Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 93-101 recibido 1-05-2009; aceptado 5-06-2009

On the definition of a prosodically balanced corpus: combining greedy algorithms with expert guided manipulation [*]

Hacia la definición de un corpus equilibrado prosódicamente: estrategia combinada de algoritmos voraces y manipulación de expertos

David Escudero-Mancebo
Universidad de Valladolid
descuder@infor.uva.es

Lourdes Aguilar
Universitat Autónoma de Barcelona
Lourdes.Aguilar@uab.cat

Antonio Bonafonte
Universitat Politècnica de Catalunya
antonio.bonafonte@upc.edu

Juan María Garrido Almiñana
Universitat Pompeu Fabra
juanmaria.garrido@upf.edu

Resumen: Este artículo presenta el proceso de definición de un corpus de texto equilibrado en términos de atributos prosódicos. Se presenta formalmente la aplicación de algoritmos voraces y se discuten sus limitaciones. Además, se propone una guía de manipulación de textos que contribuye a mejorar considerablemente los resultados. El trabajo experimental constata este hecho con la aplicación de la metodología en diversos corpus de noticias radiofónicas en español.

Palabras clave: Selección de subcorpus, algoritmos voraces, modelado prosódico

Abstract: This article reports the process of building a balanced text corpus taking into account prosodic features. We formalize the application of greedy algorithms for text selection and we discuss their limitations. We also defend an expert guideline for text manipulation that significantly improves the performance of the algorithms. The application of this methodology to a radio news corpus empirically supports the proposed strategy.

Keywords: Subcorpus selection, greedy algorithms, prosodic modelling

1 *Introduction*

Subcorpus selection is a need in various domains of speech technologies. In text-to-speech, greedy algorithms are used to build the space limited unit-selection data base (van Santen and Buchsbaum, 1997); in speech recognition the training corpus must be selected to find a representative sample (Nagorski, Boves, and Steeneken, 1992). Although some authors have proposed to randomly select the subcorpus (see (Lambert, Braunshweiler, and Buchholz, 2007)), text selection is broadly extended to ensure the representativeness and/or to maximize the units coverage of the corpus.

This contribution reports the use and the comparison of selection techniques for building a prosodically balanced corpus that intends to be a reference in the prosodic studies. This activity has been done in the framework of the research project Glissando, which main goal is to build a reference prosodic corpus for Spanish and Catalan. It is being developed for a multi-disciplinary user group, and it is going to contain speech from three situational settings, namely, news reading, conversational speech and task-oriented speech. All speech will be orthographically and phonetically transcribed, and a manually verified prosodic annotation will be provided. Given the large-scale compilation, it became clear the need of a subcorpus selection, because the reading of radio news should be limited to a time of thirty minutes.

The major milestone was to select a corpus that contains a balanced sample of prosodic phenomena in Spanish. A priori, the problem is not very different to the issue

[*] Partially founded by the Ministerio de Ciencia e Innovación, Spanish Government Glissando project FFI2008-04982-C003-02 and AVIVAVOZ project TEC2006-13694-C03

ISSN: 1135-5948

of constructing a phonetically balanced subcorpus if we have a reference prosodic unit and the set of prosodic features to characterizes it. In this paper we have chosen stress groups as this basic reference prosodic unit and texts have been labelled using it, so that greedy algorithms can devise a prosodically balanced subcorpus. For the definition of the radio news corpus, we had access to a small radio database that belongs to the *Cadena Ser Corporation* and to a huge text radio corpus from *United Nations Radio*, from now on, the mother corpus.

Variability is probably the main characteristics of prosody, with a high number of factors that affects its form and function. Under this condition, it is not easy to avoid a problem that dramatically decreases the succes rates of the greedy algorithms: the scarcity of some type of units in the mother corpus (unfrequent phonemes in the case of phonetically balanced selection). During the process of selection, the greedy algorithms can discard those rare types of units because of their low relative importance with respect to the more frequent types of units. However, as far as prosodic modelling is concerned, the appearance of unfrequent situations in the corpus is a must, as they can show relevant intonation shapes and functions. In the literature we find greedy algorithms with modified heuristics (van Santen and Buchsbaum, 1997) and specialized search strategies (Zhang and Nakamura, 2001) to select this class of rare units. In this paper, a new strategy based on the use of dynamic goal functions is proposed, and the application of different greedy heuristics and strategies for the definition of a prosodically balanced corpus are empirically compared.

