Prosodically Enriched Text Annotation for High Quality Speech Synthesis

Dimitris Spiliotopoulos[♣], *Georgios Petasis*[♦] *and Georgios Kouroupetroglou*[♣]

[♣]Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Panepistimiopolis, Ilisia, GR-15784, Athens, Greece
{dspiliot,koupe}@di.uoa.gr

[♦]Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research “Demokritos”
P.O. Box 60228, Ag. Paraskevi 153 10, Athens, Greece
petasis@iit.demokritos.gr

Abstract

Linguistically enriched text generated from natural language modules contributes significantly on the quality of speech synthesis. For all cases where such modules are not available, such enriched input needs to be produced from plain text in order to maintain quality. This work reports on a framework of several combined language resources and procedures (word/sentence identification, syntactic analysis, prosodic feature annotation) for text annotation/processing from plain text. Using that, the implementation of an automatic XML formatted output generation module produces the prosodically enriched markup.

1. Introduction

Concept-to-Speech (CtS) systems produce annotated text from the Natural Language Generator (NLG) component as input for the speech synthesis module [1]. The NLG output text is generated as error-free syntactically annotated text exhibiting full disambiguation. In addition, further linguistic information may be generated providing considerable aid to guide synthesis. CtS systems, as a result, utilize the linguistic features from the natural language generation phase in order to produce significantly improved synthesized speech [2].

On the other hand, general purpose Text-to-Speech (TtS) systems use plain text as input utilizing language processing components such as segmentation modules and part-of-speech taggers to analyze the text input prior to synthesis. These modules unavoidably exhibit inherent statistical errors that are due to their design as well as due to language ambiguity. Apart from that, the language processing modules embedded in such systems are not usually designed to identify and extract higher linguistic information. As a result, the derived synthesized speech quality suffers when compared to the respective output of a CtS system.

On the other hand, the vast majority of existing CtS systems is designed to operate in specific thematic domains and their accuracy may reduce significantly when asked to function for other domains or generalized use. This is mostly a restriction of the natural language generator components that are intended for limited domain text generation. On the other hand, TtS systems that may produce speech from existing texts are most commonly utilized for limited domains as well as generalized use.

Previous works show that linguistically annotated text input for a TtS system can lead to improved naturalness of speech output [3], [4]. When such input can be provided, the language processing from the TtS system can be superseded.

In this work, a workflow for annotating plain text is constructed, essentially producing enriched text input for synthesis, similar to the one generated by the natural language component of a CtS system.

This task necessitates the exploitation of a major set of tools and resources for language engineering as well as an expandable platform to host, manage and overview the process stages, namely the *Ellogon* Language Engineering Platform [5]. We report on the set of linguistic features and information that needs to be considered and the description of the architecture and key modules of the Ellogon platform. Furthermore, the nature of the syntactic analysis and prosodic feature incorporation are explored. Finally, the resulting markup description derivation and format are discussed in detail.

2. Enriched text input for speech synthesis

TtS systems generally accept plain (also referred to as “raw”) text as input, using specialized algorithms to internally generate the needed natural language data prior to synthesis. However, the algorithms that are usually implemented for such tasks are not powerful enough to broadly identify additional information about several linguistic phenomena from the plain text form, thus limiting the depth of text analysis and the derived description. A valuable alternative is to use preprocessed annotated text as input to the speech synthesizer. Enriched text of that kind exhibits major advantage over plain text as it retains structural, syntactic, semantic and discourse level information in the form of tags in the markup. Each of the above types of linguistic information is described by sets of features that can be used to generate improved prosody in speech synthesis. Depending on the domain as well as the type of text different sets of features may be used for maximum improvement.

As an alternative to generated text, existing plain text can be adequately processed to derive annotated NLG-similar output, essentially gaining advantage for the prosody modeling stage in speech synthesis. In order to do that efficiently, automated analysis and annotation should be made available, for the most language analysis stages as possible. A breakdown of the identifiable distinct processes is:

- Word/Sentence identification.
- Shallow syntactic analysis (part-of-speech tagging and noun-phrase identification)
- Insertion/annotation of prosodic features

As described in the following paragraphs, fully automated analysis can be achieved for all but the latter process. Linguistic phenomena such as anaphoric references, rhetorical relations and others are particularly hard to automatically derive from plain text alone. This has to be done manually allowing for flexibility of the feature sets that may be used in the description as well as fully supporting the markup in terms of description, editing and export.

