TexAFon 2.0:
A text processing tool for the generation of expressive speech in TTS applications

Juan María Garrido, Yesika Laplaza, Benjamin Kolz, Miquel Cornudella

Department of Translation and Language Sciences, Pompeu Fabra University, Barcelona, Spain
Roc Boronat 138, 08108 Barcelona, Spain
E-mail: juanmaria.garrido@upf.edu, yesika.laplaza@gmail.com, benjamin.kolz@upf.edu, cornudella@csl.sony.fr

Abstract

This paper presents TexAfon 2.0, an improved version of the text processing tool TexAFon, specially oriented to the generation of synthetic speech with expressive content. TexAFon is a text processing module in Catalan and Spanish for TTS systems, which performs all the typical tasks needed for the generation of synthetic speech from text: sentence detection, pre-processing, phonetic transcription, syllabication, prosodic segmentation and stress prediction. These improvements include a new normalisation module for the standardisation on chat text in Spanish, a module for the detection of the expressed emotions in the input text, and a module for the automatic detection of the intended speech acts, which are briefly described in the paper. The results of the evaluations carried out for each module are also presented.

Keywords: emotion detection, speech act detection, text normalisation

1. Introduction

Language and speech technological applications in general, and TTS systems in particular, have increasingly to deal, among many other aspects, with the processing and generation of expressive language (messages coming from speech-based person-machine interfaces, e-mail, SMS, chat or Twitter, for example). Current text processing modules of TTS systems have to be improved with new capabilities to process efficiently this type of text:

- **Text normalisation**. Many of these texts with expressive content are not written 'correctly' (that is, following the standard orthographic conventions of the input language): 'see U', in English, or 'a10' in Catalan, are becoming usual expressions in many contexts. Classical text processing modules are not able to handle them correctly, because they expect an orthographically correct input text. New correction and normalisation procedures have to be implemented in those systems to convert this 'incorrect' input text to a standard form.
- **Text analysis**. Expressiveness in speech is mainly transmitted through prosody, which is related to several linguistic and paralinguistic factors, such as the presence of focused words, the speech act of the utterance or the emotion being expressed. Current TTS systems do not produce good expressive speech, among other factors, because they cannot extract from texts information relative to the emotion, the speech act or the words that should be pronounced as bearing focus. New procedures should then be included in text processing modules in order to extract from the text as many information as possible relevant for the generation of expressive prosody.

This paper describes several improvements introduced in TexAFon (Garrido et al., 2012) to process correctly text with expressive content. TexAFon is a text processing module in Catalan and Spanish for TTS systems which performs all the typical tasks needed for the generation of synthetic speech from text. Its output can be directly used to generate speech using several synthesis engines, such the one by Cereproc (Garrido et al., 2008) or MBROLA (Dutoit et al., 1996), but it can also be used for other purposes (automatic phonetic transcription, building of phonetic dictionaries). The improvements described here have been focused on the normalisation of non-standard text, and the detection of emotions and speech acts in the input text. Normalisation and emotion detection has only been developed for the Spanish module, but speech act detection is already available both for Catalan and Spanish.

2. TexAfon 2.0 overview

TexAFon is a set of Catalan/Spanish text processing tools for automatic normalization, phonetic transcription, syllabication, prosodic segmentation and stress prediction from text. It has been jointly developed by researchers of the Computational Linguistics Group (GLiCom) of Pompeu Fabra University and the Speech and Language Group at Barcelona Media. Fully developed in Python, TexAFon uses linguistic knowledge-based approaches to perform the text processing tasks. This linguistic knowledge has been implemented in the form of:

- Python procedures, containing the linguistic rules;
- Python lists, containing non-editable information;
- External dictionaries, stored in text files (then editable by external users).

TexAFon has a modular architecture, which facilitates

the development of new applications using it, the addition of new languages, and the connection to other external modules and applications. This architecture clearly differentiates among:

- A general processing core, which includes the language-independent procedures.
- The language packages (two, for Spanish and Catalan), including modules and dictionaries specific of the language.
- The applications, which call the processing core depending on their needs.

