Annotating Discontinuous Structures in XML: the Multiword Case

Emanuele Pianta and Luisa Bentivogli

ITC-irst
Via Sommarive 18, 38050 Povo (Trento) - Italy
{pianta,bentivo}@itc.it

Abstract

In this paper, we address the issue of how to annotate discontinuous elements in XML. We will take discontinuous multiwords as a case study to investigate different annotation possibilities, in the framework of the linguistic annotation of the MEANING Italian Corpus.

1. Introduction

The basic data structures of XML are trees. This makes XML very suitable for linguistic annotation, as trees are a very common formalism used for various linguistic representation levels. Syntactic trees are the most clear example of such linguistic representations. Trees can also be used to represent text divisions (sentences, paragraphs, sections, chapters), the structure of the content (e.g. the RST structure, see Mann & Thomson, 1987) or the graphical layout of the text (see Bateman et al., 2002).

Unfortunately there are at least two tasks that challenge the expressiveness of XML as a formalism for linguistic annotation based on trees, which by definition, unlike graphs, don't allow branches to overlap. The first problematic task is *including annotations for multiple or alternative linguistic representations* within the same XML document. The problem is that an XML document can contain only one tree structure, whereas different representation levels can require distinct, partially overlapping trees. For instance a content unit can include the first paragraph of a text and only half of the second one. This means that the branches that encompass the first content unit will cross the branches that encompass the first and second paragraphs. Another example is given by poetry. If we want to represent within the same XML document both the structure of the poem as a sequence of lines and the division in sentences we quickly run in problems, because a line can span over two parts of sentence. The problem is even more acute if we want to include alternative representations for the same linguistic level in the same XML document. In this case the probability that alternative representations lead to overlapping tree branches is even higher. Thus, if the various linguistic representation levels do not fit in one tree, it will be very difficult or impossible to keep different levels of linguistic annotation within the same XML document.

A second group of phenomena which may be difficult to represent through a tree-based formalism such as XML, are *discontinuous units* or *long distance dependences*. Examples of discontinuous units are non-contiguous multiwords, or -in German- separable verbs, whereas long distance dependences are exemplified by the coupling of a pronoun with its textual antecedent(s).

Apparently in this second group of phenomena we are dealing with only one representation level, so we would not expect to incur the problems caused by the necessity to include distinct representation levels. However, on closer inspection, it turns out that in all non trivial annotation tasks, more than one linguistic level is involved in the representation. Even if our aim is annotating a text at one level such as the syntax or the pronoun antecedents, in most cases we also need to represent within the same XML document at least one other linguistic level which is the division of the text in an ordered sequence of graphical words. Given the necessity to include in any linguistic annotation also information about the basic sequence of graphical words, the representation of any linguistic relation involving two non continuous graphical words is problematic for a tree-based formalism such as XML.

This happens even at the most basic linguistic levels, such as lexical annotation. The sequence of graphical words in a text can be represented with a flat tree in which each leaf corresponds to a word. However, if we want to represent in the same document the fact that two non-contiguous graphical words belong to the same lexical unit, then the necessity arises to use overlapping tree branches. Thus, also this second series of representation difficulties are explained by the expressive restrictions of XML as a tree-based formalism, that is a formalism which does not allow for a natural representation of multiple overlapping hierarchies.

2. Related work

The class of problems we are dealing with has been addressed in various ways in the literature on text annotation. One clarifying formulation of the problem describes it as the difficulty of annotating both the logical and the layout (o physical) structure of the same text. These two (tree-)structures may differ in various ways, which can be described in terms of node duplication, removal/addition, reordering, break-out (Murata, 1995).

Note that in principle the annotation problems we have presented so far could at least partly be solved by resorting to SGML, where the CONCUR feature allows for specifying multiple DTDs and associated tagging on a single document instance (Sperberg-McQueen & Huitfeldt, 1999). Unfortunately the CONCUR feature is not available in XML (Clark, 1997). Also, CONCUR is an optional feature of SGML and is not supported by all SGML processors (Sperberg-McQueen & Burnard, 2001). Finally, if a solution to the problem at stake can be found within the XML formalism, this should be preferred because of the expectation that XML documents are easier to be processed, and that more and more XML-aware tools are made available to the text annotation community in the near future.

