Semi-Automatic Prosodic Transcription of Spoken Spanish in XML

Eduardo Velázquez

Ph.D. Student, Freie Universität Berlin
CONACYT Scholar (Consejo Nacional de Ciencia y Tecnología, Mexico)
utka@yahoo.com

Abstract

XML (Extensible Mark-up Language) is designed to represent hierarchical structures; in this case, it shows the structure of the prosodic components of spoken language. The XML-based transcription system proposed here allows the input of 1) the phonetic parameters of F_0, intensity and duration of each syllable, their relative variation and standard values to facilitate discrimination and comparison; 2) the distribution of feet; 3) the boundaries and characterization of intonation units and utterances, and 4) other conversational phenomena such as pauses, overlaps, interruptions, etc. This mark-up language is currently being used as an analysis tool for a corpus of digitally-recorded conversations in the Mexican and Iberian vernaculars of spoken Spanish.

1. Introduction

There are three characteristics which distinguish XML [3] from other markup languages (e.g. HTML):

a) Its *extensibility*: it does not contain a fixed set of tags.
b) Its emphasis on *descriptive* rather than procedural markup: descriptive markup allows the same document to be processed in a variety of ways by means of style sheets which use only the parts considered relevant, or work the same part of the document for different processes. XML focuses on the meaning of data, not its presentation.
c) Its *document type* concept: documents are considered to conform to different types, and every document type is formally defined by its constituent parts and their structure. Therefore, documents must be well-formed according to a defined syntax, and may be formally validated.
d) Its *independence* of any one hardware or software system: every XML document, regardless of the language and writing system employed, uses the same method to encode characters as binary data. This encoding is defined by an international standard known as Unicode, which provides a character set covering most of the past and present writing systems of the world [12].

In XML, constituents are called elements. Each element represents a document's logical component. A document must describe the logical role of its elements, the abstraction they represent. Elements may contain sub-elements and text, also called character data. Moreover, elements may be specified and described by means of attributes.

There is also another feature in XML which allows the administration of the size and complexity of documents: the external entity. Through the use of external entities, a document is able to track the pieces of bytes composing it [7].

Finally, markup is the medium used to represent the document's logical structure and the way all physical entities are linked. The general syntax rules of XML markup are given below:

- Markup differs from character data by using special characters called *delimiters*. Thus, a tag is everything between "<" and ">", or between "`&`" and "`;`". [7]
- Tag names are case sensitive. Therefore, `<TAG>`, `<Tag>` and `<TaG>` are interpreted as three different tags [7], [12].

1.1. Marking up human language

The idea of representing human language with a markup language is not new. Some specialist teams, consisting for the most part of computer scientists, are dedicated to making semantic web browsing possible (e.g. [2]), based on the conceptual classification of information, while others are working on speech synthesis (*Speech Synthesis Markup Language*, SSML, [4]) and voice recognition (*Voice Extensible Markup Language*, VoiceXML, [8]). Other applications are being developed and used by language scientists, i.e. EXMARaLDA [9], [10] and TEI [12], whose principles are related to some extent to this proposal. However, none of those systems were specifically developed to represent the prosodic structure of language.

2. Structure of prosodic transcripts

In this section, reference is made to the constituents of a transcript containing prosodic information, and the hierarchical structure in which those elements are embedded (see Fig. 1). The root element of this proposal is called `<Transcript>`. It has one optional element `<Header>` (see 2.1.) and a required element `<Text>` (see 2.2.). Their sub-elements and their respective attributes are explained below.

2.1. Header

All sub-elements of `<Header>`, except for `<Participants>`, are what are commonly known as *nodes* since they have no subordinate elements and their content is character data. Elements `<Class>` and `<Acoustic_quality>` differ from the rest of nodes because they have attributes. In Fig. 1, required attributes are displayed in regular typeface, and optional attributes in italics. Attribute *`Type`* of `<Acoustic_quality>` has three possible predefined values (`A`, `B` or `C`), while attribute *`Type1`* of `<Class>` has a character data value. Element `<Participants>` may have one or more elements `<Speaker>`, which also have a series of sub-elements, some of which have their own attributes. Element `<Header>` contains, therefore, the metadata corresponding to the sound file.