Figure 1 illustrates the general architecture of TexAFon and the new improvements included in the 2.0 version. The implementation of the emotion and dialog act detection features has been carried out by connecting two new external modules (EmotionFinder and DetectAct) to the general processing core of TexAFon. An external POS tagger and a lemmatiser, necessary for some auxiliary tasks of these two new modules, have also been integrated. Also, the text normalisation module of the processing core has been modified to allow normalization of non-standard text, and new elements have been added to the language-dependent processing package for Spanish, such as a new normalisation module for non-standard text, several auxiliary dictionaries and a new language model.

Next sections describe briefly the development of these new modules, their workflow and evaluation procedures.

3. Normalisation of non-standard text

Traditional pre-processing modules for TTS expand non-standard expressions, such as dates, hours, or URL addresses, to its corresponding orthographical form, but they assume that standard tokens (words) are correctly written. However, chat and other Internet texts contain words not written in a standard form that are not well processed by those modules. The normalisation process implemented in TexAFon 2.0 has been designed to identify and normalise (in the sense of converting to an orthographically standard representation) those forms. It has been integrated into the general normalisation module already included in TexAFon, to take advantage of the information provided by this module about the nature of the input tokens and avoid overcorrection.

The linguistic knowledge used for its development has been extracted from the computational analysis of a corpus of 8,780 real chat messages in Spanish, containing 40,676 tokens, which were manually standardised and analysed by a single annotator. This analysis allowed to define the main types of orthographic 'deviations' (such as 'character subtitution', 'character deletion', 'stress mark deletion', 'character addition') that this kind of texts show in Spanish, and its frequency, information that was later used to develop the correction rules implemented in the system. Table 1 presents some examples of the types of deviations defined. The corpus was also used to create the abbreviation dictionary and language model needed for the selection of the final form.

Input token	Standardised token	Label
Algregada	Agregada	Character addition
Q	Que	Character deletion
Ahoar	Ahora	Character transpositon
Osea	O sea	Character deletion

Table 1: Some examples of classification of non-standard tokens in the analysed corpus of chat messages.

3.1 Normalisation procedure

The implemented normalisation process involves three steps:

- Input tokens are classified into categories ('date', 'URL', 'isolated letter', 'word', 'letters&numbers', 'letters&symbols', 'smiley', etc.). This is the general process already established for standard text, but some new categories, such as 'smiley' have been added to process this non-standard texts.
- During the expansion process, tokens belonging to any of the 'correctable' categories ('letters&numbers', 'letters& symbols', 'words') are detected and submitted to normalisation. This avoids overcorrection of some types of tokens, such as email addresses or URL. A specific normalisation procedure is applied to each detected 'correctable' token, depending on its label. In the case of the 'word' category, which is the most frequent one, the input token is checked in a dictionary of standard forms; if it not there, it is identified as 'incorrect'. The normalisation module generates then a list of possible correct forms, by applying an ordered set of rules, dealing with the most typical deviation phenomena detected in the analysis of the corpus: deletion of repeated characters, character substitution, character deletion, character insertion and character transposition. At the end of this process, a list of possible correct forms for the incorrect input word is obtained.
- The selection of the corrected form to substitute the incorrect input is made from the list of candidate words, using the language model.

