Originally published in: Calzolari, Nicoletta/Choukri, Khalid/Gangemi, Aldo/Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Tapias, Daniel (Eds.): Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). 22 May - 28 May 2006, Genoa, Italy. – Paris: European Language Resources Association (ELRA). Pp. 1600-1605.

A BLARK extension for temporal annotation mining

Dafydd Gibbon*, Flaviane Romani Fernandes†, Thorsten Trippel*

*Universität Bielefeld, Germany
{gibbon,Thorsten.Trippel}@uni-bielefeld.de
† Universitade Estadual de Campinas, Brazil
f aviane@gmail.com

Abstract

The Basic Language Resource Kit (BLARK) proposed by Krauwer is designed for the creation of initial textual resources. There are a number of toolkits for the development of spoken language resources and systems, but tools for second level resources, that is, resources which are the result of processing primary level speech resources such as speech recordings. Typically, processing of this kind in phonetics is done manually, with the aid of spreadsheets multi-purpose statistics software. We propose a Basic Language and Speech Kit (BLAST) as an extension to BLARK and suggest a strategy for integrating the kit into the Natural Language Toolkit (NLTK). The prototype kit is evaluated in an application to examining temporal properties of spoken Brazilian Portuguese.

1. Introduction

The original text–based Basic Language Resource Kit (BLARK) (Krauwer, 2005) is an initial specif cation of basic tools for the development of Natural Language Processing tools which will ultimately support parsing, generation, tagging, lexicon construction, information retrieval and machine translation in local languages in a highly portable and interoperable fashion.

We extend the BLARK concept to spoken language by proposing a component for analysing temporal aspects of spoken language, with the aim of creating a "Basic Language and Speech Toolkit" (BLAST). To this end, we address the new area of Annotation Mining (AM) (Gibbon and Fernandes, 2005), proposing a subset of the basic tools required for developing databases for applications in phonetics and speech technology.

AM originated in speech technology with the treatment of annotated speech recordings in which transcription segments are aligned with segments of speech recordings in order to provide data for statistical word modelling in Automatic Speech Recognition and for unit selection techniques in Speech Synthesis.

AM uses a family of computational corpus linguistic and numerical techniques for creating second order speech resources by extracting and processing the two types of information which are available in speech and multimodal signal annotations: Categorial Annotation Mining (CAM) from annotation labels, and Temporal Annotation Mining (TAM) from annotation time–stamps. We describe components of a language–independent TAM toolkit and an initial validation comparing datasets for Brazilian Portuguese.

The background to this work is in the language resources paradigm represented by many European Commission funded projects, and in work in language archiving discussed in (Bird and Simons, 2003). The following section deals with the issue of temporal annotation mining, followed by a discussion of the functional requrements for a temporal annotation toolkit. After this, phonetic requirements are outlined in relation to a range of temporal properties of annotated speech corpora. Subsequently, design issues are discussed, and components of the toolkit are outlined. After a brief discussion of the implementation strategy, a case study is presented, in which the tools are applied to Brazilian Portuguese, followed by a conclusion.

2. Functional specif cation

The AM toolkit provides resources at a level suitable for further computational phonetic analysis and for Machine Learning (ML). This goal distinguishes AM in both content and method from conventional manual investigations of speech signals and annotations, as still practised in many traditional phonetic studies. Additionally, AM for linguistic analysis and machine learning is also distinguished from techniques of large–scale statistical processing of annotations plus signals for system training in unit–based Text–To–Speech and Automatic Speech Recognition. In specif c cases, AM methods and speech technology methods overlap, but in general linguistic analysis and speech technology have different modelling requirements with respect to annotation data. We see AM toolkits as being extensions of a BLARK, the Krauwer "Basic LAnguage Resource Kit" (Krauwer, 2005) for spoken language research and development as a "Basic Language and Speech Toolkit" (BLAST).

Examples of CAM are extensive, and are found mainly in computational corpus linguistic analyses in which transcription labels are extracted from the signal annotations and subjected to n-gram analysis, chunk parsing, concordancing procedures, etc. Examples of temporal TAM are not so common. The investigation of temporal information from speech signals has a long history in phonetics, but in general TAM techniques have only been applied sporadically to this area so far (Bird, 1999). In (Gibbon et al., 2000) temporal relations are extracted automatically from multi–tier annotations on the basis of an event logic analysis and implementation. Traditional phonetic analysis of temporal properties of speech, specif cally rhythm, are evaluated in (Gibbon and Fernandes, 2005) by applying implementations of models of these approaches to a Brazilian Portuguese corpus and comparing the results.