The output of the greedy algorithm is expected to be a balanced corpus. Nevertheless, due to the mother corpus limitations this goal is difficult or impossible to be reached. In Spanish texts the relative frequency of *paroxitones* words is tenths more than the frequency of *proparoxitones* words. In these circumstances, it is normal that the selection algorithm still outputs unbalanced results. In order to improve these results, the subcorpus has been reviewed and corrected. Some actions of this revision were semiautomatic as listed in the *Expert Guideline* that we present below. Next the results of increasing the size of the mother corpus versus the use of greedy algorithms together with expert modifications are compared.

The paper is organised as follows: first, a formalization of the different greedy algorithms is reported; next, the expert guideline is detailed and the experimental procedure and results are discussed before reaching the conclusions.

2 Text selection with greedy algorithms

2.1 Basic algorithm

The goal is to build a subcorpus from a mother corpus that has N candidates $MC = \{C_1, C_2, \ldots, C_N\}$. The subcorpus is expected to have the best M selected candidates $SC = \{S_1, S_2, \ldots, S_M\}$ so that $M < N$ and $S_i, S_j \in MC$, $S_i \neq S_j\ \forall i, j$. SC is built step by step by choosing a candidate C from MC. Initially $SC = \emptyset$ and $UC = MC$. In every step $SC = SC + \{C\}$ and $UC = UC - \{C\}$. UC contains the unused candidates.

Candidates and corpora can be characterized by a vector of integers $\bar{T} = [T_1, T_2, \ldots, T_P]$. Each element of this vector refers to a type of unit found in a corpus $\bar{T}(SC)$ or in a candidate $\bar{T}(C)$. The element $T_i(C)$ or $T_i(SC)$ is the number of units of the type i in the candidate C or in the subcorpus SC. In our case $T_i(C)$ is the frequency of prosodic units of a given type observed in C. A $solution\ function$ indicates if SC is a solution to the problem although not the optimal. A given reference target vector $\bar{T}^g$ is set so that SC is a solution iff $T_i(SC) \geq T_i^g \quad \forall i$. The problem requirements use to limit the maximum size of the corpus SC, typically with a number of candidates or a duration. Maximum size and coverage of $\bar{T}^g$ determine the stopping criteria of the algorithm.

Limitations of the mother corpus can make the coverage of $\bar{T}^g$ unfeasible. Thus, an alternative target vector $\bar{T}^f$ can be used so that:

$$\begin{aligned} \bar{T}^f &= feasibleTarget(\bar{T}^g, MC), \quad (1) \\ T_i^f &= min(T_i^g, T_i(MC)) \end{aligned}$$

A $goal\ function$ gives the value of the solution obtained in every step. We compare $\bar{T}^f$ and $\bar{T}(SC)$ using the metric

$$MissingUnits = \sum_{i=1}^{P}[max(0, T_i^f - T_i(SC))] \tag{2}$$

so that SC is a solution if $MissingUnits(SC, \bar{T}^f) = 0$.

A $selection\ function$ indicates in any moment, which of the candidates in MC is the best to be added to the set SC. As the number of elements in SC grows, $MissingUnits$ decreases but the number of units that overflow $\bar{T}^f$ increases. We compute these exceeding units as:

$$ExceedingUnits = \sum_{i=1}^{P}[max(0, T_i(SC) - T_i^f)] \tag{3}$$

An increase of $ExceedingUnits$ has a cost because it implies time to process extra information (ToBI labellig commonly takes 100-200 times real time (Syrdal et al., 2001)). Furthermore these extra units occupy the place of other units necessary to decrease $MissingUnits$. The problem then is to select the candidate C that maximizes the ratio r_C between cost and benefit:

$$C = \arg\max_{C \in UC} r_C \tag{4}$$

r_C is computed by using a heuristic rule. We have found these three heuristics in the state of the art:

$$r_C^{maxVal} = \sum_{i=1}^{P} min(L_i, T_i(C)) \tag{5}$$

$$r_C^{valVsCost} = \frac{\sum_{i=1}^{P} min(L_i, T_i(C))}{\sum_{i=1}^{P} T_i(C)} \tag{6}$$

$$r_C^{WIF} = \frac{\sum_{i=1}^{P} w_i}{\sum_{i=1}^{P} T_i(C)}, \tag{7}$$

$$w_i = \begin{cases} T_i^{-1}(MC) & ,T_i(SC) < T_i^f \\ & \&\quad T_i(C) > 0 \\ 0 & otherwise \end{cases}$$

with $L_i = max(0, T_i^f - T_i(SC))$. The heuristic $maxVal$ of equation 5 is detailed in (Matousek, Tihelka, and Rompuortl, 2008). This heuristic select the candidate that maximizes the quantity of units that left until the target without taking into account the cost entered by the exceeding units; the heuristic $valVsCost$ of equation 6 is presented in (van Santen, 1992) and it balances the ratio between valid and exceeding units. The heuristic WIF of equation 8 follows a weighting inverse frequency scheme that is discussed in (van Santen and Buchsbaum, 1997). We use the implementation of the WIF heuristic detailed in (Zhang and Nakamura, 2008). This heuristic weights unfrequent type of units so that they are selected first.