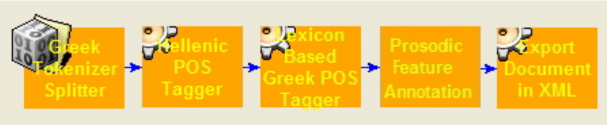


Figure1: The annotation workflow

For the purpose of this work a specific module for export to XML-like markup has been constructed. The procedure is fully automatic and easily upgradeable to support all existing tags as well as any future additions to the tagsets. The format is an adaptation-extension of the SOLE-ML description [6] that has been successfully used previously [7] as input to the *DEMOSTHeNES* speech composer system [8]. All the above processes have been implemented through the utilization of the Ellogon platform and its incorporated natural language analysis and annotation components.

3. The "Ellogon" Text Engineering Platform

Ellogon is a multi-lingual, cross-platform, general-purpose text engineering environment, developed in order to aid both researchers in the natural language field as well as companies that produce and deliver language engineering systems. Ellogon consists of mainly three subsystems:

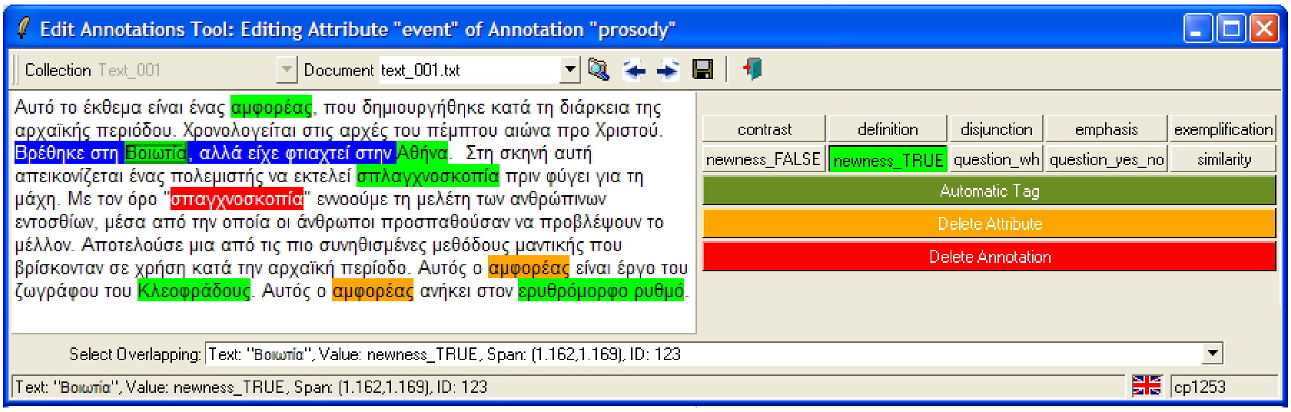
- A highly efficient core developed in C++, which implements an extended version of the TIPSTER data model. Its main responsibility is to manage the storage of the textual data and the associated linguistic information and to provide a well-defined programming interface that can be used in order to retrieve/modify the stored information.
- A powerful and easy to use graphical user interface (GUI). This interface can be easily tailored to the needs of the end user.
- A modular pluggable component system. All linguistic processing within the platform is performed with the help of external, loaded at run-time, components. These components can be implemented in a wide range of programming languages, including C, C++, Java, Tcl, Perl and Python.

Ellogon as a text engineering platform offers an extensive set of facilities, including tools for visualising textual/HTML/XML data and associated linguistic information, support for lexical resources (like creating and embedding lexicons), tools for creating annotated corpora, accessing databases, comparing annotated data, or transforming linguistic information into vectors for use with various machine learning algorithms. Additionally, Ellogon offers some unique features, like the ability to freely modify annotated textual data (with Ellogon automatically applying the required transformations on the associated linguistic information) and the ability to create stand-alone applications with customised user interfaces that perform specific tasks.