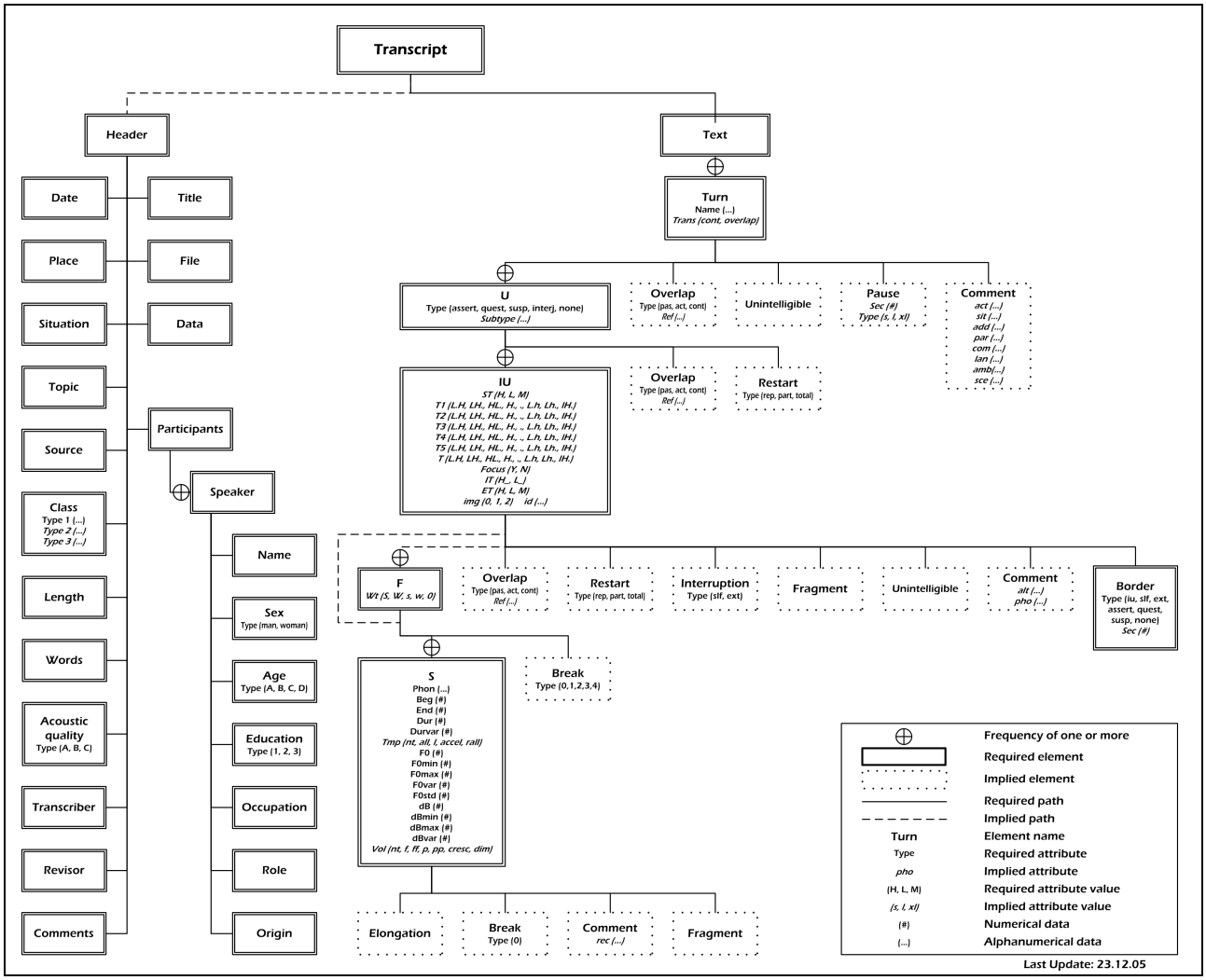

Figure 1: Hierarchical structure of prosodic information in transcripts.

2.2. Text

Compared with <Header>, where only metadata is stored, <Text> organizes all phenomena, events, actions, and spoken texts that constitute a conversation. These elements are explained by levels in the following subsections.

2.2.1. Turns

One or more of the element <Turn> are the only direct descendants of <Text>. Each <Turn> may have two attributes: *Name* and *Trans*. Attribute *Name* is required and admits any type of character as its value, which allows the introduction of the identity codes assigned to each participant in the conversation. *Trans* is optional, since transitions between turns, where a turn is the continuation of a former turn (cont) or turns produced simultaneously (overlap), are considered as exceptions and should be marked by means of those attributes.