Actually, a number of approaches have been proposed to allow for multiple overlapping annotations within XML. Sperberg-McQueen & Burnard (2001) provide a detailed description of the characteristics, the advantages and disadvantages of such approaches. Let us mention here what we think are the most important approaches:

- *Multiple encoding* of the same information. This is straightforward but redundant and bears the risk of introducing non-alignments between different but interrelated annotations, when one annotation is updated and the other not.
- Use of *milestone elements*, that is empty elements, marking the beginnings and endings of spans of text, for instance.: <start-span id='w1'/> … <end-span idref='w1'/>. This has the disadvantage that the structure of the "ghost" annotation based on milestones needs to be reconstructed with ad hoc procedures.
- *Stand-off annotation*, that is the annotation of a text is kept separate from the text itself; special pointers are used in the annotation to refer to the specific text elements which are the object of the annotation. This comes in two variants, as the annotation can be kept in a separate section of the same document, or in a separate file. See for instance the annotation format used in GATE (Cunningham et al., 2002).

Among these three approaches, in this paper we will prefer the stand-off approach, as we think that this guarantees the best compromise between advantages (elegance and clearness of the representation, conceptual simplicity of the processing) and disadvantages (physical discontinuity between the text and the annotation, complexity of the pointer processing).

As a final introductory remark, let us underline that none of these approaches to overlapping representations comes without a cost or some contraindication. As the TEI Guidelines put it, "non-nesting information poses fundamental problems for any encoding scheme, and it must be stated at the outset that no solution has yet been suggested which combines all the desirable attributes of formal simplicity, capacity to represent all occurring or imaginable kinds of structures, suitability for formal or mechanical validation, and clear identity with the notations needed for simpler cases" (Sperberg-McQueen & Burnard, 2001).

In the rest of this paper we will consider in more detail one of the cases of overlapping annotation, that have been mentioned above, that is the annotation of discontinuous multiwords. We will investigate different annotation possibilities and present the pros and cons of each of them. From such an analysis we will see that also the lexical representation level, apparently the simplest linguistic representation level, can be problematic, and requires principled solutions. More specifically we will see that lexical representation involves three more fine-grained levels, that are *tokens*, *potential words*, and *multiword expressions*.

3. The multiword case study

The term multiword is used to denotate various kinds of lexical units. Within this paper we will use it to refer to both idioms and restricted collocations. We will exemplify the problems that arise when annotating discontinuous multiwords by considering the Italian multiword "andarci piano" (Eng. "take it easy") within the following sentence:

IT: Coi superalcolici bisogna *andarci* veramente *piano*.
EN: People should *take it* really *easy* with liquors

As a first thing, this example shows that the level of graphical words can interact in interesting ways with other lexical analysis levels. On the one hand, graphical words can correspond to multiple lexicographic words, that is the kind of units that are listed as headwords in a dictionary. In the example above, the graphical word "coi" is the non-concatenating combination of a preposition (con, *with*), and an article (i, *the-plur*), whereas the graphical word "andarci" corresponds to a verb (andare, *to go*) and a clitic pronoun (ci, *there*). Composite words, such as *coi* and *andarci*, occur because two contiguous words can undergo phonological adjustment phenomena when they happen to occur one after the other in a text. Some of the adjustments are optional: for instance "coi" could be substituted by the original two words "con i", whereas the sequence of two words from which "andarci" is generated, that is "andare" and "ci", cannot occur without contraction in an Italian sentence.

