Parsing Spontaneous Speech

Rodolfo Delmonte
Department of Language Sciences
Università Ca' Foscari
Ca' Garzoni-Moro - San Marco 3417 - 30124 VENEZIA
e-mail: delmont@unive.it website - http://project.cgm.unive.it

ABSTRACT

In this paper we will present work carried out lately on the 50,000 words Italian Spontaneous Speech Corpus called AVIP, under national project API, made available for free download from the website of the coordinator, the University of Naples. We will concentrate on the tuning of the parser for Italian which had been previously used to parse 100,000 words corpus of written Italian within the National Treebank initiative coordinated by ILC in Pisa.

The parser receives as input the adequately transformed orthographic transcription of the dialogues making up the corpus, in which pauses, hesitations and other disfluencies have been turned into most likely corresponding punctiation marks, interjections or truncation of the word underlying the uttered segment.

The most interesting phenomenon we will discuss is without any doubts "overlapping", i.e. a speech event in which two people speak at the same time by uttering actual words or in some cases nonwords, when one of the speakers, usually the one which is not the current turntaker, interrupts the current speaker.

This phenomenon takes place at a certain point in time where it has to be anchored to the speech signal but in order to be fully parsed and subsequently semantically interpreted, it needs to be referred semantically to a following turn.

1. INTRODUCTION

This paper present work carried out at the University of Venice for the creation of tools for the annotation of spoken Italian which allow the user to work in a format fit for the visualization of the results in multilevels representation in commercial browsers. The specific topic of this paper will be the characterization of overlappings along the lines of what has been done in MATE project and other international projects in progress like the MEETING project. In the AVIP/API dialogues the quantity of overlapping speech is very high, as has been reported in the national conference on "Parlato Italiano" – Naples, 13-15 February, 2003. At an international level, even though everybody agrees on the relevance of the phenomenon, there is not a universal agreement on its representation from the linguistic point of view, in particular as concerns syntactic structure both at constituent and functional level. The problem of overlappings is usually associated in the English language to well-defined prosodic phenomena usually related to turn-taking by a speaker: it is our intention to study there aspects per spoken Italian. In the last few years, in the field of spoken dialogue corpus annotation, level-specific coding tools gradually emerged - for morphosyntactic annotation, co-reference annotation, dialogue acts annotation etc., as described in the MATE (Multi-level Annotation Tools Engineering) project report on the state of the art in spoken dialogue annotation tools. All of those tools, however, were either completely level-specific or very limited as regards their multi-level coding capabilities. To our knowledge, the MATE Workbench which appeared in 2000 is still the only fully multi-level and cross-level spoken language dialogue coding tool around. However, this tool still has important limitations, such as being fragile and without an appropriate user interface for the average user.

So far, however, no project has succeeded in producing a really useful general-purpose tool for coding and analysing full natural interactivity data. NITE (Natural Interactivity Tools Engineering) is one of the projects which currently address the challenge just described. NITE is a European HLT (Human Language Technologies) project which began its work in April 2001. The goal of NITE is to develop a workbench, or an integrated set of tools, for annotating and analysing full natural interactive communication among humans and between humans and systems. The annotated corpora can then be used and re-used to advance our understanding of complex natural interactive communicative behaviour, train natural interactive system components, etc. In many ways, NITE pursues the same objectives as its predecessor project MATE. The main difference is that NITE goes beyond spoken dialogue coding and analysis to full natural interactivity data annotation and analysis. The NITE objectives thus are: to develop a markup framework; identify, or develop, a number of natural interactivity best practice coding schemes to be described followingthe markup framework; and build a general-purpose natural interactivity annotation and analysis toolset which includes those coding schemes and supports the addition of new ones within the general boundaries of the markup framework. The NITE Project is funded by the European Commission to provide infrastructural technology for working with heavily cross-annotated multimodal data sets. This effort shares much in common with both the Annotation Graph Toolkit (Ma, Lee, Bird, & Maeda, 2002) and with ATLAS (Laprun, Fiscus, Garofolo, & Pajot, 2002). However, in keeping with the aim of supporting work with heavily cross-annotated data sets, NITE model allows easier access to rich structural information about the data than these other systems.