2.2 Tackling unfrequent type of units

We distinguish two stages in the operative of the basic algorithm. In the first stage, $\bar{T}(SC)$ gets to the target $\bar{T}^f$ decreasing $MissingUnits$ fast. In this stage unfrequent type of units have a secondary role unless a weighted heuristic is used. In the second stage $\bar{T}(SC)$ is close to $\bar{T}^f$ so that $ExceedingUnits$ increases. In this stage, unfrequent type of units is expected to be entered but its relative wait can be similar to the wait of other type of frequent units that are also missing. As result, there is a risk of getting a corpus SC without unfrequent units.

To avoid this we find two approaches in the literature: we can weight the elements of the vector $\bar{T}$ using a heuristic like the one presented in equation 8 or we can use the least-to-most-ordered greedy search presented in (Zhang and Nakamura, 2001). The least-to-most-ordered (LMO) greedy algorithm explodes the mother corpus into a number of P sub-corpora, so that all the candidates in the i sub-corpus have units of the type i. The search is focused in the sub-corpus that corresponds with the least frequent unit unreached of SC.

Here we propose a third alternative that consists on the use of a dynamic target vector (from now DTg greedy algorithm) in contrast to the static reference used in the basic greedy algorithm. Our greedy algorithm also searches for the target $\bar{T}^f$, but it performs the search by decomposing the target into smaller sub-targets. We sort the values of $\bar{T}^f$ in a list of B values:

$$\{T_1^s, \ldots, T_B^s\}, \quad T_b^s < T_{b+1}^s, \tag{8}$$

With $B <= P$ as we discard repetitions. We configure a list of B balanced feasible target vectors $\{\bar{T}_1^f, \dots, \bar{T}_B^f\}$ so that:

$$\bar{T}_b^f = feasibleTarget(T_b^s \cdot [1, \dots, 1], MC), \tag{9}$$

The first subproblem to solve is to get $\bar{T}_1^f$. As $\bar{T}_b^f$ is reached, we start solving $\bar{T}_{b+1}^f$. The subcorpus SC_b obtained after solving $\bar{T}_b^f$ is reused to solve $\bar{T}_{b+1}^g$. The last subproblem to be solved is, at most, $\bar{T}_B^f = \bar{T}^f$.

Small b indices are expected to refer to small T_b^s values that represent rare types of units. They are selected at the beginning so that unfrequent type of units are not discarded. The exceeding units of the first sub-problems will be reused in the following stages. In the next section we compare empirically all the proposals detailed in this section.

2.3 Quality metrics

We use the following quality metrics to compare the greedy algorithm:

$$\begin{aligned}
valUnits &= \sum_{i=1..P} min(T_i(SC), T_i^f) \\
ExcUnits &= \sum_{i=1..P} max(0, T_i(SC) - T_i^f) \\
DistTarget &= \sum_{i=1..P} |T_i(SC) - T_i^f| \\
totUnits &= \sum_{i=1..P} T_i(SC)
\end{aligned}$$

$$\begin{aligned}
\#UT &= Card(USTypes), \\
USTypes &= \{i \quad | \quad T_i^f \neq 0; T_i(SC) = 0\} \\
\#GT &= Card(GTypes), \\
GTypes &= \{i \quad | \quad T_i(SC) \geq T_i^f\} \qquad (10)
\end{aligned}$$

where $valUnits$ measures the number of the units in the target vector $\bar{T}^f$ which are obtained in SC; $ExcUnits$ measures the number of the units in SC which exceed the units specified in the target vector $\bar{T}^f$; $DistTarget$ measures the integer distance between the units in SC and the units in the target vector $\bar{T}^f$; $totUnits$ measures the number of units in SC; $Card(USTypes)$ is the number of types of units that are present in the target vector $\bar{T}^f$ but for which no unit is present in SC (UnSeen Types); and $Card(GTypes)$ is the number of type of units in SC that reach the number of units specified in the target vector $\bar{T}^f$ (Goaled Types).

	Cadena SER	United Nations	Goal Corpus
Number of news	100	3727	∼ 35
Number of sentences	618	22313	∼ 220
Number of stress groups	9812	376308	∼ 3504
Duration (minutes)	84	∼3222	≤30

Table 1: Figures of the corpora.