A large number of the functionalities provided by Ellogon have been exploited in the context of the work presented in this paper. In order to annotate corpora with prosodic information the annotation facilities provided by Ellogon have been extensively used. Ellogon provides a wide range of corpora annotation tools for annotating plain textual and HTML corpora as well as tools for annotating hierarchical related information (i.e. syntax trees). The tool for annotating textual/HTML corpora has a simple and easy to use interface: The textual/HTML rendering is presented to the user, along with a set of buttons, each of which is associated with a specific category. The user can select portions of the rendered text and classify it into one or more of the available categories. Additional facilities are provided for correcting mistakes or by automatically annotating all occurrences of specific text with the same category within a document. Furthermore, the extensive support provided by Ellogon for embedding lexical resources (like morphological lexicons) has enabled the easy construction of an accurate lexicon-based part of speech tagger by combining two independent morphological lexicons for the Greek language [9]. Finally, functionalities like XML, DOM and XSLT support as well as the various viewers created a "comfortable" environment for corpora annotators and the export of annotated information in the desired XML format.

4. Syntactic info

Before the corpus gets annotated with prosodic information, several pre-processing steps must be applied. Pre-processing mainly includes word and sentence identification, as well as part-of-speech (POS) tagging. Ellogon is equipped with ready-to-use components that can handle all these steps for the Greek language: word and sentence identification are performed by a rule-based component (HTokenizer) that presents an accuracy that approaches 100%, while a component based on machine learning (HBrill) has been employed for POS tagging, presenting an accuracy that approaches 75% (average measurement for several domains



Collection Text_001 Document text_001.txt

Αυτό το έκθεμα είναι ένας αμφορέας, που δημιουργήθηκε κατά τη διάρκεια της αρχαϊκής περιόδου. Χρονολογείται στις αρχές του πέμπτου αιώνα προ Χριστού. Βρέθηκε στη Βοιωτία, αλλά είχε φτιαχτεί στην Αθήνα. Στη σκηνή αυτή απεικονίζεται ένας πολεμιστής να εκτελεί σπλαγχνοσκοπία πριν φύγει για τη μάχη. Με τον όρο "σπαγχνοσκοπία" εννοούμε τη μελέτη των ανθρώπινων εντοσθίων, μέσα από την οποία οι άνθρωποι προσπαθούσαν να προβλέψουν το μέλλον. Αποτελούσε μια από τις πιο συνηθισμένες μεθόδους μαντικής που βρίσκονταν σε χρήση κατά την αρχαϊκή περίοδο. Αυτός ο αμφορέας είναι έργο του ζωγράφου του Κλεοφράδους. Αυτός ο αμφορέας ανήκει στον ερυθρόμορφο ρυθμό.

Select Overlapping: Text: "Βοιωτία", Value: newness_TRUE, Span: (1.162,1.169), ID: 123

Text: "Βοιωτία", Value: newness_TRUE, Span: (1.162,1.169), ID: 123

Figure 2: Prosodic feature annotation

that is tested for). This relative low accuracy of the built-in POS tagger can be easily justified, as the tagger is based on Transformation-based Error-driven learning [10] in order to classify each word of a document into a category characterized by the part of speech, gender and number of the word. As the tagger has been trained on a relatively small corpus (about 16,000 words) from a single domain (related to announcements about "management succession events") [11], it is expected to present a degradation in its accuracy when used in different domains than the one it was trained for.

In order to increase the accuracy of POS tagging, we decided to combine Ellogon's built-in component with a lexicon-based POS tagger for the Greek language. Two morphological lexicons for the Greek language have been combined in order to build a lexicon-based POS tagger with the highest possible coverage. The first lexicon is a large-scale morphological lexicon for the Greek language, developed by the Software and Knowledge Engineering Laboratory (SKEL) of NCSR "Demokritos" [9]. The lexicon consists of ~60,000 lemmas that correspond to ~710,000 different word forms. The second lexicon is property of the Speech Group, University of Athens and contains ~60,000 lemmas, which correspond to ~650,000 word forms. Both lexicons yield a word form identification span of ~880,000.