Also our work is motivated by the sorts of data modelling concerns that are raised by having many kinds of annotation, for linguistic levels ranging from phonology to pragmatics, on the same basic speech or language material. There are two reasons why such cross-annotation is prevalent: first, corpora are expensive to collect even without annotating them; projects tend to reuse collected materials where they can. Second, with the advent of statistical methods in language engineering, corpus builders are interested in having the widest possible range of features to train upon. Understanding how the annotations relate is essential to developing better modelling techniques for our systems. The AVIP/API corpus, is one example of a corpus that has been prepared to answer these questions, with annotations that range from orthography and syntax to reference and dialogue structure.

Although how annotations relate to time on signal is important in corpus annotation, it has not been targeted specifically in our previous projects for the inherent difficulty of putting in direct relation abstract representations beyond word level with those at phone/word level like phonetic, phonological and prosodic representation. In particular phrase structures and sentences, are essentially structures built on top of other annotations (in these cases, the words that make up an orthographic transcription) and have to derive their timings from the annotations on which they are based. Tree structures are common in describing a coherent sets of tags, but where several distinct types of annotation are present on the same material (syntax, discourse structure), the entire set may well not fit into a single tree. This is because different trees can draw on
different leaves (discours moves, words) and because even where they share the same leaves, they can draw on them in different and overlapping ways (e.g.,disfluency and overlapping structure and syntax in relation to words).

The problem of overlapping annotation has required a new coding of all the corpus AVIP/API in order to recover the temporal alignment of the phenomenon under study. Previously, all overlappings had been marked symbolically locally but they had been ascribed and moved in their linguistic form to the turn of their respective speaker. Our research activity has covered the items in the following preliminary list:

A. Elaboration and transformation of original texts

- normalization of texts containing dialogue transcription
- translitteration of orthophonetic transcriptions in a standard orthographic format and creation of standard translitteration protocols
- transformation of texts with overlapping organized on a dialogic basis (its content being assigned to the respective speaker), into texts with the overlapping temporally aligned with the corresponding acoustic signal
- coding of the input file for the subsequent multilevel linguistic analysis in XML format adequate for its visualization in a standard commercial browser by means of href linking
- creation of a file containing correspondences of all overlappings in XML format, between the original separate encoding of overlappings ascribed to each speaker in terms of turns and the transformed orthographic file where overlappings are encoded locally in each turn on a temporal basis.

B. Linguistic multilevel representation of each text at sentence level

- lexical annotation with association of lemmata to each wordform; association of a syntactic and a semantic class to each lemma;
- morphological annotation of each wordform with association of morphological features;
- syntactic annotation in bracketed constituents
- functional annotation in grammatical functions and transformation of the syntactic file containing wordforms of the orthographic text in a semantic representation into head lemmata and their features.
- anaphoric annotation of coreference between all referring expressions, both nominal and pronominal ones, without any restriction on the type of reference as decided in the original MapTask, including both explicit and implicit linguistic elements.

Creation of appropriate protocols for the translitteration and normalization of orthophonetic transcriptions. Description of the protocols under A. and their publication on the web. Creation of appropriate DTD and stylesheets for the coding of texts in XML format. Creation of the DTD and stylesheets used for the coding of texts in XML format elaborated under A. and the multilevel annotations created under B. and their publication on the web.

Here below we report the five level of xml annotation for the sentence *C'è un cagnolino NOISE nell'angolo sinistro*/"There's a puppy NOISE in the left corner", starting from the "orthophonetic transcription":

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<phon_orthotranscription id="a01_n">
<phw id=" pw_01 " type="yes"> C </phw>
<phw id=" pw_02 " type="yes"> e' </phw>
<phw id=" pw_03 " type="yes"> un </phw>
<phw id=" pw_04 " type="yes"> cagnolino </phw>
<phw id=" pw_05 " type="no"> RUMORE </phw>
<phw id=" pw_06 " type="yes"> nell </phw>
<phw id=" pw_07 " type="yes"> angolo </phw>
<phw id=" pw_08 " type="yes"> sinistro </phw>
<phw id=" pw_09 " type="yes"> . </phw>
</phon_orthotranscription>
```

than the orthographic transcription,

