

# The VESPA tagging manual

## Version 2.3

Magali Paquot: [magali.paquot@uclouvain.be](mailto:magali.paquot@uclouvain.be)  
Signe Oksefjell Ebeling: [s.o.ebeling@ilos.uio.no](mailto:s.o.ebeling@ilos.uio.no)  
Alois Heuboeck: [a.heuboeck@reading.ac.uk](mailto:a.heuboeck@reading.ac.uk)  
Larry Valentin : [larry.valentin@uclouvain.be](mailto:larry.valentin@uclouvain.be)





## Table of contents

Acknowledgements .....	iii
0. Introduction: the VESPA corpus collection guidelines .....	1
0.1. Collect the right type of material .....	1
0.2. Ask students to fill in a learner profile .....	1
0.3. Rename student texts .....	2
0.4. Annotate and format VESPA texts .....	2
1. Setting up the VESPA Word macros .....	4
1.1. Installing and setting up the VESPA Word macros .....	4
1.2. Changing macro security settings .....	4
1.3. Disabling the VESPA Word macros .....	4
1.4. Debugging the VESPA Word macros .....	4
1.5. Uninstalling the VESPA macros .....	5
1.6. Frequent installation problems .....	5
1.6.1. Runtime error “5834” .....	5
2. Using the VESPA macro: ‘Manual tagging’ .....	6
2.1. Before starting... ..	6
2.1.1. The docx format.....	6
2.1.2. Compatibility issue with <i>Microsoft Word 2013</i> .....	6
2.1.3. Layout issue .....	6
2.2. Run the macro .....	6
2.3. The different steps of the manual tagging .....	7
2.3.1. A useful pre-processing stage: Highlight quotes and formatted text passages.....	7
2.3.2. Select features to tag .....	9
2.3.3. Document title .....	9
2.3.4. Division types .....	10
2.3.5. List .....	15
2.3.6. Block quotes and mentioned items .....	16
2.3.7. Tables and figures .....	19
2.3.8. Formulae .....	20
2.3.9. Comments .....	21
2.4. Saving, closing and modifying the document .....	21
3. Using the VESPA macro: “Automatic tagging” .....	22
3.1. Run the macro .....	22
3.2. Start automatic tagging .....	22
3.3. Saving and closing the document.....	22
4. Post-processing: <i>Perl</i> script.....	23

4.1.	Installing <i>Perl</i> .....	23
4.2.	Converting the <i>Excel</i> database.....	23
4.3.	Setting up the <i>Perl</i> script.....	23
4.4.	Using the <i>Perl</i> script.....	24
4.5.	Troubleshooting.....	25
4.5.1.	Frequent tagging mistakes .....	25
4.5.2.	Debug process .....	27
	References .....	29
	Appendix 1: VESPA learner profile .....	30
	Appendix 2: The subset of TEI (P5) used in VESPA.....	33

## **Acknowledgements**

The VESPA macros and Perl scripts are largely based on the macros developed by Alois Heuboeck for the British Academic Written English (BAWE) corpus (cf. Ebeling & Heuboeck 2007; Heuboeck et al. 2008), and were adapted for VESPA also by Alois Heuboeck (Reading University, UK).

We are indebted to the research project 'An Investigation of Genres of Assessed Writing in British Higher Education' for developing the British Academic Written English (BAWE) corpus from which we borrowed the original Word macros and Perl scripts used to encode and format VESPA texts. The BAWE corpus was developed at the Universities of Warwick, Reading and Oxford Brookes, under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC.

See <http://wwwm.coventry.ac.uk/researchnet/BAWE/Pages/BAWE.aspx> for more information.

We gratefully acknowledge the support of the Department of Literature, Area Studies and European Languages at the University of Oslo, Norway, for funding the development of the VESPA Word macros and Perl scripts.

## 0. Introduction: the VESPA corpus collection guidelines

### 0.1. Collect the right type of material

The corpus will consist entirely of L2 academic writing in a wide range of:

- **disciplines** (linguistics, business, medicine, law, biology, etc),
- **genres** (papers, reports), and
- **degrees of writer expertise** in academic settings (from first-year students to PhD students).

Texts should be **at least 500 words long** (e.g. lab reports) but may be much longer (e.g. term papers). They should be handed in in electronic format. This reduces the time spent typing up student texts and minimizes the risk of introducing errors into the text.

*Work should be entirely the students' own*, i.e. no help should be sought from third parties, but reference tools such as dictionaries and grammar books are acceptable (use of reference tools should be indicated on the learner profile questionnaire). Texts produced by more than one student (e.g. collaborative work) and revised versions of texts (e.g. following teachers' comments) should not be included in the corpus.

Argumentative, descriptive and narrative subjects are not of interest. For this reason, the following types of titles should be avoided:

- "Crime does not pay"
- "Feminism has done more harm to the cause of women than good"
- "Pollution : a silent conspiracy"
- "The joys of the English countryside"
- "My year in America"

### 0.2. Ask students to fill in a learner profile

The VESPA learner profile has been created in order to provide researchers with information about contributors; this will enable meaningful conclusions to be drawn from the results obtained when the corpus is analysed. Using the profile, researchers will both be able to draw general conclusions about advanced learner writing in the discipline, and to examine subsections e.g. Spanish mother tongue learners, learners who speak some English at home, learners for whom German is the second language and English is the third language. It will also be possible to examine sociolinguistic aspects such as male/female comparisons. If the corpus is used as a basis for developing specifically adapted teaching tools, the potential advantages of this facility are clear.

The VESPA learner profile is available in two forms. International partners can either:

- ask their students to fill in a paper version of the questionnaire (including permission form) (see Appendix 1). In that case, VESPA partners have to encode students' answers in an Excel file (= CTXDATA.xls).
- ask their students to fill in an online questionnaire (including permission form) hosted on one of the Université catholique de Louvain's servers (contact the VESPA project director for more detail).

Each partner will have to attribute a code to each student and ask them to use this code when they fill in the learner profile (and to be very careful to type it correctly!).