The only required descendants of <Turn> are one or more utterances, <U>. All other sub-elements of <Turn> are optional. <Overlap> indicates the beginning and ending points of overlapped (pas) or overlapping (act) productions; cont is used where any of the simultaneous texts extend over more than one turn, utterance or intonation unit, as <Overlap> appears at different levels inside the structure. A reference number or name may be assigned by means of *Ref*. Element <Unintelligible> marks unintelligible texts in the transcript. Likewise, attribute *Type* of <Pause> may be specified through the values s, l and xl, while *Sec* allows the specification of duration in seconds. Finally, <Comment> has optional attributes which allow describing the different types of phenomena intervening during a conversation.

2.2.2. Utterances

Each utterance, <U>, requires one or more elements <IU>, i.e. intonation units, and its syntactical category is specified by *Type* and *Subtype*. Sub-elements <Overlap> and <Restart> may optionally appear at this level. The values of attribute *Type* of <Restart> mean repetition (rep), and partial (part) or total (total) restart.

2.2.3. Intonation units

<IU> may be specified by several attributes containing the tonal information from different positions inside the intonation unit: start tone (*ST*), end tone (*ET*), nuclear tone (*T*), and up to

five pre-nuclear tones (*T1* to *T5*). The values for these attributes, despite light modifications to avoid characters * or + so as not to interfere with XML syntax, correspond to the Sp-ToBI tone inventory [1], [11], e.g. L.H corresponds to L*+H, Lh. to L+!H*, and lH. to ¡L+H*.

There are also other attributes, like *Focus* (with positive –y– or negative –n– values) or intermediary tone, *IT*, which may be high or low. Attributes *img* and *id*, are used for the HTML rendition (see 4.).

In fact, all sub-elements of <IU> are optional, so that the structure also allows less detailed XML documents, i.e. without rhythm or syllable structures. New elements at this level are <Interruption>, <Fragment>, and <Border>. The latter takes advantage of the characters at the boundary between each unit, specifies them by means of the attribute *Type*, and even allows the input of duration with *Sec*.

2.2.4. Feet

There are two broken lines from <IU> to <S>, syllable; only one of which runs through <F>, foot. Avoiding <F> would be reasonable when the rhythm structure is not being analyzed.

When including the structure of metrical feet, attribute *Wt* (weight) could be used: s (strong), w (weak) or 0 for free feet. Upper case is used for strong feet and lower case for weak feet.

2.2.5. Syllables

Between every syllable, <S>, there is a <Break>, whose values also correspond to those of the Sp-ToBI inventory.
Each <S> has a very important series of attributes: phonetic value (*Phon*); beginning, end, duration, relative variation, and tempo of the syllable (*Beg*, *End*, *Dur*, *Durvar*, *Tmp*); its fundamental frequency, with minimum, maximum, relative variation, and standardized value [5] (*F0*, *F0min*, *F0max*, *F0var*, *F0std*), as well as its intensity, with minimum, maximum, relative variation, standardized value, and relative volume (*dB*, *dBmin*, *dBmax*, *dBvar*, *dBstd*, *Vol*).

Sub-elements of <S> are: <Elongation>, which marks the positions where a segmental elongation occurs; <Break> with value 0, in order to signal the morphological limits of two syllables in a liaison; <Comment> with attribute *rec*, as a way of orthographically reconstructing unpronounced segments, and, pointing out a fragmented production with <Fragment>.

2.3. Document Type Definition (DTD)

This model is then translated into a document type definition (DTD), which acts as the validating grammar of all documents that declare themselves pertaining to it.

Examples of this step (and other steps explained below) may be found under the following hyperlink:

<http://www.geocities.com/utka/>

3. Collecting PRAAT data into XML

3.1. Recording of conversations and basic transcription

The digital recordings adapted to this transcription system belong to a corpus of spontaneous conversations between speakers from Madrid and Mexico City, which represent the standard Spanish and Mexican vernaculars respectively.

Basic transcripts of these conversations enable their management, according to strict transcription criteria. They also provide the first input into PRAAT text grids, which will then be phonetically transcribed. Other tiers could represent, e.g. the rhythmical structure of syllables and their standardized F_0 values [5].