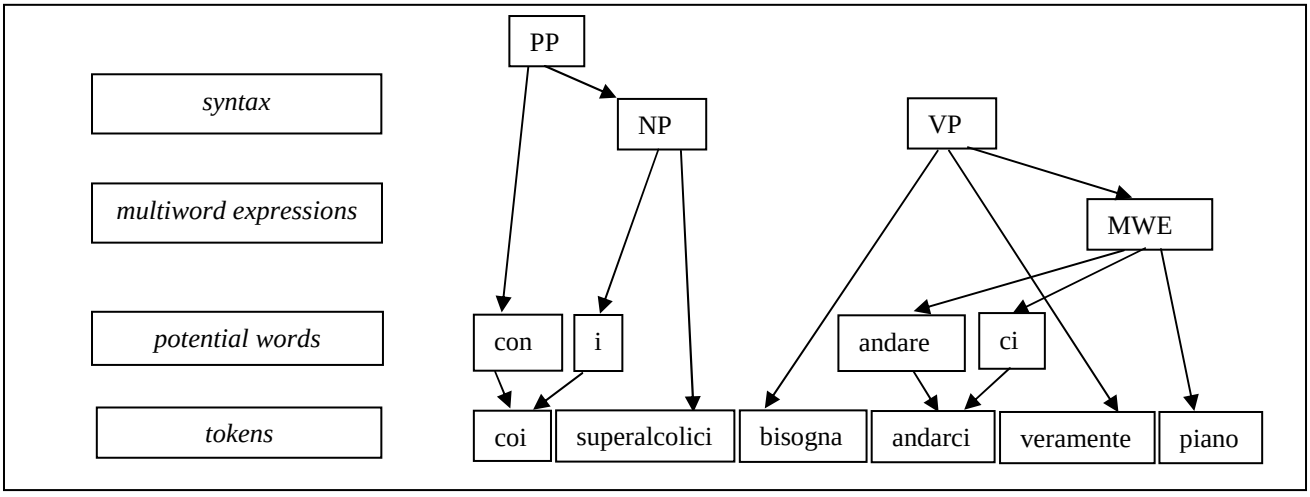

Figure 1. Interaction between lexical representation levels

On the other hand, the two graphical words "andarci" and "piano" correspond to one lexical unit with a unitary and non compositional meaning (*take it easy*). Also, note in the example that the two graphical words ("andarci", "piano") that compose the multiword, are non-contiguous, see Figure 1.

The example shows that we need to distinguish at least two other word-like units beyond graphical words. For sake of clarity, let us introduce the following notions and definitions:

a) **Token**: a graphical word (also called orthographical form), e.g.: *coi, superalcolici, bisogna, andarci, veramente, piano.*
b) **Potential word**: this notion was introduced by Pianta & Tovena (1999) to refer to an inflected word form before phonological and/or orthographic adjustment is applied to adjacent word forms, thus the notion makes sense from a generation point of view; note also that certain sequences of potential words may never occur in real texts because of obligatory adjustment rules. The potential words in our example are: *con, i, superalcolici, bisogna, andare, ci, veramente, piano.*
c) **Lexical unit**: one or more potential words carrying a unitary lexical meaning, e.g.: *con, i, superalcolici, bisogna, andare_ci_piano, veramente.*

The relations between these three levels can be complex. One token can correspond to more than one potential word, as in the examples below:

(Ita.) coi => con, i (preposition, article)
(Ita.) andarci => andare, ci (verb, clitic)
(Eng.) don't => do, not (verb, negation)
(Ger.) im => in dem (preposition, article)

On the other hand, more than one potential word can form a single lexical unit. This typically occurs with multiwords, as in "andare_ci_piano" and "take_it_easy".

As a further example consider the token "andarci" within the different sentences:

IT: Voglio *andarci* adesso
EN: I want to go there now

IT: Bisogna *andarci* piano
EN: People should take it easy

In the first sentence, the token "andarci" corresponds to two potential words (andare, ci) and to two lexical units (andare, ci). In the second sentence the token still corresponds to two potential words (andare, ci) but to only one lexical unit together with "piano".

The three levels illustrated above are conceptually distinct and, in principle, they correspond to three distinct levels of linguistic annotation.

Tokens are the basic representation level on which all the following ones are built. Generally speaking tokenization is not a trivial task. Many decisions need to be taken, and these decisions influence the analyses that are carried out at the following levels. To make some examples, tokenizing a text requires handling cases like the following ones:

- to distinguish a full stop that ends a sentence (separate token) from the full stop that ends an abbreviation (in-token character),
- to decide whether in "20%" the percentage sign is part of the preceding number or a separate token,
- to recognize that in "citta'" the "'" character is a representation of the accent on the "a" and not a quote surrounding the word, etc.