A **student code** consists of **3 letters for the institution + 4 digits for the student**. Thus, at the Université catholique de Louvain, we give students codes starting with:

- UCL0001
- UCL0002
- UCL0003

Partners may opt for the paper or the electronic version of the questionnaire. However, they need to store the resulting learner profiles in the **CTXDATA.xls** file which will serve as input data for post-processing (see Section 4).

#### Important remarks:

- Your institution code will be provided to you by the VESPA project director to make sure two institutions do not use the same code.
- If a student contributes several texts to the corpus, ...
  - She should only be given one code (and not a code per course!). This is the only way we'll be able to identify several texts written by the same student while ensuring anonymity.

- She should fill in as many learner profiles as the number of texts she submitted to make sure we have the necessary information on all texts.

### **0.3. Rename student texts**

Student texts (doc files) should be renamed before they are run through the VESPA macros.

A **text file code** consists of the **student code + hyphen + a 2-3 letter code for the course + a 2 digit code for the task**. Thus, for task 1 (e.g. report) in the course 'Business Studies' at the universit  catholique de Louvain, texts will be named:

- UCL0001-BUS-01
- UCL0002-BUS-01
- UCL0003-BUS-01

If student 0001 writes a second text (e.g. term paper) for the same course, the code of the text will be:

- UCL0001-BUS-02

If student 0001 writes a text for another course (e.g. business communication), the code of the text will be:

- UCL0001-BCO-01

**Note:** Course codes should also be agreed upon with the VESPA project director.

Text file codes should also be copied in the first column of the **CTXDATA** file.

### **0.4. Annotate and format VESPA texts**

Student texts are usually submitted to the VESPA corpus as Microsoft Word documents. However, this format proves impractical for efficient processing of a corpus. The documents need to be converted to plain text format, which in turn requires pre-processing them to avoid loss of relevant information.

Concerning the encoding of the VESPA corpus, a decision was made to apply the encoding standard proposed by the Text Encoding Initiative (TEI) P5 guidelines.

Following what was done within the framework of the BAWE project (cf. Ebeling & Heuboeck 2007; Heuboeck et al. 2008), a number of computer tools enabling semi-automatic and automatic processing of the texts collected were developed to facilitate the encoding and mark-up process.

There are 3 main steps involved in the preparation of students' texts for the VESPA corpus:

- Step 1:** Interactive annotation of titles, sections, quotes, examples, etc.
- Step 2:** Automatic conversion to XML format.
- Step 3:** A cascade of Perl scripts is used to finalize the formatting process: normalization of hyphens and dashes, transformation of Microsoft XML input to TEI-conformant tags, importation of contextual information from external spreadsheets, mark-up of sentence boundaries, etc.

An interface for interactive manual annotation (Step 1) was developed in the form of a series of Word macros, written in Visual Basic and making use of graphical user interface possibilities. This interface has been set up to guide the tagger through the annotation process step by step (see Section 2.).

As put by Ebeling & Heuboeck, it facilitates the human tagger's task in various respects:

- Operating within Word, the human tagger still has the original formatting available during the tagging process. Interpreting formatted text involves considerably less effort than interpreting unformatted text;
- Tags can be selected from options, thus avoiding any typing. The options appear as checkboxes, radio buttons, drop down lists or labelled keys;
- By organising the tagging process in two layers, i.e., first selecting from the functions available and then annotating these functions, the tagging interface, changing throughout the process, is always focussed on the function being annotated. The tagger only chooses relevant options for this function; and,

- Thus, the tagging interface is tailored to the requirements of the [VESPA] corpus: both layers of annotation, functions and specific options describing their realisation, are designed to direct and limit the annotator's choice [...]. (Ebeling & Heuboeck 2007: 251-252),

Step 2 relies on a Word macro that partners just need to run on a batch of VESPA files to convert them to XML format (see Section 3.).

When they have a batch of VESPA texts that have gone through Steps 1 and 2, partners should run files through the Perl scripts (see Section 4).



## 1. Setting up the VESPA Word macros

### 1.1. Installing and setting up the VESPA Word macros

- Step 1: open a new window and copy-paste the following path:
  - If your OS is Windows XP users: %userprofile%\Application Data\Microsoft\Word\;
  - If your OS is Windows XP, 7 or 8: %userprofile%\AppData\Roaming\Microsoft\Word\;
- Step 2: open the **Startup** folder:
  - If this **Startup** folder is absent, simply create it;
- Step 3: copy-paste the **VESPA\_Tagging.dot** file in the **Startup** folder. The macros will be automatically loaded when you use Microsoft Word.

### 1.2. Changing macro security settings

You should follow these steps only if *Microsoft Word* issue a security alert and ask whether the macro must be enabled or disabled:

- Step 1: launch **Microsoft Word**;
- Step 2: click on the **Office button** and on **Word Options**;
- Step 3: click on **Trust center**;
- Step 4: under *Microsoft Office Word Trust Center*, click on **Trust Center Settings...**;
- Step 5: click on **Macro Settings**;
- Step 6: under *Macro Settings*, select **Disable all macros with notification** and click on **OK**;

### 1.3. Disabling the VESPA Word macros

It is recommended to disable the macros when you do not need to tag texts:

- Step 1: open a new window and copy-paste the following path:
  - If your OS is Windows XP users: %userprofile%\Application Data\Microsoft\Word\Startup\;
  - If your OS is Windows XP, 7 or 8: %userprofile%\AppData\Roaming\Microsoft\Word\Startup\;
- Step 2: **cut** the **VESPA\_Tagging.dot** files and paste it in a different folder on your computer:

If you want to use the macros again, simply follow the steps described in the section **1.1. Installing and setting up the VESPA Word macros**.

### 1.4. Debugging the VESPA Word macros

You should follow these steps only if you need to debug the macros.