3.2. PRAAT scripting

The reason why PRAAT is used in this process is not just its features for phonetic analysis, but also its facility for creating scripts to automate its own functions.
By using such scripts, images for each intonation unit or utterance are created. These images, showing pitch, spectrogram, and the content of all tiers, are used for the final HTML rendition. Scripts also allow the assignation of variables to the phonetic parameters of each syllable and the looping of this process for every syllable in an intonation unit. The values of the variable may be continuously appended to a text file following an XML-like syntax.

3.3. XML document

The text file yielded by such a PRAAT script will be like this:

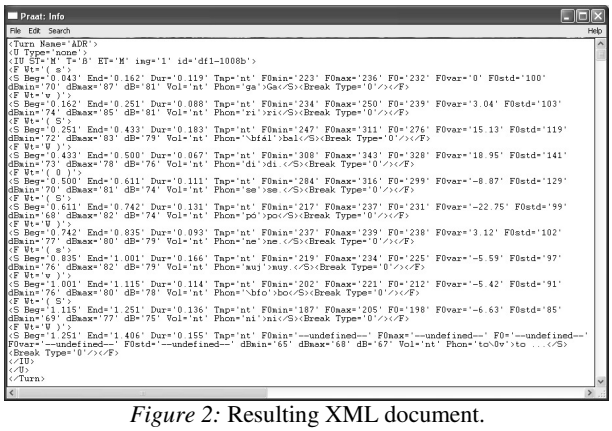

Figure 2: Resulting XML document.

Since XML documents are purely text files, there is no need of file conversions or adaptations. Possible repetitions, mistakes, or PRAAT character codes (incompatible with Unicode) may be replaced by means of a Visual Basic macro.

4. Rendering XML in HTML format

The most demanding part of the whole process of creating XML documents is the application of style data, since this requires the knowledge of several computer languages, tools and applications: XSLT, XPath, HTML, CSS, JavaScript, etc.

XML uses a set of mechanisms called XSLT (Extensible Stylesheet Language Transformations [6]). These style sheets do not just format the text to be rendered in HTML format, but also process the structure and information of elements and attributes in order to introduce, for example:

1) a table with the contents of the header, if it exists;
2) a separate table with the speakers' information;
3) a summary, where all prosodic phenomena encoded in the document are counted up, and
4) a list of conventions used throughout the text.

Moreover, by means of the style sheet, certain features may be added in order to insert a hyperlink at the beginning of each intonation unit pointing to an analysis window, where an

image showing, e.g., pitch, spectrogram and text tiers, and the corresponding sound file may be seen and listened to.

It is at this point that the `img` and `id` attributes of element `<IU>` are called upon, since `img` defines which button is displayed at the beginning of the intonation unit, and `id` provides the identity codes linking the linguistic production with the image and sound files. In the case that the button specifies that there is an analysis window linked to a particular intonation unit, the browser runs a JavaScript program, giving instructions to open another browser window displaying a web page containing the image and a link to the sound file.

Last but not least, it is necessary to point out that the ultimate aim of XML is not merely to render HTML pages, but to enrich them with a hierarchical structure and self-descriptive information. This is particularly useful in the case of `<S>`, whose attributes contain important phonetic data ready to use. Here lies the most important difference between XML and HTML: with XML all this information remains stored and may be recalled at any time.

Recalling the information may be done dynamically, for example, making each syllable in the text react when the cursor passes over them by turning red and displaying a yellow label with the most important phonetic information. If the syllable is clicked, a dialog window is opened displaying all values of the attributes of `<S>`. The final result on a web browser window appears as shown in Fig. 3. In order to test this demonstration in real time, refer to:

<http://www.geocities.com/utka/df1-04-1.html>

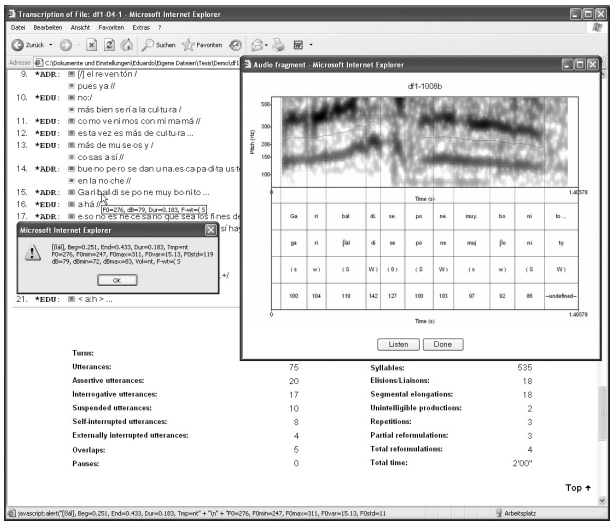

Figure 3: HTML rendition of an XML document.