We have implemented five versions of the basic algorithm. Three of them use the heuristics described in equations 5, 6, 8. The other two versions select the candidates randomly and following the rule the biggest-the-best. We are referring to these versions as *maxValue*, *valVsCost*, *WIF*, *Random* and *BiggestFirst*, respectively.

We have also implemented the *LMO* greedy algorithm and two versions of the *DTg* algorithm. *DTgV1* changes the target when the preceding target is reached. *DTgV2* changes the target in every step to the one corresponding with the least frequent unit un-reached of SC. These algorithms need a heuristic that can be any of the ones explained in the previous section.

3 The corpus

The *Glissando* project requires the recording of 30 minutes of read speech to model the characteristic prosodic features of different professional radio speakers. The goal is to select a subset of text radio news whose total duration is about half an hour optimizing the prosodic units coverage. Our mother corpus was gently supplied by Cadena SER Radio Station. This corpus is limited in size but we are interested in using it because it was reviewed by an expert according to a set of style conventions (Rodero-Antón, 2003). The original corpus is written in Spanish but we have

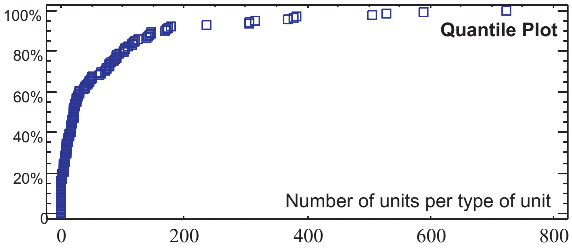

Figure 1: Quantile plot of the number of units per type of unit. Squares represent the type of unit sorted by size.

PhGInitial	SGInitial	Oxitone: 15 30 21 22	Paroxitone: 79 50 102	Proparoxitone: 0 12
	SGCentral	Oxitone: 119 104 79 108	Paroxitone: 236 316 528	Proparoxitone: 11 39
	SGFinal	Oxitone: 17 20 39 24	Paroxitone: 41 75 101	Proparoxitone: 3 10
	SGInitialFinal	Oxitone: 2 7 5 5	Paroxitone: 4 16 24	Proparoxitone: 0 0
PhGCentral	SGInitial	Oxitone: 18 45 21 26	Paroxitone: 85 64 121	Proparoxitone: 5 9
	SGCentral	Oxitone: 105 90 74 91	Paroxitone: 178 306 368	Proparoxitone: 9 26
	SGFinal	Oxitone: 9 20 27 17	Paroxitone: 64 89 144	Proparoxitone: 6 17
	SGInitialFinal	Oxitone: 5 3 7 4	Paroxitone: 16 13 28	Proparoxitone: 0 1
PhGFinal	SGInitial	Oxitone: 20 29 18 29	Paroxitone: 95 71 116	Proparoxitone: 1 7
	SGCentral	Oxitone: 145 125 117 145	Paroxitone: 307 383 589	Proparoxitone: 11 47
	SGFinal	Oxitone: 10 22 23 22	Paroxitone: 49 91 142	Proparoxitone: 10 17
	SGInitialFinal	Oxitone: 0 0 0 0	Paroxitone: 0 2 1	Proparoxitone: 1 0
PhGInitialFinal	SGInitial	Oxitone: 19 18 7 14	Paroxitone: 38 46 80	Proparoxitone: 1 5
	SGCentral	Oxitone: 171 174 140 169	Paroxitone: 377 506 724	Proparoxitone: 12 63
	SGFinal	Oxitone: 7 9 20 25	Paroxitone: 25 51 78	Proparoxitone: 3 10
	SGInitialFinal	Oxitone: 0 0 0 0	Paroxitone: 0 0 0	Proparoxitone: 0 0

Table 2: Cardinality of the type of units in the **Cadena SER** corpus. First column is the position of the phonic group in the sentence. Second column is the position of the stress group (SG) in the phonic group (PhG). Column 3, 4 and 5 correspond with the different positions of the stress in the word and size of the stress group: Oxitone (1, 2, 3 and more than 3 syllables); Paroxitone (2, 3 and more than 3 syllables); Proparoxitone (3 and more than 3 syllables).

a clone version in Catalan. To estimate the normal speed of a professional radio speaker we have used the reading of a corpus from the Cadena SER by a professional speaker.