- Step 1: disable the macros (see **1.3. Disabling the VESPA Word macros**)
- Step 2: launch **Microsoft Word**;
- Step 3: while holding down the **Alt** key, press the **F11** key to open **Microsoft Visual Basic**;

- Step 4: while holding down the **Ctrl** key, press the **R** key to open the **Project Explorer**;
- Step 5 (optional): press the **F4** key to open the **Properties Window**;
- Step 6: click on **Project (VESPA Tagging)**, then on **Modules**. This folder contains the source code of the macros
- Step 7: when debugging is complete, while holding down the **Ctrl** key, press the **S** key to save the module;
- Step 8: close **Microsoft Visual Basic** and **Microsoft Word**;
- Step 9: activate the macros (see 1.1. [Installing and setting up the VESPA Word macros](#))

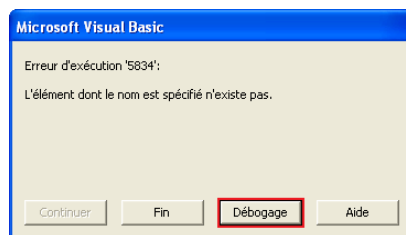
## 1.5. Uninstalling the VESPA macros

- Step 1: open a new window and copy-paste the following path:
  - If your OS is Windows XP users: `%userprofile%\Application Data\Microsoft\Word\Startup\`;
  - If your OS is Windows XP, 7 or 8: `%userprofile%\AppData\Roaming\Microsoft\Word\Startup\`;
- Step 2: delete the **VESPA\_Tagging.dot** file in the **Startup** folder.

## 1.6. Frequent installation problems

### 1.6.1. Runtime error "5834"

Runtime errors "5834" (cannot find element with this name) occur with non-English versions of Microsoft Word. These errors are due to the different style names used in various language versions of Word



Please contact the VESPA team and they will provide you with an appropriate version of the VESPA macros. Mention the language version of Microsoft Word.

## 2. Using the VESPA macro: 'Manual tagging'

### 2.1. Before starting...

#### 2.1.1. The docx format

Documents in **Office Open XML** file format (.docx) **cannot be tagged** with the *VESPA Macros*. These files must be saved in .doc format.

→ Step 1: open the .docx file to convert;

→ Step 2: click on the **Office button**, then on **Save As** and on **Word 97-2003 Document**;

→ Step 3: choose a folder, then in front of **Save as type:**, check that **Word 97-2003 Document (\*.doc)** is selected and click on **Save**.

#### 2.1.2. Compatibility issue with Microsoft Word 2013

When opening a document downloaded from the Internet, the following message may occur: "Runtime Error '4248': This command is not available because no document is open." This is due to the fact that Microsoft Word 2013 considers such a file as a potential threat for your computer. In order to solve this problem:

→ Step 0 (if required): close the downloaded file;

→ Step 1: right-click on the downloaded file and select **Properties**;

→ Step 2: under the **General** tab, in front of **Security**, click on **Unblock** and on **OK**;

#### 2.1.3. Layout issue

**!!! All manually inserted page and section breaks as well as Word styles (e.g. title formatting) need to be removed before running the automatic tagging macro.**

You should take a few minutes to browse through the text to be tagged and delete all page and section breaks. They will be easily spotted if you turn the "Show/Hide Paragraph marks and other hidden formatting symbols" option on.

## 2.2. Run the macro

→ Step 1: open the file to tag.

↳ The *VESPA* toolbar is automatically loaded with other *Microsoft Word* toolbars;

→ Step 2: click on **Add-Ins** and on ♥ **VESPA: man** to run the macro.



**!!! The macro might not work if the *Find and Replace* window is active.**

Note that the formatting marks option (¶) will be turned on automatically.

Original elements	Formatting marks
2. Theoretical framework	2. Theoretical framework¶
2.1. Language of advertising	2.1. Language of advertising¶

Table 1. Formatting marks

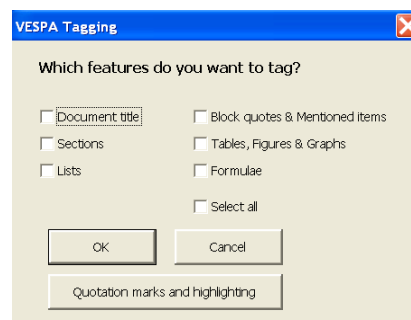
## 2.3. The different steps of the manual tagging

### 2.3.1. A useful pre-processing stage: Highlight quotes and formatted text passages

Before starting tagging, it is highly recommended that you use the 'quotation marks and highlighting' option to highlight quotation marks and formatted texts (e.g. bold characters, italics, underlined words).

The main objective of this option is to help you identify text passages that need to be tagged in a specific way (e.g. quotes, mentioned items including linguistic examples, etc).

→ Step 1: click on **Quotation marks and highlighting**;



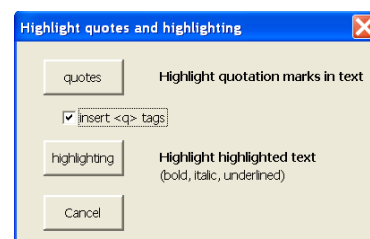
#### 2.3.1.1. Quotation marks

The 'quotes' tool highlights all quotation marks (" ", ' ') in the text. It is particularly useful to identify text passages that should be tagged as quotes or linguistic examples (see below). The tool can be used in two different ways.

##### 2.3.1.1.1. Highlighting and tagging quotation marks

As many quotation marks are used to identify quoted passages, book titles, etc. (which will need to be tagged), it is possible to automate the insertion of {q} pseudo-tags (see Section 2.4.6). A Perl script will then automatically replace all {q} pseudo-tags by TEI-conformant <q> tags.

→ Step 2a: click on **quotes** to highlight quotation marks. Make sure that **insert <q> tags** is selected.