In order to clarify terminology (which is notoriously confusing in the annotation f eld) we use a basic ontology of three parallel information types: *signal streams*, *annotation*

tracks (of empirically observed events) and *linguistic tiers* (of theory–based predictions).

3. Phonetic requirements

The linguistic and phonetic requirements for the TAM toolkit include determination of the following:

- speech rate in terms of units of different size on different annotation tracks;
- utterance–internal speech rate variation;
- timing evenness or equality (isochrony);
- periodicity of suprasegmental patterns as a basis for analysis of rhythm and other temporal structures;
- duration types: segmental (e.g. contrastive length), syllabic (e.g. in foot structure), focus–related (correlates of nuclear and related accentuation); utterance–related (e.g. f nal lengthening, parenthesis acceleration);
- precedence and overlap relations in different annotation tracks;
- syntchronisation of functionally related different annotation tracks (e.g. duration and pitch in focus analysis).

4. Design of toolkit components

The toolkit contains tools for normalisation, comparision and statistical evaluation of data and of models of the above parameters. The following tool types are currently available:

1. Format normalisation. Annotation formats such as Praat, Transcriber are converted into a standard XML format (TASX DTD). The lattice and table data structures of annotations (Bird and Liberman, 2001), particularly those with overlapping annotation tracks, are, strictly speaking, formally too complex for context–free XML–syntax, however, and require additional semantic specif cation (e.g. length of table rows, use of pointers/links). A wealth of different and essentially incompatible XML specif cations for annotation description exit. Consequently, we normalise annotations more appropriately to a relation implemented by a tabular structure, in the simplest case the classic sequence of triples of labels and time–stamps, which is interpretable as an annotation graph. Exporting to XML (or other archiving formats) is left to choices specif c to XML users. The choice of a "realistic" tabular format has the further advantage that it can easily be further processed with conventional scripting techniques or imported into a spreadsheet application for intial post–processing and easy visualisation.
2. Linear global and local models of speech timing. Variance–based phonetic models of speech timing (including PFD, RIM, PVI) are straightforward to implement, but time–consuming to investigate manually, even with basic computational help such as spreadsheet functions. These models have been shown to be inadequate as rhythm models (though this is what they have been claimed to be), but they are still useful as measures of isochrony vs. irregularity (Gibbon, 2003a), (Gibbon, 2003b). Later, the TAM toolkit will contain experimental implementations of dynamic phonetic models of speech timing (incuding those of Barbosa, Cummins, Wachsmuth), which introduce periodicity criteria into temporal structuring.
3. Hierarchical models of speech timing. Linguistic models of prosodically relevant patterns are generally tree–structured. It is not trivial to relate these to timing structures. Two algorithms for tree–building from time–stamp sequences are included, one which builds right–headed structures, one which builds left–headed structures. A tool for quantitative comparison of the resulting tree structures is included.

The f rst evaluation of the toolset was a form of f eld evaluation by applying the tools to specif c speech data and checking for linguistic and phonetic plausibility, originally designed for examining focus in Brazilian Portuguese (Gibbon and Fernandes, 2005). Currently applications to Mandarin Chinese and to Ibibio (Lower Cross, Nigeria) are in progress. Further work in progress extends the annotation-based tools with automatic sonority–based processing (**?**) in order to move from symbol–based annotation–mining to signal–based annotation mining, and to incorporate relations between the TAM and CAM dimensions.

5. Components

An overview of the architecture of the TAM toolkit library is shown in Figure 1. The tools fall into three main classes:

1. annotation preprocessing,
2. temporal models,
3. statistical tools.

5.1. Annotation preprocessing

There are many annotation formats. The use of a standard markup language such as XML is no guarantee of compatibility. XML def nes recursive tree structures with two main tree categories: objects modelled as embedded entities, and properties of objects, modelled by attribute-value pairs attached to entities. Consequently, there are two main sources of incompatibility between actual annotations in XML format:

1. XML is a metasyntax:
   (a) the choice of categories and properties implemented as entities and their properties is application–specif c; different annotating personnel may choose different representations.
   (b) the choice of entities vs. properties for representing properties of the speech signal may vary; a common example is the treatment of the annotation label as either an entity or as a value of an attribute of an entity.