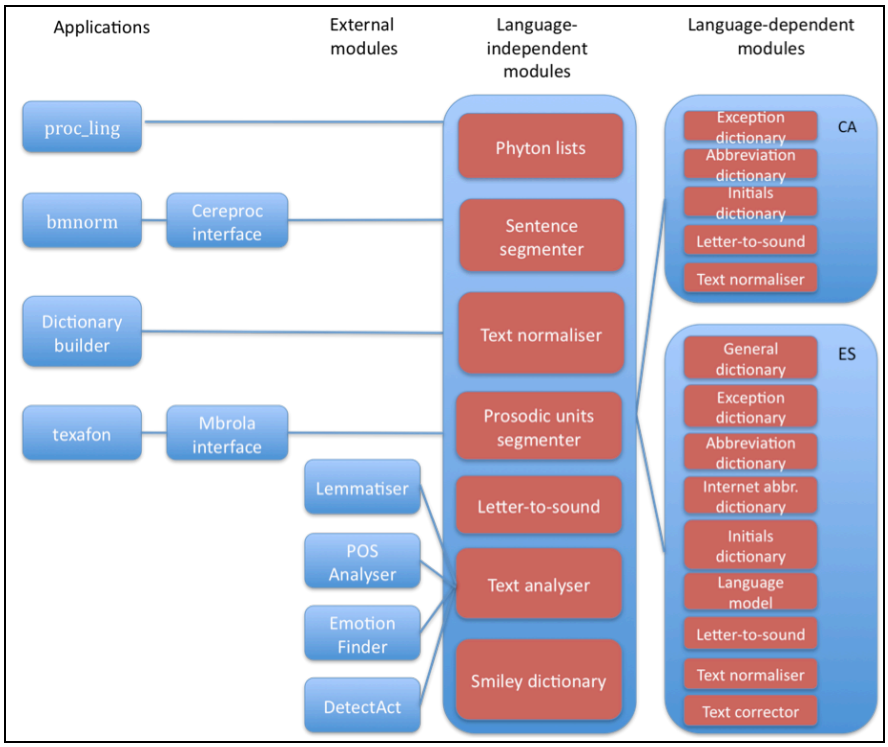

Figure 1: General TexAFon 2.0 architecture

Table 2 presents two examples of this normalisation process.

Input text	Eso k t llevas	Cuando tngas 40
Token classification	Eso [word] K [abbreviation] T [abbreviation] Llevas [word]	Cuando [word] Tngas [word] 40 [number]
Token expansion	Eso: [eso] K: [que, qué] T: [te, tu, tú] Llevas: [llevas]	Cuando [cuando] Tngas [angas, ingas, ingás, tajas, tajás, tangas, tanjas, tejas, tengas, tengo as, tijas, tingas, togas, tojas, tongas, tugas, tujas, tungas, unjas, vengas] 40 [cuarenta]
Final correct form selection	Eso: [eso] K: [**que**, qué] T: [**te**, tu, tú] Llevas: [llevas]	Cuando [cuando] Tngas [angas, ingas, ingás, tajas, tajás, tangas, tanjas, tejas, **tengas**, tengo as, tijas, tingas, togas, tojas, tongas, tugas, tujas, tungas, unjas, vengas] 40 [cuarenta]
Output Text	Eso que te llevas	Cuando tengas cuarenta

Table 2: Normalisation workflow for two sample chat utterances of the analysis corpus.

3.2 Evaluation

Table 3 summarizes the results of the evaluation carried out to analyse the performance of the normaliser. It was done on a corpus of 1,222 chat messages (4,069 tokens), different to the one used for the analysis task. The obtained results show that, by applying this normalisation procedure, 70.3% of the tokens considered as 'incorrect' by TexAFon 2.0 where correctly normalised, with a final percentage of 85.9% of correct tokens in the output normalised text (24.7% of improvement). These results are similar to the ones obtained with other similar systems described in the literature for Spanish (Armenta et al., 2003, for example).

	Standard		Non-standard	
Text	Tokens	%	Tokens	%
Original	2,490	61.2	1,579	38.8
Normalised	3,495	85.9	574	14.1

Table 3: Results of the evaluation of the new normalization module.

4. Emotion detection

EmotionFinder is the module in charge of the emotion detection task. It works at sentence level: it tries to assign a single emotional label (or none, if the text is considered to be 'neutral') to each sentence detected by TexAFon in the input text. It assumes a previous step of lemmatisation of the words making up the input sentence (EmotionFidner works only with lemmatized words, to improve its generalization power), which is carried out

by a separate module (Lemmatiser) which has also been integrated in TexAFon.