Original elements	Highlighted and tagged elements
Register here refers to Biber's definition that characterises register as "a general cover term for all language varieties associated with different situations and purposes" (in: Vestnik, 2003: 3).	Register here refers to Biber's definition that characterises register as: {q}"a general cover term for all language varieties associated with different situations and purposes" {/q} (in: Vestnik, 2003: 3).¶
Gaskell, D. & Cobb, T. (2004). "Can learners use concordance feedback for writing errors?". System, vol. 32, n° 3. pp. 301-319. <a href="http://www.lexutor.ca/cv/pdf/concordance_feedback.pdf">http://www.lexutor.ca/cv/pdf/concordance_feedback.pdf</a> .	Gaskell, D. & Cobb, T. (2004). {q}"Can learners use concordance feedback for writing errors?" {/q}. System, vol. 32, n° 3. pp. 301-319. <a href="http://www.lexutor.ca/cv/pdf/concordance_feedback.pdf">http://www.lexutor.ca/cv/pdf/concordance_feedback.pdf</a> .¶

Table 2a. Highlighted and tagged quotation marks

**!!! The drawback of this fully automatic procedure is that {q} pseudo-tags are also inserted where the marks are not quotation marks:**

Original elements	Highlighted and tagged elements
It introduces additional information that emphasize what you've just said.	It introduces additional information that emphasize what you: {/q}ve just said.¶
In most of the grammars, this phenomenon is also known as "rankshifting".	In most of the grammars, this phenomenon is also known as {q}"rankshifting" {/q}.¶

Table 2b. Incorrect examples of highlighted and tagged quotation marks

All these 'false' {q} pseudo-tags should be **deleted** from the text (cf. Section 2.4.6). Only {q} tags that are used to mark **quotes** or **book titles** should be **kept**.

### Summary: The {q} pseudo-tag

The {q} pseudo-tag is inserted automatically by the 'Quotation marks and highlighting' tool next to punctuation marks that can be used as quotation marks (“,”, ‘’, ’). It is supposed to help you identify quoted passages, book titles, etc. which we want to tag as they were not produced by the student. All {q} pseudo-tags need to be checked and 'false' {q} tags need to be removed. (Please note that quotations (excluding block quotes, cf. Section 2.4.6.2) should also be tagged in foot-/endnotes.)

Curly brackets are not usually used in tags but Word macros do not allow us to use angle brackets at this stage. A Perl script run on all VESPA files will automatically replace all pseudo-tags by TEI-conformant tags at a later stage.

As put by Ebeling & Heuboeck (2007:252), "since Word does not allow to make a distinction between text and meta-text, there is no possibility to insert genuine XML tags at this point. Any sequence of characters would be subject to transformation in the process of recoding the Word document as XML file; in particular, the tag delimiters '<' and '>' would not be recognised as metacharacters. It was thus decided to insert 'pseudo-tags' at this stage. Only after the document has been converted to XML format (...), these 'pseudo-tags' are transformed to genuine XML tags by a Perl script."

### Options:

a. If the student did not insert an end-of-quotation mark at the end of a quote, you can add it as well as a {q} pseudo-tag if you want to tag the quote:

According to Halliday, {q}"the textual metafunction is transparent [*quotation mark left out*].

→ According to Halliday, {q}"the textual metafunction is transparent"/{q}.

If this is not corrected, the quote will not be tagged as <q>.

b. If the student used italics to quote instead of quotation marks, you can add {q} pseudo-tags if you want the text passage to be tagged as a quote:

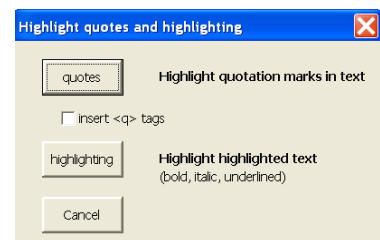
According to Halliday, *the textual metafunction is transparent*.

→ According to Halliday, {q}"the textual metafunction is transparent"/{q}.

#### 2.3.1.1.2. Highlighting quotation marks only

As the macro adds many irrelevant {q} pseudo-tags, another option is to highlight all quotation marks and manually insert {q} pseudo-tags where appropriate. However, this option is **NOT** recommended and should only be used, for example, to check whether the quotation marks are predominantly used for mentioned items rather than quoted items.

→ Step 2b: uncheck *insert <q> tags* and click on *quotes* to highlight quotation marks only.



Original elements	Highlighted elements
Register here refers to Biber's definition that characterises register as "a general cover term for all language varieties associated with different situations and purposes" (in: Vestnik, 2003: 3)	Register here refers to Biber's definition that characterises register as "a general cover term for all language varieties associated with different situations and purposes" (in: Vestnik, 2003: 3)

Table 3. Highlighted quotation marks

### 2.3.1.2. Formatted text

The 'highlighted text' tool highlights all bold characters, italics and underlined passages in the text. By default, highlighted text passages will receive a <hi>-tag in the post-processing stage (see Section 4.).

This pre-processing stage is also very useful to identify text passages that should be tagged as quotes or examples (see below).

→ Step 3: click on **highlighting** to highlight emphasis (bold), italic or underlined text.



Original elements	Highlighted elements
<i>Language in print advertising - a comparative study of French, English and German product advertisements of the brand NIVEA</i>	<b>Language in print advertising - a comparative study of French, English and German product advertisements of the brand NIVEA</b>

Table 4. Highlighted formatted marks

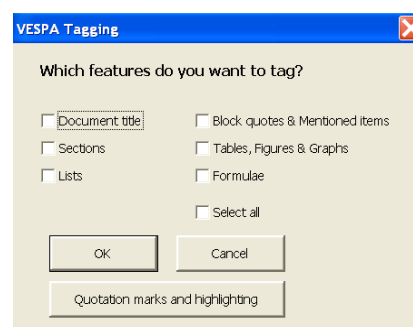
→ Step 4: click on **Cancel** to return to the previous window.

### 2.3.2. Select features to tag

The next step is to select the features that will need to be tagged. By default, you should tag all features (document title, sections, lists, block quotes and linguistic examples, tables, figures and graphs, and formulae). If you know that some features (e.g. lists, formulae) are not represented in the text you are tagging, you can deselect them and the corresponding windows will not be displayed.

→ Step 1: select the features that need to be tagged or check **Select all** in order to choose all features (recommended option in case of doubt);

→ Step 2: click on **OK**.