   Consequently, XML–formatted annotations of the same speech fragment created using exactly the same

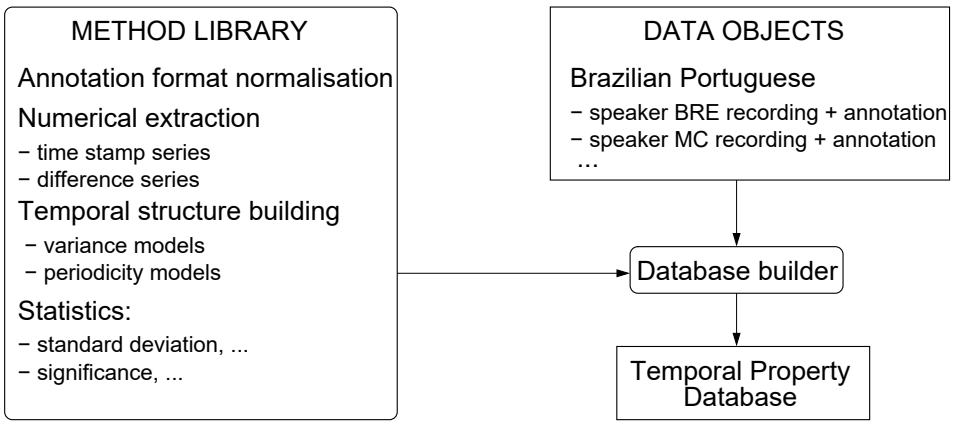

Figure 1: Temporal Annotation Mining (TAM) library overview.

criteria but by different annotation tools, such as Transcriber or the TASX Annotator, are far from isomorphic, and may only be partially converted into each other.

2. For non–tree–like objects, XML has to be supplemented by additional devices. Non–tree–like objects of the following kinds occur in speech signal annotations:

   (a) embedded tables (can be modelled informally by trees, but need more complex informal information such as identical row–branch count);

   (b) parallel annotation tiers with synchronised time–stamps;

   (c) overlapping dialogue turns.

   Each of these cases requires an additional layer of information over and above tree def nitions, adding to the formal complexity of the annotation.

The solution taken in the TAM library is to convert annotation tiers into a f at database structure rather like traditional SAM, HTK and esps/waves+ annotation conventions, consisting of triples $< label, timestamp_1, timestamp_2 >$.

5.2. Temporal models

The development of the TAM library was based on acute research and development issues in the temporal modelling of speech, not as a programme for future applications, and evaluated on specif c corpora, most extensively on Brazilian Portuguese.

The literature on timing and rhythm reveals a wide variety of phonetic models which are conventionally calculated by hand or with spreadsheets, but which can be relatively easily calculated automatically. These methods are being continually extended; a selection of specif cations of models is shown in Table 1, together with notes on their properties (see also (Gibbon and Fernandes, 2005)).

The following basic properties of rhythm models are partly based on (Gibbon and Gut, 2001), on (Gibbon, 2003b; Gibbon, 2003a) and on (Gibbon and Fernandes, 2005):

Base unit: a f nite trajectory through an n-dimensional parameter space (pitch, segment, syllable, foot sequence...).

Alternation: a dynamic, not f at base pattern trajectory, i.e. traversal through at least two positions in the parameter space (varying pitch pattern, CV syllable pattern, long–short or strong–weak syllable foot pattern,...).

Iteration the base pattern P must repeat with at least two occurrences: $P\ P^+$, i.e. any of $\{< P_1, P_2 >, < P_1, P_2, P_3 >, ...\}$.

Isochrony : equal length of the base pattern, i.e. $|P_i| = |P_{i+1}|$.

The models which have been currently implemented in the TAM library toolkit concentrate on specif c Base Units (e.g. syllable) and on the Isochrony criterion, rather than on more specif cally rhythmic properties such as Alternation and Iteration.

6. Implementation

The current implementation of the TAM library is in a combination of UNIX tools which can easily be ported to commonly used scripting languages such as Perl or Python. For current purposes, Python is the best option, since it would ensure compatibility with the widely used *Natural Language Tool Kit* (NLTK), (Bird and Loper, 2004).