EmotionFinder is able to detect eight different emotions in the input text: 'admiration', 'affection', 'disappointment', 'interest', 'happiness', 'surprise', 'rejection' and 'sadness'. It assigns also an emotion intensity label (1, 2 or 3) to the detected emotion.

Entry	Intensity	Emotion	Weight
estupendo	2	admiration	50
excepcional	2	admiration	50
extraordinario	2	admiration	60
fascinar	3	admiration	70
fascinación	3	admiration	70
fenómeno	2	admiration	70
formidable	2	admiration	60
forrarse	2	admiration	60
fuerte	2	admiration	60
genial	3	admiration	60

Table 4: Sample entries of the emotional dictionary.

The EmotionFinder module is made up of two components:

- An emotional dictionary, which includes words and expressions associated to a given emotion and intensity, and a weight indicating its reliability to detect that emotion. Table 4 shows some examples of entries of this dictionary.
- A set of functions, one per emotion, which combine searching for key words (taken from the emotional dictionary) and regular expressions with rule-based emotion inference.

Both the emotional dictionary and the rules included in the functions have been developed using the results of the analysis of a set of 4,207 utterances of real chat messages in Spanish, which is a subset of the corpus of chat conversations used for the development of the normalisation module. This corpus was labeled with emotional tags by the same annotator who carried out the analysis of orthographic deviations, using the inventory of emotions described in Garrido et al. (2012a), and then partially revised by two people different from the main annotator. This corpus was also used to determine the set of emotions to be detected by EmotionFinder, which is a subset of the list emotions more frequently expressed in the chat conversations of the corpus.

More details about EmotionFinder and its development process can be found in Kolz et al. (2014).

4.1 Emotion detection procedure

The emotion labelling process in EmotionFinder works as follows:

- The input sentence is provided as input to every emotion detection functions to check for possible cues related to the considered emotions. If a function detects one or several cues for its corresponding emotion in the input sentence, it generates as output the following information: label of the candidate emotion; the predicted intensity of the emotion (1, 2 or 3); and an associated weight indicating how reliable is the cue for the detection of that emotion, which is the sum of the weights all of hits found for the selected emotion in the sentence. So for example, the output of the function corresponding to 'happiness' for the Spanish sentence "*Estoy feliz y encantado con el plan*" would be 'ALEGRIA(3):70' (happiness with intensity level 3, and weight 70), which would be the result of the combination of the information of two different cues detected in the sentence: the presence of the word '*feliz*', labeled in the emotional dictionary as 'ALEGRIA(2):40' and '*encantado*', labeled in the dictionary as 'ALEGRIA(3):30'. This output information is added to the list of 'candidate emotions' of the sentence.
- After applying all the functions to the input sentence, the list of candidate emotions is ordered according to the weights obtained by each emotion/intensity pair. The emotion label and intensity with the highest weight is selected as the emotion label for the sentence. If the candidate list is empty, the sentence is labeled as 'neutral'.

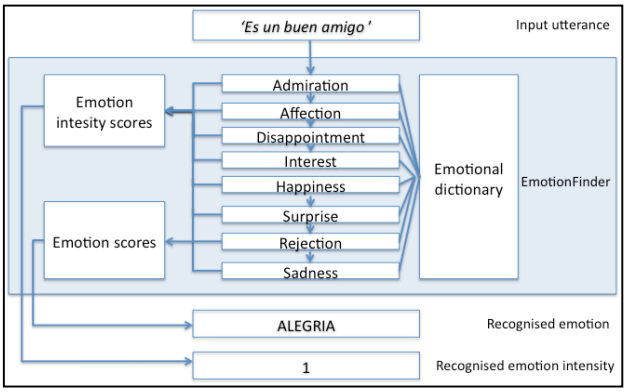

Figure 2: EmotionFinder workflow.