### 2.3.3. Document title

The document title is the title of the text that the student wrote. Do not include anything else within the document title tags. **All other information on the title page (e.g. course name, student's name, teacher's name, year) should be deleted.**

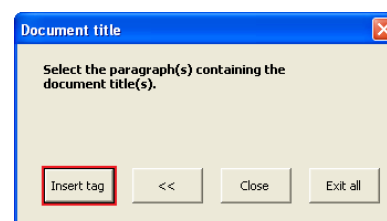
If the document includes instructions (e.g. "Comment on ... ; Discuss the relationship between ...), you can either delete them, or include them within the document title tags.

→ Step 1: place the cursor on the document title. If the document title spreads over several paragraphs, make sure you select them all.

→ Step 2: click on **Insert tag**.

→ Step 3:

- 3a: click on << to return to the previous window;
- or 3b: click on **Close** to move to the next selected feature (if no further option has been selected, the macro dialog box will be closed);
- or 3c: click on **Exit all** to close the macro dialog box.



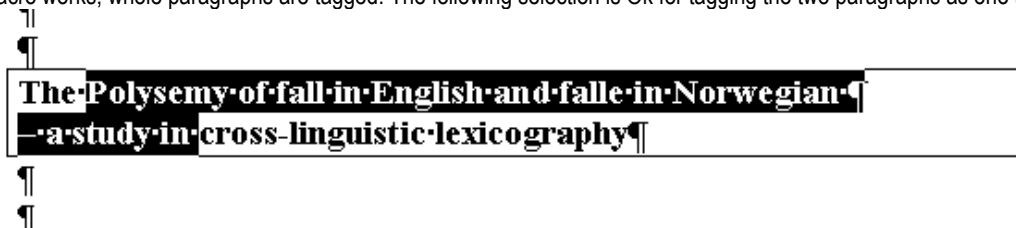
Original elements	Tagged elements
Mary G. GERM 2823: Lexicology - NOMA: 2607xxxx Academic Year 2008-2009 - May 4th, 2009  <u><i>Language in print advertising - a comparative study of French, English and German product advertisements of the brand NIVEA</i></u>	<pre>{start:vespa_documentTitle}¶ <i>Language in print advertising - a comparative study of French, English and German product advertisements of the brand NIVEA</i>¶ {end:vespa_documentTitle}¶</pre>

Table 5. Tagged document title

It is not necessary to select the whole title.

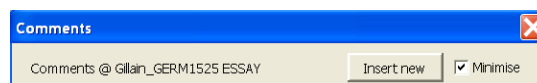
- For titles that are within one paragraph: place the cursor somewhere in that paragraph or select a bit of it
- For titles going over more than one paragraph: select at least a bit of each paragraph

As the macro works, whole paragraphs are tagged. The following selection is Ok for tagging the two paragraphs as one title:



The same comment applies to titles of division types (see Section 2.4.4).

Note that you can always add comments via the 'Comments' window if you want to specify which features you deleted or you want to keep some information (e.g. course name, instructions, etc) somewhere.

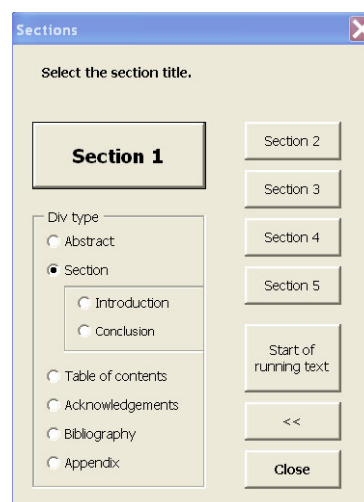


Click on 'Insert new'

### 2.3.4. Division types

Each text division should be tagged.

Division types include abstract, section (i.e. body of the text), introduction, conclusion, table of contents, acknowledgements, bibliography and appendix.



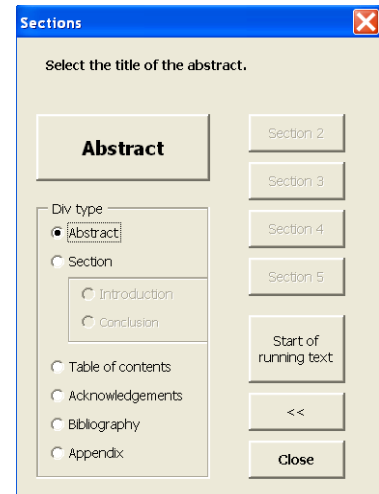
Please note that sections, the abstract, the table of contents, the bibliography, and appendices are annotated by placing opening and closing pseudo-tags around the heading only. The chunks of text appearing between these headings will be attributed to the section to which they belong later, via Perl scripts.

### 2.3.4.1. Start of running text

If there is no heading indicating the start of the first section (i.e. the beginning of running text), place the cursor before the main text (and after all other features such as document title, abstract, etc) and click on '**Start of running text**'.

### 2.3.4.2. Abstract

- Step 1: under *Div type*, select **Abstract**;
- Step 2: place the cursor on the abstract title;
- Step 3: click on the button **Abstract**.



Original elements	Tagged elements
Abstract	<code>{start:vespa_abstract-1}¶</code> Abstract¶ <code>{end:vespa_abstract-1}¶</code>

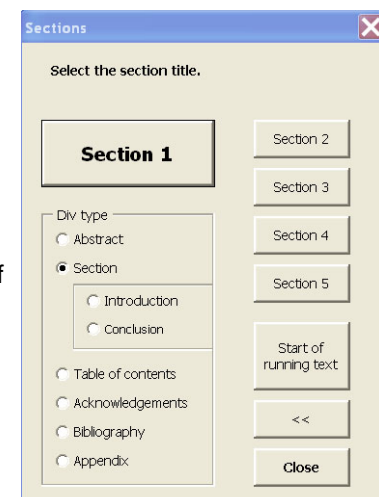
Table 6. Tagged abstract title

### 2.3.4.3. Sections

Sections are parts of the main text that are introduced by a title.