Candidate news are prosodically analysed taking into account the type of stress groups. The stress group (SG), defined as a set of words with only one lexical stress, is a unit used in traditional Spanish prosodic studies (Navarro-Tomás, 1944), and it has been revealed to be a good prosodic unit of reference to model prosody (see for example (Garrido, 1996; Escudero and Cardeñoso, 2007)) both for Spanish and Catalan (Escudero, Cardeñoso, and Bonafonte, 2008). Other unit that serves us to count the prosodic coverage of texts is the phonic group, that refers to the stretch of speech within two pauses. It is worthwhile noting that the results obtained with the textual analysis will not probably coincide with the final reading of the speakers, since it is well known that speakers attend to more clues than punctuation marks to prosodically organise sentences.

Results obtained in previous studies (Escudero and Cardeñoso, 2007) led us to use the following prosodic features to characterize the stress groups: position of the phonic group in the sentence (Initial, Central, Final and Initial-Final), SG's position in the phonic group (Initial, Central, Final and Initial-Final), stressed syllable position within the SG (Oxiton, Paroxiton, Proparoxiton) and the number of syllables that contains the SG (one, two, three or more than three syllables). Of course other features and/or more values could be used but we were selective to avoid the combinatorial explosion and the drastic reduction of the number of samples per feature combination. The total number of possible feature combinations is 144 (some of the combinations are impossible); in other words, 144 different types of stress groups should be found in the corpus.

The corpus was automatically processed with the text analysis module of the Ogmios Text-to-speech system (Bonafonte et al., 2008) to obtain a prosodic labelling that takes into account the stress groups and their prosodic features. Table 1 shows the figures of the corpus with estimations of the size of corpus to build. The table also includes the figures of the **Radio ONU** news corpus retrieved from http://www.unmultimedia.org/radio/spanish. This corpus is used in this paper to compare the goodness of the *Expert Guideline* approach versus the classical *the biggest the corpus the best the selection* approach.

Table 2 depicts the contents of the SER corpus in terms of type of units. Figure 1 reveals the main limitation of this data: 12% of the classes has no samples and only 20% of the type of units has more than 20 units. The corpus is clearly unbalanced.

4 *Expert guideline for corpus modification*

A first analysis of the texts selected by the algorithms has revealed that without some kind of manual revision and correction the results will not improve. After the application of the greedy algorithms the selected corpus is still unbalanced. Several factors can explain it:

- as it is well-known oxytone stress word pattern is the most frequent in Spanish, followed by the paroxitone one, and very far away in the scale, by the proparoxitone one (see among others, (Canellada and Madsen, 1987). This is the reason why the number of appearances of proparoxitone stress groups is very low compared with the other types. In the selected **Cadena SER** corpus, we have 945 oxitones groups, 2316 paroxitones and only 141 proparoxitones.
- as it has been mentioned before, it is not possible to predict form the text the final segmentation in phonic groups carried out by readers, because some of the pauses introduced are not induced by puntuaction mark but by other factors such as syntactic structure and even individual decisions. For this reason, the automatic estimation of phonic groups for this task, based exclusively in punctuation marks, is necessarily temptative, only a reference to guide the selection. Using this approach, the theoretical phonic groups detected in the mother corpus tend to be rather long. This fact could explain the low number of short phonic groups in the analysed corpus (that is, GTInicialFinal, or phonic groups containing only one stress group). In the selected **Cadena SER** corpus, we have 511 initial stress groups, 2313 central, 511 final and only 67 initial-central ones. Furthermore, we have 230 one-syllable, 794 two-syllable, 1014 three-syllable and 1364 more-than-three-syllables groups.
- radio news style has their own convention to mark prosodic boundaries: texts can have long sentences without any punctuation. Besides this, a careful reading of the texts has revealed that the punctuation conventions for Spanish are not always respected in the texts written for radio news, since professional radio speakers have their o wn ones.

To balance as much as possible the number of represented stress groupsand (theoretical) phonic groups in the selected texts, without loosingthe naturalness of the contents of the original corpus, two strategieswere applied:

- modification of the text to include a greater number of proparoxytone words, while preserving as much as possible the naturalness of the contents. The procedure was straightforward: first, a list of proparoxitones Spanish names (6 entries), surnames (10 female and 6 male entries) and names of cities (9 entries) was built; then the proparoxitones proper nouns in the corpus were identified by using FreeLing [1] (58, 17, 48 and 18 entries respectively) and systematically replaced by the names of the list. The resulting texts were carefully reviewed to avoid repetitions in a given text, and strange results affecting its naturalness. Some extra names were also added or substituted manually where possible.
- slight modification of the punctuation of the texts, including some extra marks (specially in appositions and long restrictive relative clauses), but trying to keep the balance between control and naturalness of the text (too many punctuation marks in the texts whould make them unrealistic, specially considering that they are supposed to be representative of radio news). Some punctuation marks were added in some texts in order to obtain shorter phonic groups (by means of periods, semicolon an colon) and more SGs in positions other than central (by means of the use of commas, whenever the grammatical sense permits).