The representation of potential words and lexical units is also crucial for other representation levels. For instance, recognizing *potential words* is crucial for a correct syntactic annotation and also for word level alignment of parallel texts. If we do not distinguish the two potential words that compose the token "coi" we will not be able to annotate the prepositional phrase "coi superalcolici" with the correct syntactic structure: [*PP* con [*NP* i superalcolici]] (see Figure 1). We need potential words also to properly carry out word alignment of parallel texts:

andare [*align with*: go], ci [*align with*: there]

On the other hand, recognizing *lexical units* is crucial for lexical semantic annotation (e.g. with WordNet word senses), and for syntactic annotation as well.

In the current practice the first representation level is handled by the so-called *orthographic annotation*, which describes the actual tokens as they are found in the text. As for potential words and lexical units, they are usually represented together in one single annotation level, which is currently referred to as *morphosyntactic annotation*. Note that in this annotation approach, not only are the two levels based on an in-line annotation approach, but also no distinction is made between potential words and lexical units. This practice is due to the fact that in most cases lexical units and potential words coincide. When this is not the case, some problematic issues arise. Multiwords are the typical exception to the one-to-one correspondence between potential words and lexical units. The only proposal for representing multiword expressions that we could find in the literature is due to Ide & Romary (2002). However, this proposal has some limitations that we will examine in the next section.

4 Annotation of multiwords

The study on the annotation of multiwords that is presented in this section has been carried out in the framework of the MEANING project, more specifically in the context of the development of the Italian MEANING Corpus, a multi-level linguistically annotated corpus, having domain representativeness as main text selection criterion (Bentivogli et al., 2003). In designing the annotation scheme of the corpus we adhered as much as possible to the proposals for the new ISO/TC 37/SC 4 standard for linguistic resources (Ide and Romary, 2002), which are based on *annotation structures* (nestable <struct> elements) and *data categories* (<feat> tags). Different representation levels are contained in separate documents. Also, we use the XLink and XPointer syntax to represent relations between elements in different XML documents, and IDREFs attributes for relations within the same document.

