (00%)	00 (00%)
L-	24 (03%)	**09 (21%)**	02 (29%)	08 (06%)	00 (00%)	00 (00%)	00 (00%)	00 (00%)
M-	00 (00%)	02 (05%)	**01 (14%)**	04 (03%)	00 (00%)	00 (00%)	00 (00%)	00 (00%)
H-	50 (06%)	08 (19%)	04 (57%)	**71 (53%)**	00 (00%)	00 (00%)	00 (00%)	00 (00%)
L%	00 (00%)	00 (00%)	00 (00%)	00 (00%)	**123 (95%)**	03 (75%)	00 (00%)	03 (04%)
LH%	00 (00%)	00 (00%)	00 (00%)	00 (00%)	03 (02%)	**00 (00%)**	00 (00%)	01 (01%)
HH%	00 (00%)	00 (00%)	00 (00%)	00 (00%)	00 (00%)	00 (00%)	**07 (14%)**	42 (63%)
H%	00 (00%)	00 (00%)	00 (00%)	00 (00%)	03 (02%)	01 (25%)	42 (86%)	**21 (31%)**
Total	855	43	7	133	129	4	49	67

Table 2: Confusion matrices (raw scores and percentages)

are manifested phonetically as a falling movement to a mid tone target or as a mid level plateau when it occurs after a high tone". A discussion about the phonological status of the M% level can be found in [16], and according to this, the final pitch height is independent of any syntagmatic reference to preceding pitch accents. Nevertheless, the results of the test, with a high rate of confusions L- vs M-, show that the decisions about pitch height are not so straightforward.

To conclude, although the Spanish ToBI labeling system has been proved as an effective system to annotate intonation for Spanish (since the labels cover the prosodic phenomena encountered in the corpus), the inter-transcriber reliability metrics has shown that a broader consensus among transcribers is needed in order to have a valid transcription system. To achieve this, further studies of the perceptual and acoustic properties of tones are needed. This is specially true for the following pitch accents: high pitch accent (H*) vs rising pitch accent (L+H*), downstepped pitch accents versus non-downstepped counterparts. With respect to boundary tones, the worst distinguished categories are the mid tones. A related issue is the difficulty to decide in a very low pitch range if a tone is present or if the syllable has been unaccented.

The approach suggested here is to further deepen the study of the phonetic dimensions of intonational categories to better understand the acoustic and perceptual cues that are associated to them, and, as a consequence, to achieve a greater coherence in their operational definitions in the labeling system. Our data are useful in pointing out which distinctions are most amenable to future research. For the purposes of the construction of the Glissando corpus, a large speech database from different communicative sources, to reduce the inventory of labels according to their confusability so as to develop a tool that at least speeds manual labeling of prosody seems a valid alternative to full manual annotation.

5. Acknowledgements

We want to acknowledge the labeling task undergone by L. Astruc, M. Cabrera, E. Estebas, C. de-la-Mota and F. Vizcaíno within the Glissando project. Antonio Bonafonte gently offered the speech database from where the corpus was excerpted.

6. References

[1] D. R. Ladd, *Intonational Phonology*. Cambridge University Press, 1996.

[2] E. Campione and J. Veronis, "A multilingual prosodic database," in *Proceedings of ICSLP 98*, 1998.

[3] E. Estebas and P. Prieto, "La notación prosódica del español. una revisión del Sp-ToBI," *Estudios de Fonética Experimental*, vol. XVIII, pp. 263–283, 2009.

[4] J. I. Hualde, "El modelo métrico autosegmental," in *Teorías de la entonación*, P. Prieto, Ed. Ariel, 2003, pp. 155–184.

[5] M. Q. Wang and J. Hirschberg, "Automatic classification of intonational phrasing boundaries," *Computer Speech and Language*, no. 6, pp. 175–196, 1992.

[6] A. K. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling prosody," *Speech Communications*, no. 33, pp. 135–151, 2001.

[7] M. E. Beckman, M. D. Campos, J. T. McGregory, and T. A. Morgan, "Intonation across Spanish, in the tones and break indices framework," University of Ohio, Tech. Rep. http://www.ling.ohio-state.edu/ tobi/sp-tobi/, 2000.

[8] C. de-la Mota, L. Aguilar, J. Borras-Comes, and R. Sichel-Bazin, "Development of Catalan and Spanish ToBI online training materials," in *Proceedings of International Conference Phonetics and Phonology in Iberia (PaPI)*, 2009.

[9] M. Grice, R. Reyelt, R. Benzsuller, and A. Batliner, "Consistency in transcription and labelling of German intonation with GToBI," in *Proceedings of ICSLP 1996*, 1996, pp. 1716–1719.

[10] A. Syrdal and J. McGory, "Inter-transcriber reliability of ToBI prosodic labeling," in *Proceedings of ICSLP*, 2000, pp. 235–238.

[11] T. Yoon, S. Chavarria, J. Cole, and M. Hasegawa, "Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," in *Proceedings of Interspeech 2004*, 2004.

[12] S. Jun, S. Lee, K. Kim, and Y. Lee, "Labeler agreement in transcribing Korean intonation with K-ToBI," in *Proceedings of ICSLP 2000*, 2000.

[13] R. Herman and J. McGory, "The conceptual similarity of intonational tones and its effects on intertranscriber reliability," *Language and Speech*, vol. 45, pp. 1–36, 2002.

[14] S. Calhoun, "Phonetic dimensions of intonational categories: the case of L+H* and H*," in *Proceedings of Prosody 2004*, 2004.

[15] J. Pitrelli, M. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *Proceedings of ICSLP 94*, 1994, pp. 123–126.

[16] P. Prieto, "The intonational phonology of Catalan," in *Prosodic Typology*, S.-A. Jun, Ed. Oxford University Press, 2009.