5 *Experimental results*

We apply greedy algorithms to the **Cadena SER** corpus and to the two other corpora: the **Radio ONU** corpus and **Modified Cadena SER** corpus. The goal is to measure

[1] FreeLing 2.1 An Open Source Suite of Language Analyzers http://garraf.epsevg.upc.es/freeling

Corpus Cadena SER

Greedy algorithm	valUnits	excUnits	DistTarget	totUnits	#UT	#GT	Duration	#SC
Random	1549 / 46%	1818/ 53.9%	2517	3367	7	4%	1732 sec	35
Basic+BiggestFirst	1597 / 45.1%	1943/ 54.8%	2594	3540	3	46%	1749 sec	26
Basic+maxVal	1696 / 47.9%	1842/ 52%	2394	3538	4	46%	1790 sec	29
Basic+valVsCost	1716 / 49.4%	1753/ 50.5%	2285	3469	4	48%	1763 sec	38
Basic+WIF	1674 / 48.8%	1755/ 51.1%	2329	3429	0	50%	1760 sec	36
DTgV1+valVsCost	1629 / 47.2%	1820/ 52.7%	2439	3449	0	51%	1773 sec	35
DTgV2+valVsCost	1645 / 48.3%	1757/ 51.6%	2360	3402	0	50%	1792 sec	37
LMO+valVsCost	1600 / 46.6%	1833/ 53.3%	2481	3433	1	54%	1772 sec	33

Corpus Radio ONU

Greedy algorithm	valUnits	excUnits	DistTarget	totUnits	#UT	#GT	Duration	#SC
Random	1580 / 46.2%	1839/ 53.7%	3401	3419	25	17%	1772 sec	36
Basic+BiggestFirst	1642 / 49.8%	1650/ 50.1%	3150	3292	18	33%	1709 sec	11
Basic+maxVal	1760 / 51.7%	1644/ 48.2%	3026	3404	17	36%	1768 sec	13
Basic+valVsCost	1910 / 55.4%	1536/ 44.5%	2768	3446	16	40%	1794 sec	52
Basic+WIF	1645 / 47.6%	1804/ 52.3%	3301	3449	8	36%	1785 sec	34
DTgV1+valVsCost	1692 / 49.1%	1748/ 50.8%	3198	3440	0	36%	1782 sec	44
DTgV2+valVsCost	1769 / 51.3%	1679/ 48.6%	3052	3448	5	31%	1792 sec	57
LMO+valVsCost	1668 / 48.8%	1750/ 51.1%	3224	3418	10	36%	1770 sec	26

Corpus Cadena SER modified with the expert guideline (changing proparoxitones proper names)

Greedy algorithm	valUnits	excUnits	DistTarget	totUnits	#UT	#GT	Duration	#SC
Random	1596 / 46.8%	1814/ 53.1%	2591	3410	5	3%	1782 sec	35
Basic+BiggestFirst	1663 / 46.9%	1876/ 53%	2586	3539	1	44%	1749 sec	26
Basic+maxVal	1784 / 49.2%	1839/ 50.7%	2428	3623	2	49%	1776 sec	29
Basic+valVsCost	1795 / 51.6%	1682/ 48.3%	2260	3477	7	47%	1751 sec	37
Basic+WIF	1762 / 50.4%	1733/ 49.5%	2344	3495	0	47%	1797 sec	35
DTgV1+valVsCost	1719 / 48.7%	1810/ 51.2%	2464	3529	0	48%	1780 sec	35
DTgV2+valVsCost	1750 / 51%	1676/ 48.9%	2299	3426	0	47%	1779 sec	36
LMO+valVsCost	1659 / 47.8%	1805/ 52.1%	2519	3464	0	52%	1775 sec	33

Table 3: Quality metrics for the different greedy algorithms. Greedy algorithm column refers to *strategy+heuristic*). *valUnits* and *excUnits* are also expressed as a percentage with respect to *totUnits*. *#GT* percentage with respect to the the total type of units.

whether the *Expert Guideline* has the potential to improve the final corpus as much as the use of a bigger corpus has.

The target vector was set to obtain an ideal balanced subcorpus that lasts at most half an hour. The total number of units in this ideal corpus is estimated taking into account the reference of the **Cadena SER** corpus (see table 1). This figure is divided by P to obtain a balanced target $\bar{T}^g = [T^g_1, .., T^g_P]$ with $T^g_i = T^g_1 \quad \forall i = 1..P$.

