A Tag For Punctuation

Alexei Lavrentiev

TEI MM 2008
MS SIG working paper panel

Alexei Lavrentiev
Université de Lyon, ENS LSH, ICAR

A Tag for Punctuation?

In this paper I will argue that it may be useful to introduce a special tag (*e.g.* **<punct>**) for punctuation marks, which could join the TEI Analysis module along with **<c>**, **<w>**, **<phr>** and other "segLike" elements. I will first discuss the reasons why punctuation marks may need tagging, and then consider the TEI tags that might be used for that purpose. None of them appears to be perfect for this job. After discussing linguistic properties of punctuation marks, I will propose a tentative formal definition for the **<punct>** element.

Tagging punctuation marks may be useful, if not necessary; for several reasons. The first reason is that these marks are often relevant for automatic syntactic and semantic analysis of a text. In fact, punctuation marks are a kind of "natural tags" that make the text structure explicit. In modern European languages, a dot usually indicates a sentence boundary, a comma is generally used to separate some syntactic units inside a sentence. Many "chunkers" and other automatic language processing tools actually rely on punctuation marks for primary segmentation. However, like any other natural language object, punctuation marks are subject to synonymy and homonymy. A dot used as an abbreviation mark does not necessarily mark a sentence boundary (although it can fulfil both functions if the abbreviation is situated at the end of a sentence). The marks like colon may or may not signify a sentence boundary depending on semantic and structural reasons that can hardly be formalized. In some cases Unicode provides different codes for functionally distinct homonymous punctuation marks (*cf.* 2212 "minus sign" *vs.* 2010 "word-breaking hyphen" *vs.* 00AD "soft hyphen", etc.) but this only works with very few particular glyphs.
One could argue that the correct way to mark up syntactic units in a text is to tag those units: using either "seg-like" surrounding tags or "milestone-like" empty elements. However, this kind of tagging is likely to be performed by automatic language processing tools, and their efficacy can be considerably ameliorated by pre-tagging punctuation marks, especially those that are not used in their typical role (like dots in abbreviations). For example, we could manually tag as "weak" punctuation all exclamation and interrogative marks that are not situated at sentence boundaries, and then use a script to tag as "strong" punctuation all the other occurrences of these marks. A syntactic parser will be able to take these data into account.

The second reason for tagging punctuation marks is more specific and related to the edition of texts where the original punctuation diverges considerably from the modern rules. In many editorial traditions, the punctuation of source texts is modernized tacitly, whereas the original spelling is more frequently respected. The TEI does offer a generic mechanism for encoding regularizations (using **<orig>**, **<reg>** and **<choice>** elements), but in the case of punctuation marks this solution may appear to be unpractical. On the one hand, the heavy mechanism using **<choice>** does not seem "cost-effective" for these short (typically one-character) marks with a limited list of different forms. On the other hand it may be desirable to treat differently the regularization of word spelling and that of punctuation marks. A reader may want to view a text with original spelling and regularized punctuation or vice versa.

The first two reasons for tagging punctuation marks we mentioned are purely practical, but these marks can also be an object of a research interest and thus need to be annotated using appropriate

tools. Punctuation is a part of the writing system which does not have a direct correspondence with segmental units in the oral speech and is therefore inherently different from alphabetic characters (or letters). In some cases punctuation marks can be considered to be a part of a word form (dots in abbreviations or dots surrounding numbers in Medieval manuscripts) but as a rule they are situated at the same level as whole words in the hierarchy of segmental text structure. Unlike letters, punctuation marks can have their own functions and/or semantics. Sometimes punctuation marks can be paraphrased by words or morphemes but very often their function can hardly be expressed verbally. Studying punctuation in old texts, where no strict rules of usage were imposed, may produce interesting results on the way a writing person (a scribe, for example) perceived and modelled syntactic structures.

Any of the reasons stated above may justify the tagging of punctuation marks in a particular project. If the project is concerned with being TEI-conformant, a number elements seem to be available for that purpose.

A small section 3.2 of TEI P5 Guidelines is dedicated to the treatment of punctuation. The primary objective of this section is to "solve problems" the punctuation marks may cause. Punctuation is considered here to be a technical problem of minor importance. The solutions suggested consist in using appropriate Unicode characters where available, explicitly tagging sentences, quotations, abbreviations and other segments with "ambiguous" punctuation, or in using the **<c>** element.
As I have already mentioned, the first solution is only applicable to a small number of cases, and the second one implies manual text segmentation, a process which could be automated if the ambiguous punctuation marks were pre-tagged.

The **<c>** element is an obvious solution for tagging punctuation. The only example provided in the definition of this element is **<c type="punctuation">**?**</c>**, so it looks like *the* element TEI recommends for punctuation marks. As one of the practical reasons for tagging punctuation consists in distinguishing strong (sentence boundary) and weak punctuation, an attribute (**type** or **function**) can be used to specify this parameter. The **<c>** element is actually introduced by the Guidelines as a specialized form of **<seg>** that can only contain CDATA or a **<g>**, and is allowed inside larger segmentation elements like **<s>**, **<w>**, **<m>**, etc.
However, there is a substantial linguistic difference between characters like letters or diacritics and punctuation marks. The former are distinctive units used to construct meaningful units like morphemes or words. The latter are functionally independent units acting at the level of syntactic units. A word can consist of a single letter (like *I* in English), which does not mean that we can use **<c>** instead of **<w>** to mark it up. Some punctuation marks can be decomposed into several distinctive units (*e.g.* a triple interrogative mark *???*)[1], other marks, like quotes or parentheses, are composed of a pair of distant elements. Theses marks, not only designate a boundary of a text segment but also have a certain directionality, as they signify either the beginning or the end of a syntactic unit.
Using the same tag for elementary constructive units like letters and for functionally independent (and potentially decomposable) units like punctuation marks does not seem therefore linguistically correct. Furthermore, some "punctuation-like" characters (for example, apostrophes and hyphens) are in reality parts of word forms and should be clearly distinguished from ordinary (or "syntactic") punctuation marks.