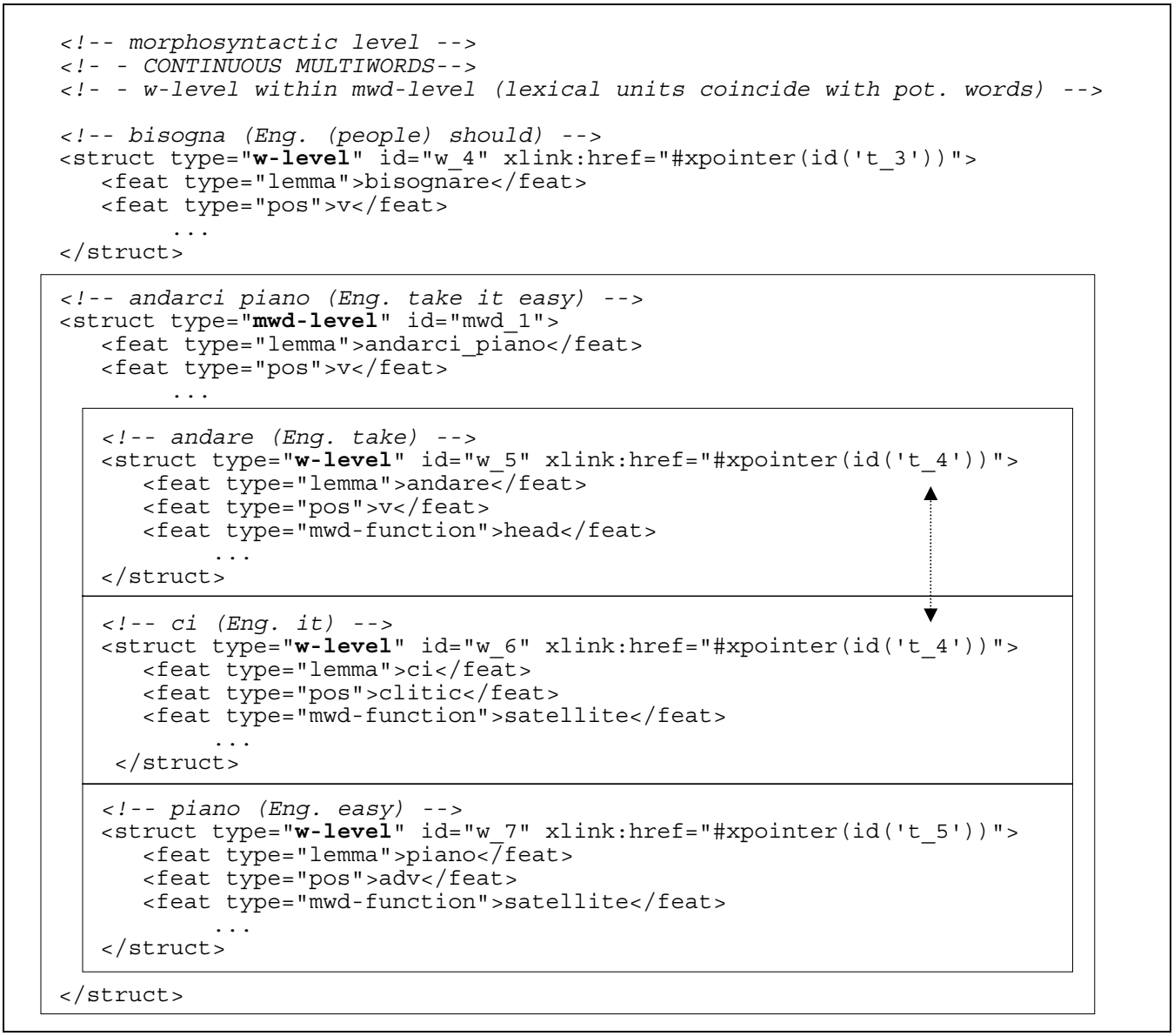

Figure 2. Annotation Scheme A, for continuous multiword expressions

In the actual annotation phase of the Italian MEANING Corpus, we faced the task of annotating multiwords, and we realized that the current annotation schemes available in the literature do not always allow to distinguish between potential words and lexical units, and do not provide satisfactory solutions for the annotation of discontinuous multiwords.

4.1 Continuous multiwords

If all the elements of each multiword were adjacent, we could still easily represent both the potential word and lexical unit levels through in-line annotation, following the proposal by Ide and Romary (2002). For instance, we can annotate the sentence "bisogna andarci piano" (Eng. "people should take it easy") as shown in Figure 2 above.

In Annotation Scheme A, simple lexical units (where potential words and lexical units coincide) are annotated with w-level structures, whereas complex lexical units are annotated in-line with mwd-level structures. Each mwd-level structure encompasses the w-level structures describing the single potential words which constitute the multiword. Note that in the example above, even if the annotation of multiword expressions is in-line with respect to potential words, the annotation of potential words at morphosyntactic level is stand-off with respect to the token level. This is in fact the only way to specify that the two potential words *andare* and *ci* correspond to the one token *andarci*.

This annotation scheme is slightly different from the original proposal by Ide and Romary, in which both simple and complex lexical units are annotated with w-level structures. We think that the w-level and the mwd-level are to be kept distinct, because certain pieces of information only pertain to the mwd-level. For instance, the lemma and the PoS of the multiword can only be annotated at the mwd-level. This is an important point that should be kept in mind to understand some of the proposals that will follow.

```
<!-- morphosyntactic level -->
<!-- DISCONTINUOUS MULTIWORDS -->

<!-- andare (Eng. take) -->
<struct type="w-level" id="w_5" xlink:href="#xpointer(id('t_4'))">
   <feat type="lemma">andare</feat>
        ...
   <feat type="mwd-element" IDREFS="w_6 w_8">head</feat>
</struct>

<!-- ci (Eng. it) -->
<struct type="w-level" id="w_6" xlink:href="#xpointer(id('t_4'))">
   <feat type="lemma">ci</feat>
        ...
   <feat type="mwd-element" IDREFS="w_5 w_8">satellite</feat>
</struct>

<!-- veramente (Eng. really) -->
<struct type="w-level" id="w_7" xlink:href="#xpointer(id('t_5'))">
   <feat type="lemma">veramente</feat>
        ...
</struct>

<!-- piano (Eng. easy) -->
<struct type="w-level" id="w_8" xlink:href="#xpointer(id('t_6'))">
   <feat type="lemma">piano</feat>
        ...
   <feat type="mwd-element" IDREFS="w_5 w_6">satellite</feat>
</struct>
```