Table 3 shows that both **Radio ONU** and **Modified Cadena SER** improve results significantly with respect to **Radio ONU**. The selection from **Modified Cadena SER** seems to be better that the selection from **Radio ONU** (only the $DTgV2+valVsCost$ algorithm outputs better results for the **Radio ONU** corpus). Note that this table only reflects the modifications on the **Cadena SER** Corpus that concerns the proparoxitones proper names substitutions as it was explained in the *Expert Guideline* section.

Concerning to the application of the different greedy algorithms, we include *Random* and *Basic+BiggestFirst* as a worst case reference. The two versions *Basic+maxVal* and *Basic+valVsCost* improve *Random*, among other things in the number of covered classes (*#GT* metric). The two heuristics also improve *Basic+BiggestFirst* increasing the number of the valid units (*valUnits* metric). The version *Basic+valVsCost* is the best to optimize *valUnits*. The version *Basic+maxVal* is very efficient in terms of the number of candidates needed to get the final result (*#SC* metric). The use of the *WIF*, *LMO* and *DTg* strategies succeed the goal to cover unfrequent classes ($\#UT \approx 0$). In the final configuration, *DTg* outputs better results in terms of the *valUnit*, *excUnits* and *DistTarget* metrics. Both versions are more efficient in the selection of unfrequent type of units (*#UT*) metric. *DTgV2* offers

PhGInicial			
SGInitial	Oxit.: (9,13,7)(12,14,12)(10,8,14)(9,10,21)	Parox.: (28,36,16)(17,29,19)(30,40,35)	Proparox.: (0,6,3)(5,8,5)
SGCentral	Oxit.: (36,43,36)(35,44,39)(29,23,43)(26,29,71)	Parox.: (71,79,78)(100,117,110)(154,157,199)	Proparox.: (4,5,6)(15,23,33)
SGFinal	Oxit.: (4,9,4)(6,14,5)(12,16,16)(6,7,14)	Parox.: (15,20,9)(33,32,22)(37,53,52)	Proparox.: (3,6,5)(4,7,5)
SGInitialFinal	Oxit.: (2,2,3)(4,6,3)(1,3,7)(1,3,3)	Parox.: (4,5,3)(7,9,5)(9,12,8)	Proparox.: (0,1,0)(0,1,3)
PhGCentral			
SGInitial	Oxit.: (8,25,15)(27,32,7)(13,20,19)(11,22,17)	Parox.: (32,58,33)(31,56,25)(45,66,35)	Proparox.: (3,10,3)(3,16,5)
SGCentral	Oxit.: (43,63,22)(37,33,24)(24,40,29)(32,48,48)	Parox.: (65,84,64)(120,141,86)(142,196,154)	Proparox.: (5,11,6)(11,32,9)
SGFinal	Oxit.: (4,10,6)(7,25,5)(15,17,13)(10,20,15)	Parox.: (32,51,14)(38,42,34)(57,88,56)	Proparox.: (3,19,3)(7,33,9)
SGInitialFinal	Oxit.: (3,3,7)(3,4,5)(3,7,4)(3,7,3)	Parox.: (7,11,10)(4,11,7)(12,22,12)	Proparox.: (0,0,3)(1,7,4)
PhGFinal			
SGInitial	Oxit.: (8,9,11)(9,17,16)(5,9,22)(9,16,12)	Parox.: (34,33,20)(36,42,18)(38,58,32)	Proparox.: (1,6,3)(5,6,6)
SGCentral	Oxit.: (53,58,27)(41,47,26)(46,34,44)(51,42,68)	Parox.: (92,95,72)(137,117,110)(181,160,179)	Proparox.: (4,5,8)(21,29,30)
SGFinal	Oxit.: (4,8,5)(11,8,11)(10,12,11)(6,9,12)	Parox.: (18,26,8)(34,38,23)(53,67,46)	Proparox.: (5,9,9)(4,19,15)
SGInitialFinal	Oxit.: (0,0,0)(0,0,3)(0,0,3)(0,0,3)	Parox.: (0,0,3)(1,2,3)(1,5,4)	Proparox.: (1,1,0)(0,1,2)
PhGInitialFinal			
SGInitial	Oxit.: (5,6,5)(5,3,4)(3,1,4)(6,0,6)	Parox.: (10,7,15)(14,9,15)(27,12,28)	Proparox.: (1,0,3)(2,0,4)
SGCentral	Oxit.: (48,16,56)(61,21,51)(42,8,50)(57,10,67)	Parox.: (113,34,94)(170,52,132)(219,58,213)	Proparox.: (5,2,7)(23,4,18)
SGFinal	Oxit.: (3,1,3)(3,0,4)(7,2,6)(7,3,10)	Parox.: (12,7,7)(16,8,14)(20,9,30)	Proparox.: (1,3,5)(4,5,5)
SGInitialFinal	Oxit.: (0,0,2)(0,0,0)(0,0,0)(0,0,0)	Parox.: (0,0,0)(0,0,0)(0,0,0)	Proparox.: (0,1,0)(0,1,0)