Due to the introduction of the lexicon-based tagger, the POS tagging preprocessing step can be separated into two independent sub-steps: The built-in POS tagger as well as the lexicon-based POS tagger are both applied independently. If a word is contained in any of the two lexicons and thus is assigned a POS category by the lexicon-based tagger, this categorization becomes the final POS of the word, ignoring any categorization performed by the built-in POS tagger. On the other hand, if a word is not found in any of the two lexicons, the categorization performed by the built-in POS tagger is used.

The tagset employed by both POS taggers contains information about the POS category of a word (ARTICLE, ADJECTIVE, NOUN, VERB, CONJUNCTION, PARTICIPLE, ADVERB, PARTICLE, PREPOSITION, PRONOUN) and not additional information like gender or number. Ambiguity in the lexicon-based tagger (i.e. word forms found in more than one POS categories) is resolved by selecting the category that comes first in the tagset described above. For example, if a word form is found to be an adjective and a participle, it will be classified as adjective.

5. Prosodic tagging

Part-of-speech and phrase type information alone cannot always infer certain intonational focus points since those are not only affected by syntax but also by semantics and pragmatic factors [12]. These factors are accounted for by enriching the text corpus accordingly. For the manual annotation of prosodic features the extensive editing tool of the platform is used. That is a fully flexible system that can use subsets of selected features as well as new sets that may be edited at any time. The current list of prosodic features currently used is given below:

contrast
definition
disjunction
emphasis (explicit)
exemplification
newness_TRUE
newness_FALSE
similarity
question_yes_no
question_wh

Prior experience [4] shows that there may be more than one feature associated with each word or set of words (or phrase) needed for successful description used for prosody modeling. The editing module fully supports unlimited overlapping of annotations for each token as well as nesting.

Annotations are shown in different color as illustrated in Figure 2. Overlapping/nesting can be viewed (details showing at the bottom of Figure 2) and edited at any time.

6. Export to XML

XML markup is a well tested means of representing enriched text. The SOLEML description was built as an annotation scheme for CtS synthesis [6], used as markup for the enriched text output of the ILEX generator [1]. It has been used with great success in previous works and is now a standard input of the DEMOSTHeNES speech composer. A special module for automatic extraction to an extended XML description based on SOLEML from Ellogon has been constructed and successfully integrated to the workflow. Figure 3 shows the XML output for a part of the text shown in Figure 2.

```
<utterance>
<relation name="Word" structure-type="list">
<wordlist>
<w id="w23">Βρέθηκε</w>
<w id="w24">στη</w>
<w id="w25" punct=",">Βοιωτία</w>
<w id="w26">αλλά</w>
<w id="w27">είχε</w>
<w id="w28">φτιαχτεί</w>
<w id="w29">στην</w>
<w id="w30" punct=".">Αθήνα</w>
</wordlist>
</relation>
<relation name="Grouping" structure-type="list">
<elem punct-type="double-quote" href="words.xml#id(w49)"/>
</relation>
<relation name="Syntax" structure-type="tree">
<elem phrase-type="S">
<elem phrase-type="prosody" contrast>
<elem lex-cat="VERB" href="words.xml#id(w23)">Βρέθηκε</elem>
<elem lex-cat="ARTICLE" href="words.xml#id(w24)">στη</elem>
<elem phrase-type="prosody" newness="true">
<elem lex-cat="NOUN" href="words.xml#id(w25)">Βοιωτία</elem>
</elem>
<elem lex-cat="CONJ" href="words.xml#id(w26)">αλλά</elem>
<elem lex-cat="VERB" href="words.xml#id(w27)">είχε</elem>
<elem lex-cat="VERB" href="words.xml#id(w28)">φτιαχτεί</elem>
<elem lex-cat="ARTICLE" href="words.xml#id(w29)">στην</elem>
<elem phrase-type="prosody" newness="true">
<elem lex-cat="NOUN" href="words.xml#id(w30)">Αθήνα</elem>
</elem>
</elem>
</elem>
</relation> </utterance>
```

Figure 3: The SOLEML export description

A wordlist of all tokens (words) and punctuation values takes up the first part, followed by the syntax and high-level information. The SOLEML element is the final part of the process and is automatically updated when the tagsets from the earlier steps are modified.