```
<orthotranscription id="a01_n">
<turn id="G001">
<w id=" w_01 " case="cap"> G001 </w>
<w id=" w_02 " case="cap"> C </w>
<w id=" w_03 " case="low"> e' </w>
<w id=" w_04 " case="low"> un </w>
<w id=" w_05 " case="low"> cagnolino </w>
<w id=" w_06 " case="low"> nell </w>
<w id=" w_07 " case="low"> angolo </w>
```

<w id=" w_08 " case="low"> sinistro </w>
<w id=" w_09 " case="punt"> . </w>
</turn>
</orthotranscription>

Here is the morphological, syntactic and semantic features and lemmatized transcription,

<mword_file id="mfeats.xml">
<mw id="mw_0000" pos="I" mfeats="turn" href="orton.xml#id(w_01)"> G001</mw>
<mw id="mw_0001" pos="N" mfeats="ms" lemma="ci" sfeats="nh" sems="hum" href="orton.xml#id(w_02)"> C</mw>
<mw id="mw_0002" pos="V" mfeats="KL3s" lemma="essere" sfeats="vc" sems="cop" href="orton.xml#id(w_03)"> e'</mw>
<mw id="mw_0003" pos="D" mfeats="fs" lemma="un" sfeats="art" sems="ind" href="orton.xml#id(w_04)"> un</mw>
<mw id="mw_0004" pos="N" mfeats="ms" lemma="cagnolino" sfeats="n" sems="com" href="orton.xml#id(w_05)"> cagnolino</mw>
<mw id="mw_0005" pos="P" mfeats="fs" lemma="in" sfeats="partd" sems="def" href="orton.xml#id(w_06)"> nell</mw>
<mw id="mw_0006" pos="N" mfeats="ms" lemma="angolo" sfeats="n" sems="com" href="orton.xml#id(w_07)"> angolo</mw>
<mw id="mw_0007" pos="A" mfeats="ms" lemma="sinistro" sfeats="ag" href="orton.xml#id(w_08)"> sinistro</mw>
<mw id="mw_0008" pos="PU" mfeats="punto" href="orton.xml#id(w_09)"> .</mw>
</mword_file>

then the syntactic constituency transcription,

<frase init ="mw_0000" end= "mw_0008">
<nodo id="sw_0001" type="CP"><foglia href="mfeats.xml#id(mw_0000)"> G001</foglia>
<nodo id="sw_0002" type="F">
<nodo id="sw_0003" type="IBAR"><foglia href="mfeats.xml#id(mw_0001)">c</foglia><foglia href="mfeats.xml #id(mw_0002)"> e'</foglia></nodo>
<nodo id="sw_0004" type="COMPC">
<nodo id="sw_0005" type="SN"><foglia href="mfeats.xml#id(mw_0003)">un</foglia><foglia href="mfeats.xml# id(mw_0004)"> cagnolino</foglia></nodo>
<nodo id="sw_0006" type="SP"><foglia href="mfeats.xml#id(mw_0005)">nell</foglia>
<nodo id="sw_0007" type="SN"><foglia href="mfeats.xml#id(mw_0006)">angolo</foglia>
<nodo id="sw_0008" type="SA"><foglia href="mfeats.xml#id(mw_0007)">sinistro</foglia><foglia href="mfeats.xml#id(mw_0008) ">.</foglia>
</nodo></nodo></nodo></nodo></nodo></nodo>
</frase>

2. Incremental Shallow-to-Deep Parsing

Shallow or partial parsing produces minimal and incomplete syntactic structures, often in an incremental descriptive schema. In order to repeat some if not all of the features successfully analysed by full GETARUNS, we need to extend shallow parsing to deeper language analysis, while preserving robustness. In order to tackle deeper linguistic aspects we assume the following are essential requisites to fulfill:

- structural information must be extended in order to recover clause-level structure safely;
- lexical information should be tapped in order to help differentiate arguments from adjuncts; i.e. the lexicon should contain full subcategorization frames for most if not all verb, adjective, noun predicates that require them;
- grammatical functions should also be mapped onto the syntactic representation in order to take advantage of fundamental distinctions these descriptions afford: predicative vs. non-predicative functions are distinguished thus allowed a correct semantic mapping to take place.

As in most shallow parsers, we use a sequence or cascade of transducers: however, in our approach, since we intend to recover sentence level structure, the process goes from partial parses to full parses. Sentence and then clause level is crucially responsible for the right assignment of arguments and adjuncts to a governing predicate head. This is clearly paramount in our scheme which aims at recovering predicate-argument structures, besides performing a compositional semantic translation of each semantically headed constituent.