Table 4: Cardinality of the type of units in the **Cadena SER** corpus. First column is the position of the phonic group in the sentence (in bold face) or the position of the stress group (SG) in the phonic group (PhG). Column 3, 4 and 5 correspond with the different positions of the stress in the word and size of the stress group: Oxitone (1, 2, 3 and more than 3 syllables); Paroxitone (2, 3 and more than 3 syllables); Proparoxitone (3 and more than 3 syllables). In parenthesis results corresponding to the **Cadena SER** (valUnits=1645) **Cadena SER Modified** (valUnits=1833) and **Radio ONU** corpora (valUnits=1769)

the best compromise between value and cost (maximum percentage for *valUnits* and minimum percentage for *excUnits* and minimal $DistTarget$). *WIF* improves *DTg* when applied to the **Cadena SER** corpus.

The news selected by the $DTgV2 + valVsCost$ algorithm (after changing proparoxitones proper names) have been analyzed according by the *Expert Guideline* obtaining the stress groups classification displayed in table 4. Although the modifications task is still in process, these results show that the selection from **Modified Cadena SER** can easily improve the selection from **Radio ONU** (the metric *valUnits* is 1833 for **Modified Cadena SER** versus 1769 for **Radio ONU**.

6 Conclusions

We conclude that greedy algorithms are useful to select text corpus tackling efficiently the problem of unfrequent type of units. Nevertheless, the modifications of the output of these selection algorithms are a need due to the intrinsic characteristics of the language. The application of a very simple *Expert Guideline* shows to be efficient to improve the selected corpus.

As a future work we plan to extend the guideline to refine results. We also expect to enrich the selected corpus with the Catalan version. The inclusion of the Catalan version is challenge to propose combined greedy algorithms and bilingual expert guidelines.

7 Acknowledgements

The authors thank César González and Enma Rodero their support producing the corpus material.

References

Bonafonte, Antonio, Asunción Moreno, Jordi Adell, Pablo D. Agüero, Eleftherios Banos, Daniel Erro, Ignasi Esquerra, Javier Pérez, and Tatyana Polyakova. 2008. The UPC TTS system description for the 2008 blizzard challenge. In *Proc of the Blizzard Challenge*, Brisbane, Australia, September.

Canellada, M. and J. Madsen. 1987. *Pronunciación del español. Lengua hablada y literaria*. Madrid: Catalia.

Escudero, D. and V. Cardeñoso. 2007. Applying data mining techniques to corpus based prosodic modeling speech. *Speech Communication*, 49:213–229.

Escudero, D., V. Cardeñoso, and A. Bonafonte. 2008. On the comparison of catalan-spanish intonation systems using statistical corpus modeling and objective metrics. In *Proceedings of Prosody 2008.*

Garrido, J. M. 1996. *Modelling Spanish Intonation for Text-to-Speech Applications.* Ph.D. thesis, Facultat de Lletres, Universitat de Barcelona, España.

Lambert, T., N. Braunshweiler, and S. Buchholz. 2007. How (not) to select your voice corpus: Random selection vs. phonologically balanced.

Matousek, J., D. Tihelka, and J. Rompuortl. 2008. Building of a speech corpus optimised for unit selection tts synthsis.

Nagorski, A., L. Boves, and Y. Steeneken. 1992. Optimal selection of speech data for automatic speech recognition systems, 2473-2476.

Navarro-Tomás, T. 1944. *Manual de Entonación Española.* Madrid, Guadarrama.

Rodero-Antón, E. 2003. *Locución radiofónica.* IORTV, 1 edition.

Syrdal, A. K., J. Hirschberg, J. McGory, and M. Beckman. 2001. Automatic tobi prediction and alignment to speed manual labeling prosody. *Speech Communications*, (33):135–151.

van Santen, J. 1992. Diagnostic perceptual experiments for text-to-speech system evaluation.

van Santen, J. and A. Buchsbaum. 1997. Methods for optimal text selection.

Zhang, J.S. and S. Nakamura. 2001. Least-to-most ordered search for minimum sentence set for collecting speech database. *Proceedings of ASJ*, pages 145–146, October.

Zhang, J.S. and S. Nakamura. 2008. An improved greedy search algorithm for the development of a phonetically rich speech corpus. *IEICE Transactions of Information and Systems*, E91-D(3):615–630, March.