Figure 3. Annotation scheme B: w-level structures with IDREFs

Also, we explicitly mark the head and the satellites of the multiword (see the feature mwd-function), assuming that at least some features of the head (for instance agreement features) are passed over to the all multiword. Note also that the two potential words "andare" and "ci" point to the same token "andarci" in the orthographic file through XLink and XPointer links.

4.2 Discontinuous multiwords

Unfortunately, multiwords can be discontinuous, as is shown in the sentence of our case study "Coi superalcolici bisogna *andarci* veramente *piano*" (Eng. "People should *take it* really *easy* with liquors"). The adverb "veramente" (really) can be inserted within the multiword, but is by no means part of the multiword. Annotation Scheme A seems not to be suitable to represent this case. More specifically there seems not to be any way to represent both the fact that "andare", "ci", and "piano" compose a single lexical unit, and the fact that the adverb "veramente" occurs between the potential words "ci" and "piano", but is a distinct lexical unit.

In the rest of this section we will illustrate two alternative solutions based on in-line annotation (Annotation Schemes B and C), and another solution which requires a stand-off annotation (Annotation Scheme D).

The first solution is given in Annotation Scheme B (see Figure 3 above). All potential words are represented by w-level structures, and we do not use an explicit mwd-level. However we represent the fact that a potential word is part of a multiword through the feature tags in the w-level structure. When a potential word is part of a multiword, its w-level structure contains a `<feat>` tag like the following:

```
<feat type="mwd-element"
      IDREFS="w_6 w_8"> head </feat>
```

The advantage of this solution is its structural simplicity. We don't need to introduce a new type of structure to represent multiwords: all we need are pointers inter-connecting the various parts of each multiword, and discontinuity is not an issue. The disadvantages of this solution are on one side the proliferation of pointers, on the other side the lack of a specific structure to represent information that pertains to the multiword as a unit and not to its components, e.g. the lemma and the PoS. The lack of a specific multiword level structure is a problem also for higher level linguistic annotations. For instance, at the syntax level we would like be able to refer to a multiword as a unit (see the pointer that links the VP node to the multiword verb in Figure 1). It is hard to see how this could be done within Annotation scheme B.

On the other hand Annotation Scheme C (Figure 4) resorts to the explicit representation of the mwd-level. However, the strategy here is the opposite of the one used in Annotation Scheme A: instead of representing simple structures within complex ones, i.e. w-level structures within mwd-level structures, we represent information about complex structures within the simple ones.

Figure 4. Annotation Scheme C: mwd-level within w-level structures

In Annotation Scheme C, we include the mwd-level structure, containing the information pertaining to the multiword, within the w-level structure representing the *head* of the multiword ("andare"). This mwd-level structure contains also the pointers to the possibly discontinuous satellites of the multiword (through the IDREF attribute). The w-level structures describing the *satellites* of the multiword include a mwd-level structure each, containing a pointer to the head of the multiword.

Also this annotation scheme has some drawbacks. First, it may be incorrect or at least inelegant to nest conceptually complex structures within simple ones. Second, the description of the function of each element of the multiword (head vs. satellites) has been put at the mwd-level, even if it logically pertains to the w-level. Finally, selecting information about multiwords is somehow awkward, as it is contained within simple words.

There is a further solution (Annotation Scheme D represented in Figure 5) which solves these drawbacks resorting to stand-off annotation.

In Annotation Scheme D, the potential word level and the multiword level are represented in two different sections. The first section represents potential words through w-level structures and their ordering in the text. Information about multiwords is easily accessible in the second section, where each mwd-structure contains the relevant multiword information and pointers to the multiword constituents in the first section. The status of a word as element of a multiword is marked explicitly in the potential word section, whereas the information pertaining to the multiword level can be retrieved starting from the first section, by following the ID-IDREF link backward with an XPATH expression. On the other hand the stand-off syntactic annotation can point to unitary multiword level structures in the multiword section of the annotation.