So the first parser receives the input sentence split by previous processors, which is recursively/iteratively turned into a set of non-sentential level syntactic constituents - some of which can incorporate a PP headed by "of". Other operations solved at constituent level is that of collecting under the same constituent structure head level coordinate stuctures separated by "and/or".

Non-sentential level constituents, can be interspersed by heads beginning subordinate clause markers, like subordinating conjuctions, or parentheticals - by punctuation, indirect interrogative clauses - by interrogative pronouns. The final output is a list of headed syntactic constituents which comprise the usual set of semantically translatable constituents, i.e., ADJP, ADVP, NP, PP, VC (Verb Cluster). In addition to that, sentence level markers interspersed in the output are the following:

- FINT, interrogative clause marker;
- DIRSP, direct speech clause marker;
- FP, parenthetical clause marker;
- FC, coordinate clause marker;
- FS, subordinate clause marker;

- F2, relative clause marker.

The task of the following transducer is that of collapsing into the corresponding clause the clause material following the marker up to some delimiting indicator that can be safely taken as not belonging to the current clause level. In particular we assume that at each sentence level only one VCluster can appear: we define the VC as IBAR indicating that there must be a finite or tensed verb included in it. VClusters containing non-tensed verbal elements are all defined separately,

- SV2, for infinitive VCs;
- SV5, for gerundive VCs;
- SV3, for participial VCs.

The second transducer has also two additional tasks: it must take care of ambiguity related to punctuation markers such as COMMA, or DASH, which can either be taken as beginners of a parenthetical or indicators of a list, or simply as separators between main clause and subordinate/coordinate clause. It has also the task of deciding whether conjunctions indicated by FC or by FS are actually starting a clause structure or rather an elliptical structure.

The third pass is intended to produce an improvement on the sentence-level full parse, by transducing each constituent label into a corresponding grammatical function label. The rules are the following, and are taken from the inventory LFG theory and follow its rules and principles. In order to account for the ambiguous labelling of NPs, we use a logical flag associated to IBAR: it is set to false at the beginning of the parser; when the first NP is met and ibar(false) has success, it will be turned into SUBJ. When the IBAR is taken the flag is set to true so that the following NP will be turned into OBJ. We also compute another important feature of IBARs: their passivity. So whenever a passive IBAR is taken, we do not expect a following NP to belong to that clause level, but rather to the following one. Grammatical functional labels are then the following:

- ADJPs are turned into ACOMP;
- ADVPs are turned into ADJ;
- NPs are turned into SUBJ, in case the ibar flag is set to false; and into OBJ in case the ibar flag is set to true;
- PPs are turned into OBL;
- SV2, SV5, SV7, are all turned into VCOMP.

Some of these functional labels may undergo further changes when subcategorization is looked up in the lexicon: in particular,

- OBJs may become NCOMP;
- OBLs may become PCOMP;
- ADJs may become ADVCOMP.