It should be noted that language processing software often considers punctuation marks to be separate tokens (*i.e.* units of the same level as word forms). As word forms, they can carry their own morphological annotation. Some projects even use **<w>** element for punctuation for the sake of processing commodity.

1 An ellipsis mark (...) is often considered to be a single character.

It is of course always possible to use a generic **<seg>** element and its attributes **type**, **subtype** and **function** to mark up the punctuation marks, but if specialized tags exist for sentences, phrases, words, morphemes and characters, why not allow one for punctuation marks? Having a special tag for punctuation will make it possible to define a specific content and attribute value models and thus have a better control over the data.

Are there other alternatives for tagging punctuation marks within the existing TEI tagset?

As punctuation marks often act as "natural" text-chunkers, it may be possible to consider using an empty **<milestone>** element with a unit attribute to specify what chunk a given mark is used to delimit. This element is however not intended for tagging segmental text units and only has a limited set of attributes, which may cause problems for annotation.
The **<g>** element was mentioned during the discussion on the TEI-MS-SIG list as an option for encoding "non-standard" (*i.e.* non-Unicode) punctuation marks that can be found in Medieval manuscripts (see Parkes 1992 and Haugen 2006 for some examples). This element should actually be used to deal with particular glyphs of ancient and medieval punctuation marks but it is not intended for use in linguistic segmentation and annotation. The **<punct>** element should allow **<g>** as its content (as all the linguistic segmentation elements do).

Some projects like Menota (http://www.menota.org) and BFM-Manuscripts (http://bfm.ens-lsh.fr) are already using a **<punct>** element defined in their own namespace.

If the TEI council accepted the utility of introducing a special tag for punctuation marks, this element could join the TEI Analysis module and the **segLike** model class. In addition to global and **segLike** class attributes, it should have a **force** attribute (to distinguish "strong", "weak" and, possibly, "medium" punctuation marks), a **unit** attribute (to specify the type of syntactic unit a given mark is delimiting), a **direction** attribute (to specify, if applicable, whether the mark is used to open or to close a remarkable text segment) and, possibly, a **status** attribute (as a simple mechanism for dealing with the regularization of punctuation)[2]. Standard linking mechanisms can be used to connect the opening and closing marks working together.
As for its content, it can either be restricted to **macro.xtext** model (like that of the the **<c>** element) or allow a larger number of elements (including **model.pPart.edit** and **<c>**). The second option seems to be more adequate, but the problem of content models is probably worth discussing for all of the **segLike** elements. Such a discussion would however go far beyond the scope of this paper, and I will only mention an example of definitions that need reconsideration.
One of the aims of creating specialized linguistic segmentation elements was to define specific content models for each of them, but as they all belong to as segLike model, it is allowed to have **<s>** (sentences) inside an **<m>** (morpheme), which is absurd. At the same time, the current content model does not allow elements like **<am>** (abbreviation mark) and **<ex>** (supplied letters in the expansion of an abbreviation) inside a **<w>**, although these elements are word-internal by definition. The latter problem will soon be corrected by adding **model.pPart.edit** to the content of **<w>** but further reflection on which elements can occur inside a word and which cannot seems necessary.

An alternative solution to creating a new element could consist in redefining the <c> element, which seems actually be rarely (if ever) used in practice for anything else than tagging punctuation. I think that this would be wrong, as both characters and punctuation marks are real linguistic objects having different properties. Even though tagging individual characters may only be necessary in a

2 A standard mechanism with **<orig>**, **<reg>** and **<choice>** should nevertheless be preferred.

limited number of projects with a special interest in the analysis of graphic systems, it would be a shame to lose this possibility.

As a conclusion, I would like to propose a tentative formal declaration for a **<punct>** that I hope can serve as a basis for further discussion:

```
element punct
{
   att.global.attributes,
   att.segLike.attributes,
   attribute force { data.word }?,
   attribute unit { data.word }?,
   attribute direction { "before" | "after" | "unknown" |
"inapplicable" }?,
   attribute source { data.word }?,
   ( text | model.gLike | model.cLike | model.pPart.edit )*
}
```

This declaration contains a new model (**model.cLike**) introduced to avoid a direct reference to a single element. If the **model.segLike** were used instead, sentences, phrases and words would be allowed inside punctuation marks, which would be even more absurd than a clause inside a morpheme. *En revanche*, none of the elements belonging to **model.pPart.edit** seems theoretically impossible inside a punctuation mark.

References

HAUGEN, Odd Einar (ed.) (2006). *MUFI Character Recommendation. Characters in the Official Unicode Standard and in the Private Use Area for Medieval Texts Written in the Latin Alphabet. Part 1: Alphabetical order. Version 2.0,* Bergen : Medieval Unicode Font Initiative, http://www.mufi.info.

PARKES, Malcolm B. (1992). *Pause and Effect: an Introduction to the History of Punctuation in the West,* Aldershot: Scolar Press.