Figure 5. Annotation Scheme D: w-level and mwd-level structures in two files (or sections of file)

It is worthwhile to note that if the potential word level and the multiword level are to be represented in two different files instead of the same file, annotation scheme D can still be applied substituting the IDREFs with XLinks and XPointers.

In annotation scheme D, as well as in the previous ones, when simple lexical units coincide with potential words, they are represented with plain w-level structures.

We think that the stand-off approach illustrated by Annotation Scheme D can be considered the best compromise to represent discontinuous multiwords, in terms of structural clarity, expressive power and conciseness, so this solution will be applied to the annotation of multiwords in the Meaning Italian Corpus.

5 Conclusions

In this paper we analyzed the problem of linguistically annotating discontinuous elements with XML-based annotation schemes. The difficulty of this task seems to have the same grounds as the difficulty to include multiple or alternative linguistic annotations in the same XML document, that is the fact that an XML document cannot represent multiple branch-crossing trees.

Whereas stand-off annotation is the standard solution proposed to solve the multiple (alternative) annotation problem, less attention has been paid in the literature to the issue of representing discontinuous elements. We analyzed this issue by taking as case study the representation of discontinuous multiwords.

To this extent, first we pointed out the opportunity of conceptually distinguishing between tokens (graphical words), potential words (words before phonological adjustment) and lexical units (lexical semantic units), by showing that the objects of these three levels do not always correspond in a one-to-one way. Second, we showed that annotation schemes available in the literature do not allow to represent discontinuous multiwords. Finally, we proposed four different annotation schemes in XML for representing the three linguistic levels introduced above, by taking into account the most recent proposals for linguistic annotation standards, and by making explicit the distinction between potential words and lexical units whenever they do not correspond in a one-to-one way. Three of the proposed annotation schemes allow to represent discontinuous multiwords. However we got to the conclusion that stand-off annotation is the most suitable approach to represent discontinuous multiwords, and, more generally, to represent the complex relationships that hold between tokens, potential words, and multiwords.

References

Bateman, J., Henschel, R. & Delin, J. (2002). A brief introduction to GeM annotation schema for complex document layout. In Proceedings of the 2nd Wordshop on NLP and XML (pp. 13--20) Taipei, Taiwan.

Bentivogli, L., Girardi, G. & Pianta, E. (2003). The MEANING Italian Corpus. In Proceedings of Corpus Linguistics 2003(pp. 103--112) Lancaster, UK.

Clark, J. (1997). Comparison of SGML and XML. World Wide Web Consortium Note 15 December 1997. http://www.w3.org/TR/NOTE-sgml-xml.html

Cunningham, H., Maynard, D., Bontcheva, K. & Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, USA.

Harold, H. R. (2001). *XML Bible (second edition)*.

Ide, N. & Romary, L. (2002). Standards for Language Resources. In Proceedings of LREC 2002 (pp. 59--65) Las Palmas, Canary Islands, Spain.

Mann W. C. & Thomson S. A. (1987). Rhetorical Structure Theory: a Theory for Text Organisation. Technical Report RS-87.190, USC/Information Science Institute.

Murata, M. (1995). File format for documents containing both logical structures and layout structures. In Electronic Publishing, 8(4), 295--317.

Pianta, E. & Tovena, L. M. (1999). Mixing representation levels: The hybrid approach to automatic text generation. In Proceedings of the AISB'99 Workshop on Reference Architectures and Data Standards for NLP (pp. 8--13) Edinburgh, UK.

Sperberg-McQueen, C. M. & Burnard, L. (Eds) (2001). TEI P4: Guidelines for Electronic Text Encoding and Interchange: XML-compatible edition. The TEI Consortium. http://www.tei-c.org/P4X/.

Sperberg-McQueen C. M. & Huitfeldt, C. (1999). Concurrent Document Hierarchies in MECS and SGML. In Literary and Linguistic Computing 14, 29--42.

XCES: Corpus Encoding Standard for XML. http://www.cs.vassar.edu/XCES/