Finally the fourth pass has the task of splitting complex sentences into simplex ones, or clauses. This may require recovering IBAR and complement structures following a relative clause or a subordinate clause functioning as noun complement, and rejoining it to its subject while preserving control information. This level as the previous ones may lead to failures, which is recovered by simply considering all functions as belonging to the same clause and using IBARs as filters, by means of subcategorization.

The output of the four transducers is passed to the algorithm that takes care of the creation of predicate-argument structures which has the additional task of taking into due account interclausal relations. To do that, semantic indices of governing predicates are used to assert dependencies between two adjacent clauses. This may also apply to a main clause and a clause-like adjunct like a gerundive or a participial.

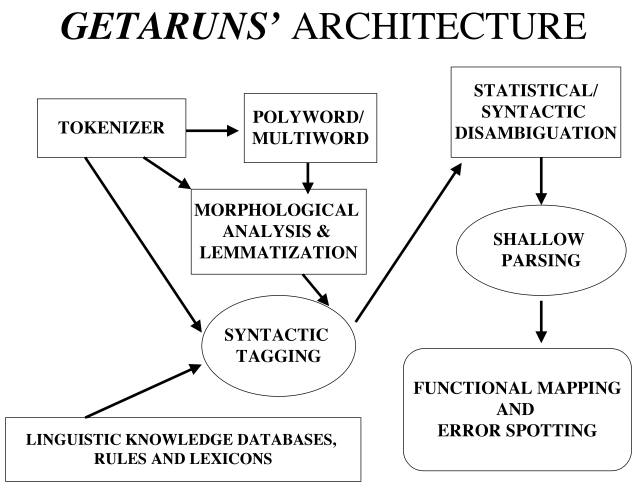


Fig. 1 GETARUNS Shallow Parser Architecture

3. Overlappings

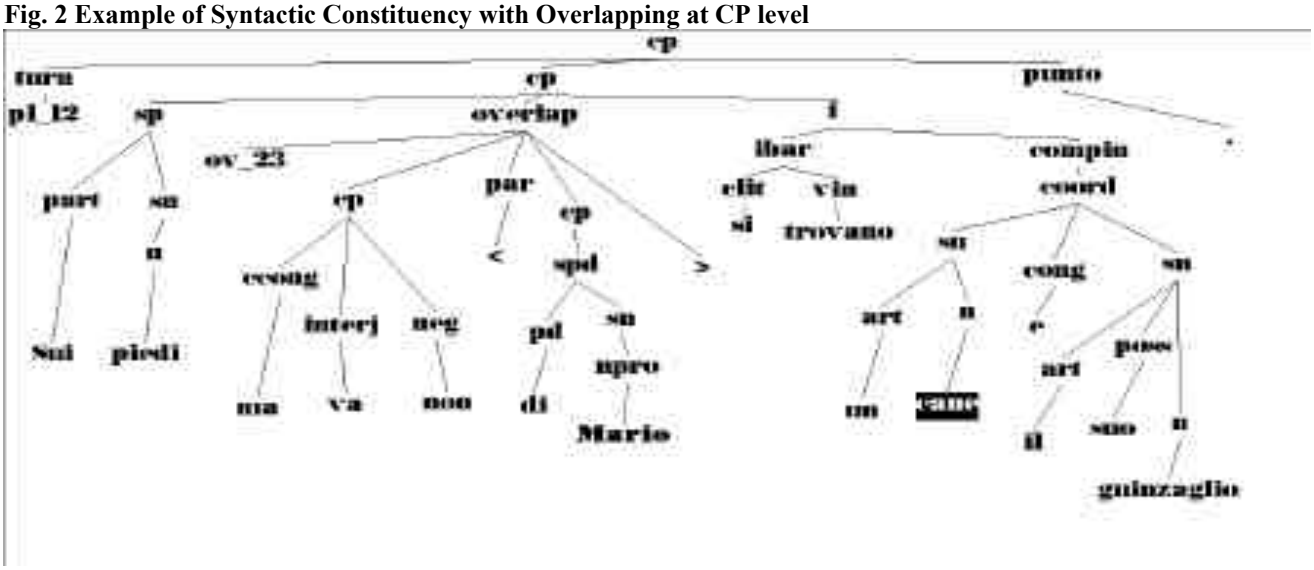
One of the distinctive characteristics of naturalistic conversation (in contrast to monolog situations) is the presence of overlapping speech. Overlapping speech may be of several types, and affects the flow of discourse in various ways. An overlap may help to usurp the floor from another speaker (e.g., interruptions), or to encourage a speaker to continue (e.g., back channels), or simply end up just in an attempt at usurping the floor without success (Vain Interruption as defined by Bazzanella). In our work we have explored types of overlaps and their physical parameters, including prosodic aspects. Note that for the purpose of this research we use the terminology proposed by E.Shriberg et al. who individuate spurt units based solely on observable temporal stretches of overlapping speech where we use the neutral terms "jump-in points" and "jump-in words" to specify overlap onsets of spurts. This is to avoid any confusion with terminology taken from pragmatics and the turn-taking literature that refers to turn units, since there is not a one-to-one mapping between spurts and turns.

Speaker overlaps, are directly observable in our data, since by definition overlaps occur at points of simultaneous speech on more than one of the (individually recorded) channels, besides their explicit indication in the orthophonetic transcription thus translitterated into the orthographic transcription. What we are interested in is

finding out whether there is any correlation between the onset of overlaps and their possible characterization from the point of view of syntactic structure, which we have preliminarly proposed to treat by introducing a node of discourse constituency called OVLP (overlapping), from where the two temporally aligned components of overlapping branch, the overlappee and the overlapper stretch of speech/text. The typologies proposed in the
English literature and those suggested by Bazzanella will be verified in relation to their treatment at the level of syntactic constituency.Both punctuation and overlap have been discussed in the literature as correlating with prosodic cues. For example, past computational work has discussed prosodic features for sentence boundaries as well as disfluency boundaries. Past work in conversation analysis, discourse analysis, and linguistics has shown prosody to be a useful cue in turn-taking behavior.