4.2 Evaluation

EmotionFinder was evaluated using two different corpora: the same corpus used for the development of the system, and a second corpus of 609 labelled chat messages, different from the ones of the development corpus, but belonging also to the same general corpus of chat messages previously described. Tables 5 and 6 show the results of these two evaluations. In the first one, a mean precision of 0.54 was obtained, with a recall of 0.49, but strong differences among emotional labels were observed. Best results are obtained in the case of the 'interest' label (0.67), followed by the 'neutral' label (0.65). Labels showing the worst results are 'disappointment' (0.05) and 'surprise' (0.04). In the

second one, the obtained mean precision for all the considered emotional labels plus 'neutral' (absence of emotion) was 0.6, and recall 0.58, slightly better than in the previous evaluation, but individual emotional labels show clearly lower values than in the previous evaluation, with two labels ('happiness' and 'surprise') having a precision score of 0, and a maximum of 0.33 in the case of 'rejection'. These results reveal a dependency on the training corpus of the dictionary and the rules. Anyway, these results are slightly better than those of the system chosen as reference, oriented also to the detection of emotions in Spanish for TTS purposes (García and Alías, 2008), for a more complex identification task (nine emotional labels in EmotionFinder versus the six labels of the reference system).

Label	True positive	False positive	False negative	Recall	Precision	F1
Neutral	1216	762	540	0.69	0.61	0.65
Happiness	71	141	424	0.14	0.34	0.2
Admiration	49	128	41	0.54	0.28	0.37
Affection	96	205	124	0.44	0.32	0.37
Rejection	214	175	517	0.29	0.55	0.38
Surprise	4	54	118	0.03	0.07	0.04
Interest	276	143	131	0.68	0.66	0.67
Sadness	22	30	89	0.2	0.42	0.27
Disappointment	2	20	57	0.03	0.09	0.05
TOTAL	**1950**	**1658**	**2041**	**0.49**	**0.54**	**0.51**

Table 5: Results obtained in the evaluation with the development corpus.

Label	True positive	False positive	False negative	Recall	Precision	F1
Neutral	308	90	72	0.81	0.77	0.79
Happiness	0	9	11	0	0	0
Admiration	2	28	8	0.07	0.2	0.1
Affection	20	43	47	0.30	0.32	0.31
Rejection	6	12	48	0.11	0.33	0.17
Surprise	0	2	34	0	0	0
Interest	14	38	11	0.56	0.27	0.36
Sadness	1	5	21	0.05	0.17	0.08
Disappointment	1	6	5	0.17	0.14	0.15
TOTAL	**352**	**233**	**257**	**0.58**	**0.60**	**0.59**

Table 6: Results obtained with the evaluation corpus

5. Speech act detection

DetectAct is the external module responsible of assigning a speech act label to every input sentence. It works in a similar way to EmotionFinder, in the sense that it uses lexical information, contained in a speech act dictionary, to determine speech acts, but it shows also some noticeable differences, as for example that the list of speech acts labels is not closed and predefined, as in the case of EmotionFinder, but open (labels are read from the speech act dictionary).

DetectAct is then made up of two components:

- A speech act dictionary per language, which contains the words relevant for the detection of each considered act, with a weight associated to each one. Table 7 shows some examples of the information contained in these dictionaries.
- A decision function, which uses the information contained in the dictionary to determine the speech act label that best fits the input sentence.

Speech act dictionaries are obtained from the automatic analysis of a corpus annotated with speech act labels, and not manually, as in the case of EmotionFinder. This automatic training generates a list of words for each one of the speech act labels detected in the corpus, and calculates a weight for each word in the list, which is actually the sum of occurrences of each word in the training corpus associated to that speech act label. The same word can appear then several times in the dictionary, but associated to a different speech act label and with a different weight.

