

Bernard Desgraupes, [Sylvain Loiseau](#) and [Benoît Habert](#)

Contents

- [1 Can Wikipedia be used as a corpus?](#)
- [2 Wiki2Tei: a converter of wiki syntax into standard corpus markup](#)
- [3 Conversion strategy](#)
- [4 References](#)
- [5 Links](#)

Can Wikipedia be used as a corpus?

Articles in the online encyclopedia *Wikipedia* use a special syntax, called the *mediawiki syntax*, for their formatting and typesetting. For online rendition on the Wikipedia web site, articles are transformed into HTML on the fly by a mediawiki parser. The simplicity of the mediawiki syntax is a key feature of Wikipedia, and probably a condition of its success, since it allows everyone to edit Wikipedia articles with very few technical skills.

Beside its intended usage as an encyclopedia, Wikipedia is more and more often used as a linguistic resource. Wikipedia provides a clean base of texts (much more carefully written than average texts found on the internet), free of charge, indexed into thematic categories, with various contextual pieces of information, and it provides aligned texts in various languages. It is used for instance in several evaluation campaigns, or for description of web genres.

However, the mediawiki syntax is very unsuited for any other task than online rendition and edition of articles. Neither the mediawiki syntax, nor its HTML equivalent, allow us to identify the different components of the text (title, array, lists, templates, links, images, etc.). Without any handling of these components, it is not possible to take full advantage of the Wikipedia database: texts are corrupted with various components irrelevant for a given linguistics task. Moreover, any fine study of the textual properties of the Wikipedia articles is limited by the difficulty to address the different sections and components of the text.

Wiki2Tei: a converter of wiki syntax into standard corpus markup

The wiki2Tei parser addresses this problem by parsing the mediawiki syntax and converting it into a standard data format. The format used is the XML-based [Text Encoding Initiative](#) vocabulary, which is the most elaborated standard for corpus encoding. The conversion is intended to preserve as far as possible all information available in the wiki syntax. The Wiki2Tei converter tries to express the logical content of the wiki markup rather than its rendition on the Wikipedia web site.

The Wiki2Tei parser converts highlight (see (1) on figure), links (2), title (3), inferring the division of the text, list (4) and template (5), or category tag (8), among many other features.

Conversion strategy

Rather than parsing the wiki text with a new parser, our choice is to modify the mediawiki software:

- Firstly, because the parser already tries to produce an XML syntax (in the HTML vocabulary), so the major part of the existing software can be reused. But much work was needed in order to make the conversion more strict and to ensure that every output document is a well-formed XML document.
- Secondly because the mediawiki software is the de facto normative definition of the mediawiki format: all human readable documentation of the parser is trying to guess and describe the behaviour of the mediawiki parser and is neither complete nor reliable;
- Thirdly, because the mediawiki parser makes much useful information available, such as resolving templates and variables. The parser is deeply coupled with the text and the pages themselves and contains a major part of their content: the texts in this corpus are not independent from the software which helps to generate them.

The Wiki2Tei parser is an open source piece of software available on [sourceforge](#)

Converting some components of the wiki syntax

References

Habert B. (2005) *Instruments et ressources électroniques pour le français*, Paris : Ophrys.

Loiseau S. (2007) « CorpusReader : un dispositif expérimental pour construire des observables », *Corpus*, n°6 : « Contexte, interprétation, codage ».

Poudat C. & Loiseau S. (2006) *Représentation et caractérisation lexicale des sciences dans Wikipédia*, Revue Française de Linguistique Appliquée, n°2007-2, « Lexique de la langue scientifique ».

Links

LIR