Here below is an example of a syntactic constituency structure with the sentence *Sui piedi ov_23 ma va non di Mario > si trovano un cane e il suo guinzaglio*/"On the feet ov_23 but come on not Mario's ones > you find a dog and its lead" where the main discourse constituent OVERLAP has been integrated in the CP level constituency. This implements principles of linguistic representatiion expressed in our previous work, in particular in Delmonte, 1987, where syntactic structure was to interact with conceptual and pragmatic structure in order to take into due account phenomena like Contrastive and Emphatic Focus.

Fig. 2 Example of Syntactic Constituency with Overlapping at CP level



4. References

Carla Bazzanella, 1994. "LE INTERRUZIONI", in Le facce del parlare. Un approccio pragmatico all'italiano parlato, Cap 8, Firenze/Roma: La Nuova Italia, ristampa 1996.

Bernsen, N. O., Dybkjaer, L. and M. Kolodnytsky (2002). The NITE Workbench - A Tool for Annotation of Natural Interactivity and Multimodal Data. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas.

Bard, E.G., Anderson, A.H., Sotillo, C., Aylett, M., Doherty-Sneddon, G. & Newlands, A. (2000) Controlling the intelligibility of referring expressions in dialogue, Journal of Memory and Language , 42(1) , 1-22.

Flammia, G., & Zue, V. (1995). N.b.: A Graphical User Interface for Annotating Spoken Dialogue. In J. Moore, M. Walker, M. Hearst, L. Hirschman, & A. Joshi (Eds.), Working Notes from the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation (pp. 40-46). Palo Alto: AAAI.

Hindle D. 1993. "Deterministic parsing of syntactic nonfluencies", In Proc. ACL, pages 123-128.

Laprun, C., Fiscus, J. G., Garofolo, J., & Pajot, S. (2002, May). A Practical Introduction to ATLAS. Paper presented at the 3rd International Conference on Language Resources and Evaluation (LREC), Las Palmas.

Lickley,R., "Detecting Disfluency in Spontaneous Speech", Thesis, Department of Linguistics, University of Edinburgh 1994.

Ma, X., Lee, H., Bird, S., & Maeda, K. (2002). Models and Tools for Collaborative Annotation. Paper presented at the Third International Conference on Language Resources and Evaluation.

McKelvie,D., "The Syntax of Disfluency in Spontaneous Spoken Language", HCRC Research Paper HCRC/RP-95, Edinburgh, 1998.

N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The ICSI Meeting Project. In Proceedings of the Human Language Technology Conference, San Diego, 2001, p.10-18.

Readings in Corpus Linguistics, ed. G. Sampson and D. McCarthy, London and NY: Continuum International, 2002. Originally circulated on the web in 2000.

Shriberg, E.; R. Bates, and A. Stolcke. A prosody-only decisiontree model for disfluency detection. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, Proc. EUROSPEECH, vol. 5, pp. 2383–2386, Rhodes, Greece, 1997.
Shriberg, E.; A. Stolcke, and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, Proc. EUROSPEECH, vol. 2, pp. 1359–1362, Aalborg, Denmark, 2001.
Wheatley, B., G. Doddington, C. Hemphill, J. Godfrey, E.C. Holliman, J. McDaniel, and D. Fisher, "Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech", Proc. ICASSP-92, Vol. I, 533-53
Delmonte R. (2000), SLIM Prosodic Automatic Tools for Self-Learning Instruction, *Speech Communication* 30, 145-166.
Delmonte R., (2002), A Prosodic Module for Self-Learning Activities, *Proc. SpeechProsody2002*, Aix-en-Provence, 243-246.
Delmonte R., (1987), The Realization of Semantic Focus and Language Modeling, in *Proc. Xith ICPhS*, 1987, Tallinn (Estonia), Vol.2, 101-104.
http://www.icp.grenet.fr/ELRA/home.html
http://www.iet
http://www.ilc.pi.cnr.it/EAGLES/
http://www.cs.rochester.edu:80/research/trains/annotation/
http://www.hcrc.ed.ac.uk/maptask.html
http://mate.nis.sdu.dk/
http://www.darmstadt.gmd.de/ rostek/tatoe.hml
http://www.nis.sdu.dk
http://www.icsi.berkeley.edu/speech/mtgrcdr.html
http://www.etca.fr/CTA/gip/Projets/Transcriber/