The corpora used to develop the dictionaries currently used by DetectAct are not representative of chat text: they are made up of a set of automatic telephone service messages, both in Spanish and Catalan (716 messages for each language), extracted from the I3Media database (Garrido et al., 2012a). This corpus was annotated by two different annotators using a set of nine different dialogue act labels ('greeting', 'acknowledgment', 'apology', 'action request', 'information request', 'confirmation request', 'warning', 'confirmation', 'information') inspired in those defined for the SPAAC project (Leech and Weisser, 2003). However, other dictionaries, representing other types of text, could be easily built using this procedure it an annotated corpus was available.

Word	Label	Weight
benvingut	Greeting	31
	Information	1
	Apology	1
gràcies	Acknowledgement	6
	Information	1
premi	Information	129
	Action_request	32
	Confirmation_request	10
teclegi	Action_request	19
	Information	5

Table 7: Examples of weights associated to some words in the speech act dictionary as a function of the speech act label

5.1 Speech act detection procedure

The processing workflow of DetectAct, illustrated in

figure 3, is quite straightforward:

- Words of each input sentence are first labelled with POS tags using the POS tagger
- Every word not classified as content word in the previous step is checked in the speech act dictionary, in order to retrieve all the speech act labels and weights associated to it.
- At the end of the previous process, a grade for each candidate speech act is calculated, which is the sum of the weights of the key words associated to it found in the dictionary. The selected label is the one with the highest grade at the end of this process.

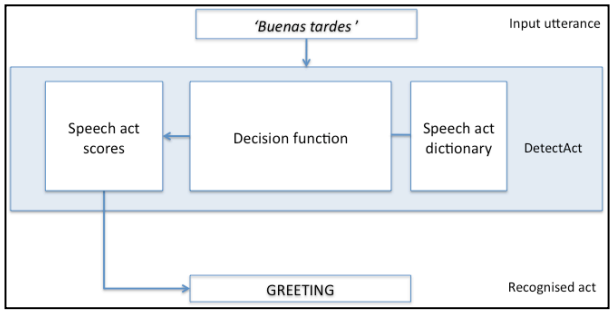

Figure 3: DetectAct workflow.

5.2 Evaluation

DetectAct was evaluated using two corpora, one for Spanish and one for Catalan, different to the used for the training of the dictionaries, but also recorded during the I3Media project. These corpora contained 1,220 (Spanish) and 1,288 (Catalan) telephone service messages similar to the ones of the training corpora, and had been annotated with similar, but not fully coincident, speech act labels (only 911 and 976 messages, respectively, of these corpora are labeled with one of the speech act labels appearing at the training corpora). The results of this evaluation, presented in table 8, gave a 55.90% of correct assigned labels for Spanish and 57.84% for Catalan. These results are also close to the ones of other similar systems (Webb et al., 2005, for example). It is worth to notice that these results are noticeably improved if only those messages labeled with speech act labels appearing in the training corpus are considered: 74.86% of correct assigned labels for Spanish and 76.33% for Catalan.

Language	Precision (%)	Recall	F1
Spanish	55.90	0.62	0.59
Catalan	57.84	0.62	0.60

Table 8: Results of the evaluation of DetectAct.

A deeper analysis of the evaluation data reveals strong differences among speech act labels (tables 9 and 10). The best results are obtained in the case of the 'neutral' label in both languages (0.79 and 0.80 for Spanish and Catalan, respectively) and in the 'greeting' label for Spanish (0.69). Low values are observed in both languages in individual speech act labels. Even more, there are labels for Spanish ('information request', 'confirmation request' and 'warning') and Catalan ('greeting' and 'information request') that have a precision score of zero. These results reveal a dependency on the training corpus, as in the case of the Emotion detection module.

Label	True positive	False positive	False negative	Recall	Precision	F1
Neutral	609	99	222	0.73	0.86	0.79
Greeting	33	7	23	0.59	0.83	0.69
Acknowledgment	8	2	55	0.13	0.80	0.22
Apology	13	0	33	0.28	1.00	0.44
Action Request	3	26	0	1.00	0.10	0.18
Information Request	0	0	0	0	0	0
Confirmation Request	0	0	36	0	0	0
Warning	0	0	0	0	0	0
Confirmation	6	0	51	0.11	1.00	0.20
Information	10	404	0	1.00	0.02	0.04
TOTAL	**682**	**538**	**420**	**0.62**	**0.56**	**0.59**

Table 8: Results obtained for Spanish in the evaluation of the speech act detection module.