- Step 1: under *Div type*, select **Section**;
- Step 2: place the cursor on a section title
- Step 3: click on **Section x**. Select Section 1 if it is the title of a main section; Section 2 if it is the title of a sub-section; etc.

Repeat those steps for all section titles.



Please note that the number that follows the section name on the 'Sections' tool window represents the level (e.g. main section, sub-section), not the number of the section (see examples in Table 7.).

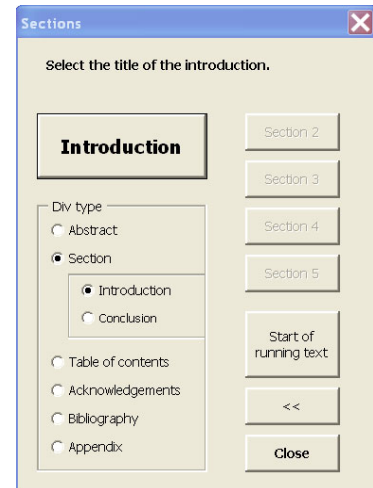


Original elements	Tagged elements
3. Methodology	{start:vespa_section-1}¶ 3. Methodology¶ {end:vespa_section-1}¶
4.4. Words clusters analysis	{start:vespa_section-2} 4.4. Words clusters analysis {end:vespa_section-2}

Table 7. Tagged section title

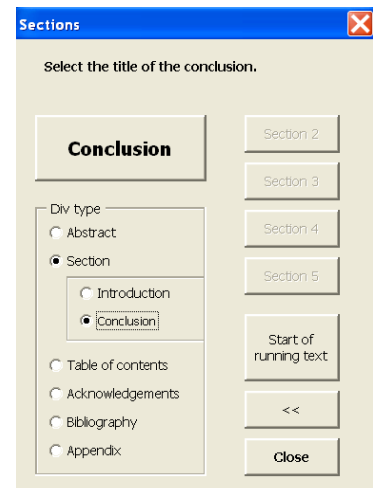
If the introduction is clearly identified (e.g. via a title 'Introduction'), you can specify that the section you selected is the introduction: Do not use the 'introduction' and 'conclusion' sections if they are not clearly identified in the text (e.g. by a title).

- Step 1: under *Div type*, select **Section** and **Introduction**;
- Step 2: place the cursor on the introduction title
- Step 3: click on the button **Introduction**.



If the conclusion is clearly identified (e.g. via a title 'Conclusion'), you can specify that the section you selected is the conclusion:

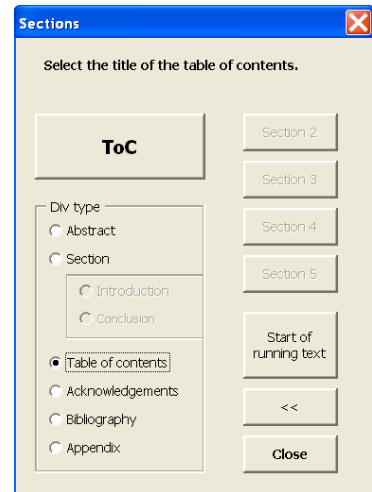
- Step 1: under *Div type*, select **Section** and **Conclusion**;
- Step 2: place the cursor on the conclusion title
- Step 3: click on the button **Conclusion**.



**Note:** The division types 'introduction' and 'conclusion' are optional. However, tagging the introduction and the conclusion will allow users to query specific parts of the students' texts.

#### 2.3.4.4. Table of contents

- Step 1: under *Div type*, select **Table of contents**;
- Step 2: place the cursor on the title of the table of contents;
- Step 3: click on **ToC**.

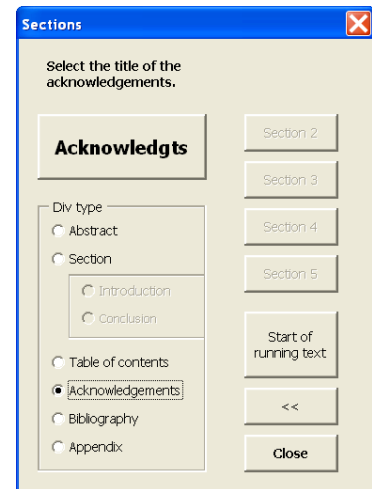


Original elements	Tagged elements
Table of contents	<code>{start:vespa_toc-1}</code> ¶ Table of contents¶ <code>{end:vespa_toc-1}</code> ¶

Table 8. Tagged title of the table of contents

#### 2.3.4.5. Acknowledgements

- Step 1: under *Div type*, select **Acknowledgements**;
- Step 2: place the cursor on the title of the acknowledgements;
- Step 3: click on **Acknowledgts**.

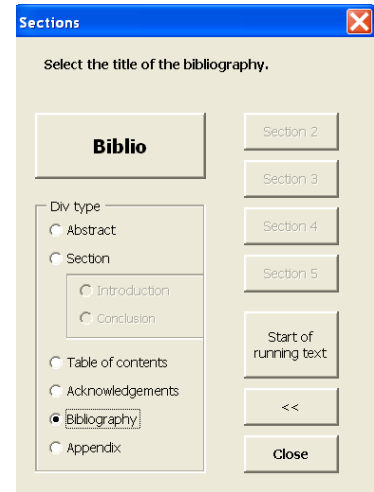


Original elements	Tagged elements
Acknowledgements	<code>{start:vespa_acknowledgements-1}</code> ¶ Acknowledgements¶ <code>{end:vespa_acknowledgements-1}</code> ¶

Table 9. Tagged title of the acknowledgements

### 2.3.4.6. Bibliography

- Step 1: under *Div type*, select **Bibliography**;
- Step 2: place the cursor on the title of the bibliography;
- Step 3: click on **Biblio**.