The implementation of the Roach timing model (cf. Table 1) is implemented in `awk` (a less arcane predecessor of Perl):

```
#!/usr/bin/gawk -f
# stat-roach-pfd.sh
# D. Gibbon
# 2005-04-05
# Roach foot deviation model

BEGIN {
  n=ARGC-1;
  for (i=1;i<ARGC;i++) {
          sum=sum+ARGV[i];
          };
  mean=sum/n;
  for (i=1;i<ARGC;i++) {
          meandiff=mean-ARGV[i];
          if(meandiff<0) meandiff=0-meandiff;
          meandiffsum=meandiffsum+meandiff
```

Table 1: Formal comparision of Linear Rhythm Models with reference to rhythm modelling conventions.

Name	Model	*Base unit*	*Alternation*	*Iteration*	*Isochrony*
PIM	$\Sigma_{i \neq j} \left\lvert \log \frac{I_i}{I_j} \right\rvert$	foot	no	no	yes
PFD	$100 \times \frac{\Sigma \lvert MFL - len(foot_i) \rvert}{n \times MFL}$, MFL $= \frac{\Sigma_{i=1}^{n} \lvert foot_i \rvert}{n}$	foot	no	no	yes
PVI	$100 \times \Sigma_{k=1}^{m-1} \left\lvert \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right\rvert / (m-1)$	V stretches	no	yes	yes

```
        };
  avdev=meandiffsum/n;
  print 100*meandiffsum/sum;
}
```

The implementation of the Scott, Isard and Boysson-Bardies timing model (cf. Table 1) is also implemented in `awk`:

```
#!/usr/bin/gawk -f
# stat-pim.sh
# D. Gibbon
# 2005-04-05
# Scott et al Pairwise Irregularity Measure

BEGIN {
n=0;
sum=0;
ratio=0;
logratio=0;
}
$0 != "" {
n++;
interval[n]=$0;
}
END {
accum=0;
for (i=1;i<=n;i++) {
    for (j=i+1;j<=n;j++) {
        accum++
        logratio=interval[i]-interval[j];
        if (logratio<0)
            logratio=0-logratio;
            sum=sum+logratio;
           };
       };
size=n
# Incorrect in original model:
# should be size=(n*n)/2-(n/2)
average = sqrt(sum)/size
print "N:      ", size, accum
print "Average:", average
}
```

The implementation of the PVI model (cf. Table 1) also implemented in `awk`:

```
#!/usr/bin/gawk -f
# D. Gibbon
# 2005-04-05
# phon-pvi.sh

BEGIN {
sum=0;
previous=0;
n=0;
}
NR==1 && $0 != "" {
  previous=$0;
  n++;
  }
NR>1 && $0 != "" {
  difference=$0-previous;
  both=previous+$0;
  average=both/2
  normalised=difference/average
  if (normalised<0)
     normalised=0-normalised;
  sum=sum+normalised;
  previous=$0;
  n++;
  }
END {
mean=sum/(n-1);
pvi=100*mean
print pvi
}
```

7. A case study

7.1. The problem

The TAM tools described here were originally developed in order to investigate basic claims in the phonetics literature on rhythm modelling, in particular in the context of the work of (Frota and Vigário, 2000), (Duarte et al., 2001) and (Galves et al., 2002) on the rhythm of European Portuguese (EP) and Brazilian Portuguese (BP). This issue was also related to a research programme designed to investigate the role of timing in neutral and focussed sentences in BP (Sândalo and Truckenbrodt, 2002).

7.2. Corpus

The corpus used here is based on the Sândalo and Truckenbrodt corpus of Brazilian Portuguese (Sândalo and Truckenbrodt, 2002).
The corpus is constituted by 49 sentences which represent different syntactic structures: SV, SSV, SVO, SSVO, SSVOO, SVOO, SSVAO, SS and SS (Scomp)V, where

- S corresponds to subject,
- V corresponds to verb,
- O corresponds to object,
- A corresponds to adverb,
- SS corresponds to simple subject constituted by two elements (a substantive and an adjective, for instance, 'o sofá preto' - 'the black sofa'),

- OO corresponds to an only object constituted by two elements to and SS,
- and SS (Scomp) corresponds to a complex subject constituted by four elements (two substantives and two adjectives, for example: 'o tatu russo e a abelha rainha'- 'the Russian armadillo and the queen bee').

The sentences contain transitive verbs, unaccusative verbs or unergative verbs.

7.3. Speakers

The sentences were produced by f ve educated (graduate) female Brazilian speakers (age range 26 to 51 years) from different regions of Brazil:

1. Rio de Janeiro city (Rio de Janeiro State),
2. São José do Rio Preto city (São Paulo State),
3. Garça city (São Paulo State) and
4. Itumbira city (Goiás State).