5. Conclusions

In this paper, I have presented an XML-based transcription system designed to be applied as a tool to study several prosodic phenomena in spoken conversations pertaining to the Spanish vernaculars of Madrid and Mexico City.

This process begins with the basic transcription of recordings of interactive communications, where the relevant prosodic phenomena are minimally marked. Such information constitutes the most primary input source when segmenting the sound file in PRAAT. Other analyses are aggregated to the text grid as different tiers. The phonetic and textual information corresponding to each analyzed segment (turn, utterance, intonation unit or syllable) are then extracted and written into a text file with an XML-like syntax. Once the resulting XML document is free of errors, well-formed and valid according to its document type definition, it is processed and formatted by means of a style sheet, which yields an enriched and dynamic HTML document.

The resulting document then becomes something very different from the commonly-held conception of text or web page. Not only is its appearance important, but its treatment of data is rendered in such a way that it facilitates the analysis of spoken corpora through the coordination and combination of text, databases, sound files and images, offering a very powerful but easy-to-use platform.

6. References

[1] Beckman, M.E.; Díaz-Campos, M.; Tevis McGory, J.; Morgan, T.A., 2002. Intonation across Spanish in the Tones and Break Indices framework. *Probus* 14: 9-36.

[2] Berners-Lee, T.; Hendler, J.; Lassila, O., 2001. The Semantic Web. *Scientific American*: 17/05/2001 [http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21]

[3] Bray, T.; Paoli, J.; Sperberg-McQueen, C.M. ; Maler, E.; Yergeau, F. (ed.), 2004. Extensible Markup Language (XML) 1.0 Second Edition. *W3C Recommendation*: 04/02/2000 [http://www.w3.org/TR/REC-xml]

[4] Burnett, D.C.; Walker, M.R.; Hunt, A. (ed.), 2004. Speech Synthesis Markup Language (SSML) Version 1.0. *W3C Rec.*: 07/09/2004 [http://www.w3.org/TR/2004/REC-speech-synthesis-20040907/]

[5] Cantero, F.J., 2002. *Teoría y análisis de la entonación*. Barcelona: Universitat de Barcelona.

[6] Clark, J. (ed.), 1999. XSL Transformations (XSLT) Version 1.0. *W3C Rec.*: 16/11/1999 [http://www.w3.org/TR/1999/REC-xslt-19991116]

[7] Goldfarb, C.F.; Prescod, P., 2000. *The XML Handbook*, 2nd Ed. London *et al.*: Prentice Hall.

[8] McGlashan, S.; Burnett, D.C.; Carter, J.; Danielsen, P.; Ferrans, J.; Hunt, A.; Lucas, B.; Porter, B.; Rehor, K.; Tryphonas, S., 2004. Voice Extensible Markup Language (VoiceXML) Version 2.0. *W3C Rec.*: 16/03/2004. [http://www.w3.org/TR/2004/REC-voicexml20-20040316/]

[9] Schmidt, T., 2002. EXMARaLDA - ein System zur Diskurstranskription auf dem Computer. *Arbeiten zur Mehrsprachigkeit*, B (34), Hamburg. [http://www.rrz.uni-hamburg.de/exmaralda/Daten/4D-Literatur/AZM.pdf]

[10] Schmidt, T., 2005. Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. *Arbeiten zur Mehrsprachigkeit*, B (62), Hamburg. [http://www.rrz.uni-hamburg.de/exmaralda/Daten/4D-Literatur/SFB_AzM62.pdf]

[11] Sosa, J.M., 2003. La notación tonal del español en el modelo Sp-ToBI. In *Teorías de la entonación*, P. Prieto (ed.). Barcelona: Ariel, 185-208.

[12] Sperberg-McQueen, C.M.; Burnard, L. (ed.), 2004. TEI P5. Guidelines for Electronic Text Encoding and Interchange. *The TEI Consortium*. [http://www.tei-c.org/P5/Guidelines/]