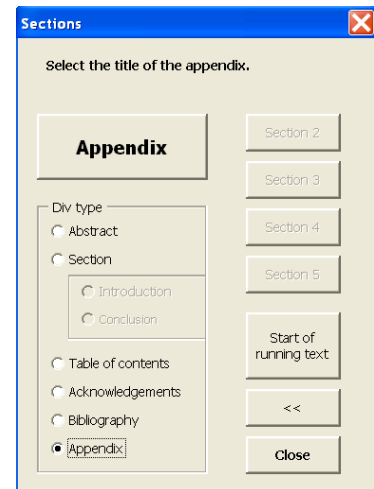


Original elements	Tagged elements
References:	{start:vespa biblio-1}¶ References: ¶ {end:vespa biblio-1}¶

Table 10. Tagged title of the bibliography

### 2.3.4.7. Appendix

- Step 1: under *Div type*, select **Appendix**;
- Step 2: place the cursor on the title of the appendix;
- Step 3: click on **Appendix**.



Original elements	Tagged elements
<u>Appendix 1 – Product headlines</u>	{start:vespa appendix-1}¶ <u>Appendix 1 – Product headlines</u> ¶ {end:vespa appendix-1}¶

Table 11. Tagged title of the appendix

- Step 4: when all titles have been tagged:
  - 4a: click on << to return to the previous window;
  - or 4b: click on **Close** to move to the next selected feature (if no further option has been selected, the macro dialog box will be closed).

### 2.3.5. List

A list is a set of things, names, numbers, etc. usually written one below the other (example 1), numbered (example 2) or introduced with bullet points, hyphens or dashes (example 3).

Examples of lists
<p>An essay must include the following sections:</p> <p>Introduction Content Conclusion Bibliography</p>
<p>An essay must include the following sections:</p> <ol style="list-style-type: none"> <li>1. Introduction</li> <li>2. Content</li> <li>3. Conclusion</li> <li>4. Bibliography</li> </ol>
<p>An essay must include the following sections:</p> <ul style="list-style-type: none"> <li>• Introduction</li> <li>• Content</li> <li>• Conclusion</li> <li>• Bibliography</li> </ul>

Typically, a list item contains a rather small amount of text, consisting of only one word or phrase but it may also contain full sentences or list-like formatted paragraphs.

Example of a list containing full sentences
<p>Based on the aims stated above, three main hypotheses were explored:</p> <ol style="list-style-type: none"> <li>1. There will be a high positive correlation (<math>r &gt; 0.75</math>) between the level achieved for academic text comprehension in English and the students' level of proficiency in English.</li> <li>2. There will be a medium positive correlation (<math>r &gt; 0.50</math> and <math>&lt; 0.74</math>) between the level achieved for academic text comprehension in English and the students' degree of disciplinary expertise.</li> <li>3. There will be a high positive correlation (<math>r &gt; 0.75</math>) between the level achieved for academic text comprehension in English and the level achieved for academic text comprehension in Spanish.</li> </ol>
Example of list-like formatted paragraphs
<ol style="list-style-type: none"> <li>1. the crossing gates were coming down. → “The crossing gates” is plural so, for rules of concord between subject and verb, the singular verb was is wrong and must be replaced by the plural verb form in order to match with the plural subject.</li> <li>2. The relative pronoun who is used to refer back to people. The reference here, though, is to an animal. The relative pronoun who must be replaced by the relative pronoun which, which is used to refer back to non-personal nouns.</li> <li>3. Draw is an irregular verb. Its past form is not formed by adding the regular plural ending -ed, but through a vowel change in the middle of the stem. The correct form, thus, is drew.</li> <li>4. It has to be replaced by there, as the empty subject it is used to refer to something later in the clause and is usually followed by a definite noun phrase, while there is used to introduce new information or to say that something exists or happens, and can be followed by an indefinite noun phrase, as it is the case in a man in a sports car.</li> </ol>

However, lists included in the main text should **NOT** be tagged as <list> as the <s> tag (sentence) cannot include a <list> tag:

“False lists” examples
An essay must include the following sections: (1) Introduction; (2) Content; (3) Conclusion and (4) Bibliography
An essay must include the following sections: (a) Introduction; (b) Content; (c) Conclusion and (d) Bibliography

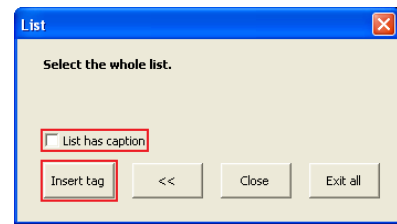
→ Step 1: select the whole list to tag in the document, then select *List has caption* (if necessary);

→ Step 2: click on *Insert tag*.

Repeat those steps if required;

→ Step 3:

- 3a: click on << to return to the previous window;
- or 3b: click on **Close** to move to the next selected feature (if no further option has been selected, the macro dialog box will be closed);
- or 3c: click on **Exit all** to close the macro dialog box.



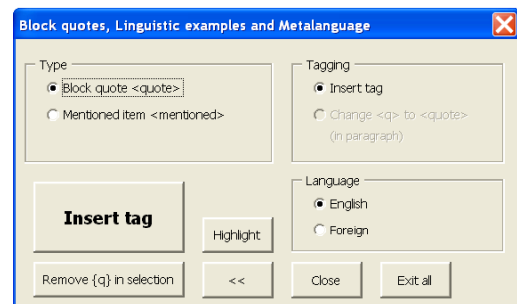
Original elements	Tagged elements
An essay must include the following sections:	An-essay must include the following sections:¶
Introduction	{start:vespa_list}¶
Content	Introduction¶
Conclusion	Content¶
Bibliography	Conclusion¶
	Bibliography¶
	{end:vespa_list}¶

Table 12. Tagged list

### 2.3.6. Block quotes and mentioned items

#### 2.3.6.1. Highlight quotes and highlighting

→ Click on **Highlight** to open the *Highlight quotes and highlighting* window if quotation marks and highlighting have not been highlighted and/or tagged yet (see 2.4.1. *A useful pre-processing stage: Highlight quotes and formatted text passages*);



Remember that the drawback of this fully automatic procedure is that **{q} pseudo-tags are also inserted where the marks are not quotation marks**. All these 'false' {q} pseudo-tags should be **deleted** from the text, by using the button "**Remove {q} in selection**".