7. Conclusions

The procedure of creating enriched linguistic environment for high-level speech synthesis has been presented as a workflow of combined linguistic resources and text engineering workbed. From an initial plain text corpus an enriched text XML markup description is derived, ready to be used by the speech synthesizer. Utilizing and enhancing the Ellogon platform modules, a combination of automatic text analysis and manual prosody annotation as well as an implementation of the XML export module produces fully descriptive annotated text input. The generated XML description is compatible with the one from the ILEX generator (tested for a subset of ILEX-produced text) with the added flexibility of use for other limited domain or general purpose plain text. Future work involves the addition of phrase information by using the Ellogon noun phrase chunker for the Greek language, the only missing element from the current analysis when compared to the text produced from ILEX.

8. Acknowledgements

The work described in this paper has been partially supported by the HERACLITUS project of the Operational Programme for Education and Initial Vocational Training (EPEAEK) of the Greek Ministry of Education under the 3rd European Community Support Framework for Greece.

9. References

[1] O'Donnel, M., Mellish, C., Oberlander, J., and Knott, A., "ILEX: An architecture for a dynamic hypertext generation system", *Natural Language Engineering*, vol.7, no.3, pp. 225-250, 2001

[2] Hitzeman, J., Black, A., Taylor, P., Mellish, C., and Oberlander, J. "On the Use of Automatically Generated Discourse-Level Information in a Concept-to-Speech Synthesis System", *Proc. 5th International Conference on Spoken Language Generation (ICSLP)*: 2763-2768, 1998

[3] Pan, S., McKeown, K., and Hirschberg, J., "Exploring features from natural language generation for prosody modeling" *Computer Speech and Language*, 16:457-490, 2002

[4] Xydas, G., Spiliotopoulos, D., and Kouroupetroglou, G., "Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora". *IEICE Trans. of Inf. and Syst.*, Special Section on "Corpus-Based Speech Technologies", vol. E88-D, no 3, March 2005, pp. 510-518.

[5] Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos, I., and Spyropoulos, C.D., "Ellogon: A New Text Engineering Platform". *Proc. 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 72-78, Las Palmas, Canary Islands, Spain, May 2002.

[6] Hitzeman, J., Black, A., Mellish, C., Oberlander, J., Poesio, M., and Taylor, P., "An annotation scheme for Concept-to-Speech synthesis", *Proc. 7th European Workshop on Natural Language Generation*, Toulouse France, pp. 59-66, 1999.

[7] Xydas, G., Spiliotopoulos, D., and Kouroupetroglou, G., "Modeling Prosodic Structures in Linguistically Enriched Environments". *Lecture Notes in Artificial Intelligence* 3206, Springer-Verlag Berlin Heidelberg, pp. 521-528, 2004, ISBN 3-540-23049-1

[8] Xydas, G. and Kouroupetroglou, G., "The DEMOSTHeNES Speech Composer", *Proc. 4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland, pp.167-172, 2001.

[9] Petasis, G., Karkaletsis, V., Farmakiotou, D., Androutsopoulos, I., and Spyropoulos, C.D., "A Greek Morphological Lexicon and its Exploitation by Natural Language Processing Applications". *Lecture Notes on Computer Science*, vol.2563, Springer Verlag, 2003.

[10] Brill, E., "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging". *Computational Linguistics*, 21, 1995.

[11] Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D., and Androutsopoulos, I., "Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques". *In Fakotakis, N. et al. (Eds.), Machine Learning in Human Language Technology*, pp. 29-34, 1999.

[12] Bolinger, D., Intonation and its Uses: Melody in grammar and discourse, Edward Arnold, London, 1989