Four speakers are in the age range 26 to 40 years old, and one of the speakers is 51 years old.

7.4. Methodology

The corpus creation procedure is as follows:

1. recording of the corpus sentences,
2. syllabic segmentation, phonetic transcription and annotation of these sentences.

The speakers were asked to read two kinds of prompt sentence:

1. neuter sentences, in which no word was specif cally constrained for focus by the context,
2. sentences with narrow focus on the subject.

These two types of sentences were produced by speakers on the basis of appropriate contextual triggers. The sentences with narrow focus on subject were produced as answers to questions like the following:

A:	Quem	correu?	
	'Who	runs?'	
B:	O	José	correu.
	The	Joe	runs.
		'Joe runs.'	

The neutral sentences were produced as answers to questions like:

A:	O que	aconteceu?	
	'What	happened?'	
B:	O	José	correu.
	The	Joe	runs.
		'Joe runs.'	

The speakers were instructed to produce sentences naturally, neither slowly nor quickly. The complete corpus consists of 490 sentences (49 neuter sentences and 49 sentences with focus on subject, both types produced by the 5 speakers).

The recording of the sentences was made on a PC using the Praat speech analysis workbench[1] and with a good acoustic capture microphone, Leson SM 48.
The syllabic phonetic annotation was made using Praat, according to conventions of the alphabet of International Phonetic Association (IPA) represented by the SAMPA ASCII conventions. Two special conventions were adopted:

1. 'N' represents nasalization of the preceding vowel,
2. 'R' represents 'r' consonant in coda position.

The choice for 'R' to indicate every variation of 'r' in coda position was made because the speakers of different regions of Brazil showed some variations in the production of 'r' in syllabic coda and this detail was not considered important to the objectives of the study.

7.5. Evaluation

In the present context, evaluation means the evaluation of the toolkit, rather than evaluation of the data; evaluation of the data is a means to an end, here, rather than an end in itself.
The data were successfully processed by the implementations of speech timing (rhythm) models, producing the partly surprising result that the different models of speech rhythm did not correlate too well with each other. The result was unsurprising, in that different models might be expected to yield different results; on the other hand, the result was surprising in that the models claimed to model the same relationships between speech segments in terms of variation from isochrony.
The results are reported in (Gibbon and Fernandes, 2005).

8. Conclusion and future prospects

In this study, we set out to extend the idea of the originally text–based Basic Language Resource Kit (BLARK) (Krauwer, 2005) to spoken language resources proposing a set of tools for analysing the main characteristic formal property of spoken language, namely timing. In this respect, the study has continued work in the new area of Annotation Mining (AM), in proposing a subset of the basic tools required for developing databases for applications in phonetics and speech technology.
The tools were both motivated by and evaluated against a Brazilian Portuguese corpus within the context of projects for investigating timing in neutral and focussed sentences, and also differences in rhythm-determining factors in European Portuguese and Brazilian Portuguese. In the course of this investigation, it was discovered that the formalisation which the implementation constraint on the traditional approaches imposed led to the uncovering of both formal and empirical problems with the implemented models, due to incomplete specif cations of concepts such as rhythm in the original approaches.
The broader implications of this work are situated within two main methodological areas:

[1] www.fon.hum.uva.nl/praat/praat5133.html

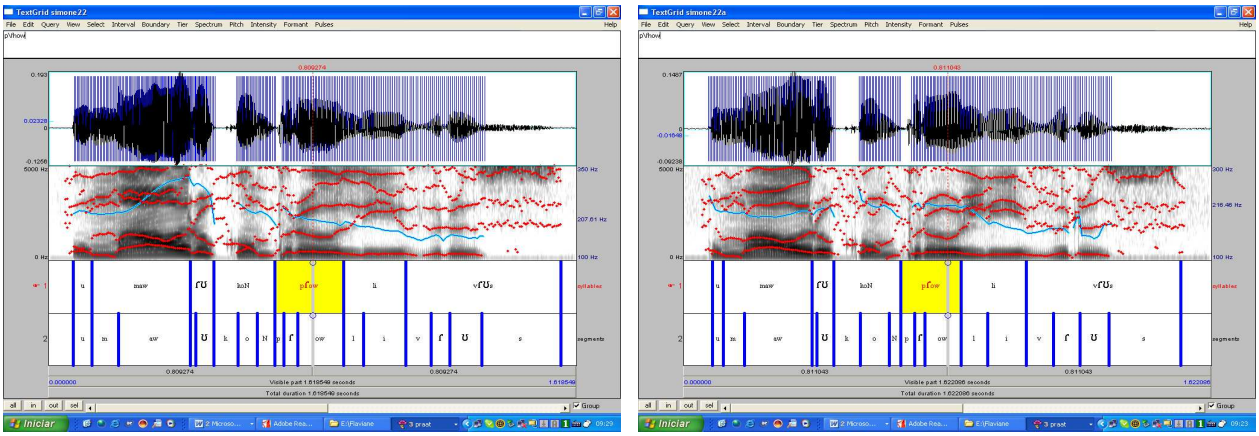

Figure 2: Annotated Brazilian Portuguese sentence: O Mauro comprou livros. '[the] Mauro bought books.' in focussed (left) and neutral (right) contexts, showing different temporal patterning.

1. the spoken language resource paradigm, with the goal of extending toolkit concepts such as the BLARK and NLTK which were discussed in the introduction, and, in the mid–term, integrating the tools into a seamless environment with well–def ned databases as interfaces between tools, and interoperability determined by implementation in a generally used scripting language such as Python;

2. the subdiscipline of Computational Phonetics, with the goal of replacing laborious manual measurement and calculation with automatic procedures as far as possible, initially on the basis of previously created annotations, but ultimately also using automatic spoken language segmentation algorithms.

The current prototype toolkit is being extended by the addition of other types of temporal analysis tool; on completion of the initial development, the toolkit can be ported to Python and the feasibility of interfaces with NLTK tools will be examined, with a view to providing a spoken language oriented modules for NLTK. It is envisaged that the BLARK concept will be effectively def ned as a subset of the extended NLTK, and that the spoken language oriented extensions to BLARK will constitute a "Basic Language and Speech Toolkit" (BLAST), comprising both written and spoken language elements of the NLTK.

9. References

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23–60.

Steven Bird and Edward Loper. 2004. Nltk: The natural language toolkit. In *Proceedings 42nd Meeting of the Association for Computational Linguistics (Demonstration Track)*, pages 214–217, Barcelona.

Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582.

Steven Bird. 1999. Multidimensional exploration of online linguistic f eld data. *Proceedings of the 29th Meeting of the North–East Linguistic Society*, 26:33–50.

D. Duarte, A. Galves, N. Lopes, and R. Maronna. 2001. The statistical analysis of acoustic correlates of speech rhythm. The Sciences of Complexity Conference, ZiF, Bielefeld.

Sonia Frota and Maria Vigário. 2000. Aspectos de prosódia comparada: ritmo e entoação no PE e no PB. In Rui v. Castro and Pilar Barbosa, editors, *Actas do XV Encontro Nacional da Associação Portuguesa de Lingüistica*, volume 1, pages 533–555, Braga. APL.

A. Galves, J. García, D. Duarte, and C. Galves. 2002. Sonority as a basis for rhythmic class discrimination. In *Speech Prosody 2002*, Aix–en–Provence.

Dafydd Gibbon and Flaviane Romani Fernandes. 2005. Annotation–mining for rhythm model comparison in brazilian portuguese. In *Proceedings of INTERSPEECH–EUROSPEECH 2005*, Lisbon.

Dafydd Gibbon and Ulrike Gut. 2001. Measuring speech rhythm in varieties of English. In *Proceedings of EUROSPEECH 2001*, pages 91–94, Aalborg.

Dafydd Gibbon, Harand Lüngen, and Andreas Witt. 2000. Enhancing speech corpus resources with multiple lexical tag layers. In *Proceedings of LREC 2000*, Athens.

Dafydd Gibbon. 2003a. Computational modelling of rhythm as alternation, iteration and hierarchy. In *Proceedings of ICPhS 2003*, Barcelona.

Dafydd Gibbon. 2003b. Corpus–based syntax-prosody tree matching. In *Proceedings of EUROSPEECH 2003*, Geneva.

Steven Krauwer. 2005. ELSNET and ELRA: a common past and a common future. `http://www.elda.org/article48.html (14.10.2005 06:27:33)`.

M. F. Sândalo and Hubert Truckenbrodt. 2002. Some notes on phonological phrasing in Brazilian Portuguese. *MIT Working Papers in Linguistics*, 42.