Label	True positive	False positive	False negative	Recall	Precision	F1
Neutral	695	76	264	0.72	0.89	0.80
Greeting	0	13	40	0	0	0
Acknowledgment	8	2	51	0.16	0.80	0.27
Apology	14	0	28	0.33	1.00	0.50
Action Request	2	8	0	1.00	0.20	0.33
Information Request	0	10	0	0	0	0
Confirmation Request	14	4	27	0.34	0.78	0.47
Warning	1	4	0	1.00	0.20	0.33
Confirmation	4	0	44	0.08	1.00	0.15
Information	7	426	0	1.00	0.02	0.04
TOTAL	**745**	**543**	**454**	**0.62**	**0.58**	**0.60**

Table 9: Results obtained for Catalan in the evaluation of the speech act detection module.

6. Conclusion

Improvements described here make TexAFon 2.0 a useful tool for the processing of expressive text in Spanish and Catalan, and a good example of the capabilities of knowledge-based approaches in the generation of expressive synthetic speech. The evaluation results outlined here, although different for each module, are encouraging, with scores at the level of similar state-of-the-art systems, in some cases for more complex tasks, as for EmotionFinder. More research is still in progress to explore the possibilities of improving the performance of these modules by using exclusively knowledge-based approaches.

7. References

Armenta, A., Escalada, J. G., Garrido, J. M. & Rodríguez, M. A. (2003). Desarrollo de un corrector ortográfico para aplicaciones de conversión texto-voz. *Procesamiento del Lenguaje Natural*, 31, pp. 65-72.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vrecken, O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. *Proceedings ICSLP 96, Philadelphia*, 3, pp. 1393-1396.

García, D. & Alías, F. V. (2008). Identificación de emociones a partir de texto usando desambiguación semántica. *Procesamiento del Lenguaje Natural*, 40, pp. 75-82.

Garrido, J. M., Bofias, E., Laplaza, Y., Marquina, M., Aylett, M., & Pidcock, Ch. (2008). The CEREVOICE speech synthesiser. *Actas de las V Jornadas de Tecnología del Habla (Bilbao, 12-14 noviembre 2008)*, pp. 126-129.

Garrido, J. M., Laplaza, Y., Marquina, M, Pearman, A., Escalada, J. G., Rodríguez, M. A. and Armenta, A. (2012a). The I3MEDIA speech database: a trilingual annotated corpus for the analysis and synthesis of emotional speech, *LREC 2012 Proceedings*: pp. 1197-1202. Online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/865_Paper.pdf, accessed on 21 March 2014.

Garrido, J. M., Laplaza, Y., Marquina, M., Schoenfelder, C., and Rustullet, S. (2012b). TexAFon: a multilingual text processing tool for text-to-speech applications. *Proceedings of IberSpeech 2012, Madrid, Spain, November 21-23, 2012*, pp. 281-289.

Kolz, B., Garrido, J. M., Laplaza, Y. (2014). Automatic prediction of emotions from text in Spanish for expressive speech synthesis in the chat domain, *Procesamiento del Lenguaje Natural*, 52. Online: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4904/2918, accessed on 21 March 2014.

Leech, G. and Weisser, M. (2003). Generic speech act annotation for task oriented Dialogues. *Proceedings of the Corpus Linguistics 2003 conference*, University Centre for Computer Corpus Research on Language, Technical Papers 16.1, pp. 441-446.

Python Language Website, http://www.python.org/.

Webb, N., Hepple, M., & Wilks, Y. (2005). Dialogue act classification based on intra-utterance features. *Proceedings of the AAAI Workshop on Spoken Language Understanding*.