In particular, the human tagger should make sure that there are no {q} within {q}- pseudo tags.

#### 2.3.6.2. Block quote

A block quote is a quote from a book, an article, etc, **separated from the main text** by indentation or a new line character (example 1). Block quotes will be marked with a pseudo {quote} tag, later to be replaced by a <quote> tag via a Perl script.

Example of block quote
As quoted in Granger (1976:50), G.O. Curme (1935) stated:  "Attention is called to the fact that English "one" has a meaning somewhat different from that of the corresponding indefinite in other languages, such as German "man", French "on", etc. The force of English "one" is more indefinite (...). These German and French forms are very convenient expressions, for they make it possible to refer to a definite person of definite person without taking the time or trouble to name or describe the person or persons. (...) Thus on account of the lack of an appropriate indefinite pronoun, the passive has become a favorite form of expression in English"

**Why should I bother?**

Student writing often contains many quotes from published material. If we include excerpts from expert writing in our analysis of learner language, we may skew the results of many automatic CL tools (word lists, keyword analysis, collocation analysis, etc).

**How will annotating the corpus help me in future studies?**

CL tools such as WordSmith Tools allow you to exclude text between specific tags (e.g. <quote> </quote>) from your analysis.

A block quote (<quote>) differs from a quote in running text (<q>) in that it should only be used to tag quotes in separate paragraphs. A block quote cannot appear within a passage tagged with an <s> tag (sentence). In the following example, for instance, the quotation is part of the running text, and the {q} pseudo-tag will have to be used (later replaced by a <q> tag by a Perl script (see Section 4.)).

Use of the {q} pseudo-tag
L2 students take EAP/ESP courses everywhere around the world and analyzing the VESPA learner corpus could be an effective way of {q}“operationalizing writing difficulties”{/q} (Bitchener and Basturkmen 2006, p. 14).

The <quote> tag should not be used in foot-/endnotes.

<ul style="list-style-type: none"> <li>➔ Step 1: under <i>Type</i>, select <b>Block quote &lt;quote&gt;</b>;</li> <li>➔ Step 2: select the block quote to tag in the document;</li> <li>➔ Step 3: under <i>Language</i>, select <b>English</b> or <b>Foreign</b> according to the language of the selected element;</li> <li>➔ Step 4: click on <b>Insert tag</b>;</li> <li>➔ Step 5 (optional): select the tagged quote and click on <b>Remove {q} in selection</b> to delete any remaining {q} tags.</li> </ul> <p>Repeat steps 2 to 5 if required.</p>	
---	--

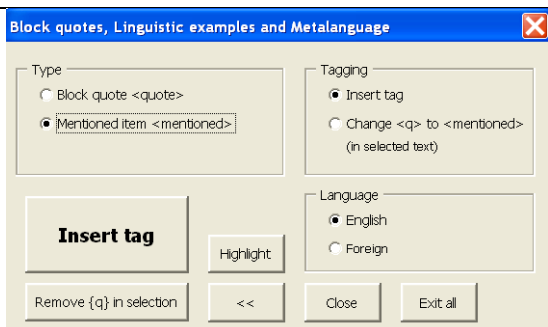
Original elements	Tagged elements
As quoted in Granger (1976:50), G.O. Curme (1935) stated:  “Attention is called to the fact that English “one” has a meaning somewhat different from that of the corresponding indefinite in other languages, such as German “man”, French “on”, etc. The force of English “one” is more indefinite (...). These German and French forms are very convenient expressions, for they make it possible to refer to a definite person of definite person without taking the time or trouble to name or describe the person or persons. (...) Thus on account of the lack of an appropriate indefinite pronoun, the passive has become a favorite form of expression in English”	As·quoted·in·Granger·(1976:50),·G.O.·Curme·(1935)·stated:¶ ¶ {start:vespa_quote}¶ “Attention·is·called·to·the·fact·that·English·“one”·has·a·meaning·somewhat·different·from·that·of·the·corresponding·indefinite·in·other·languages,·such·as·German·“man”,·French·“on”,·etc.·The·force·of·English·“one”·is·more·indefinite·(...).·These·German·and·French·forms·are·very·convenient·expressions,·for·they·make·it·possible·to·refer·to·a·definite·person·of·definite·person·without·taking·the·time·or·trouble·to·name·or·describe·the·person·or·persons.·(...)·Thus·on·account·of·the·lack·of·an·appropriate·indefinite·pronoun,·the·passive·has·become·a·favorite·form·of·expression·in·English”¶ {end:vespa_quote}¶

Table 13. Tagged block quote

**Note:** Ignore {q} pseudo-tags in block quotes as they will be ignored by the Perl scripts.

### 2.3.6.3. Mentioned items

Apart from quotes, student texts often include other types of mentioned items, e.g. cited works, foreign words, linguistic examples and passages of texts analyzed by students. All these mentioned items should be tagged with the <mentioned> tag.

<p>→ Step 1: under <i>Type</i>, select <b>Mentioned item</b> &lt;mentioned&gt;;</p> <p>→ Step 2: select the text to tag in the document;</p> <p>→ Step 3: under <i>Tagging</i>, select <b>Insert tag</b> to simply add tags <b>or</b> <b>Change &lt;q&gt; to &lt;mentioned&gt;</b> (in selected text) to convert {q} tags between the selected example;</p> <p>→ Step 4: under <i>Language</i>, select <b>English</b> or <b>Foreign</b> according to the language of the selected element;</p> <p>→ Step 5: click on <b>Insert tag</b>;</p> <p>→ Step 6 (optional): select the tagged example and click on <b>Remove {q} in selection</b> to delete any remaining {q} tags.</p> <p>Repeat steps 2 to 6 if required.</p>	
---	--

**Be careful! The <mentioned> tag should be within sentence boundaries and not used to tag full paragraphs. The <mentioned> tag should also not be used within the <docTitle> tag.**

The **Change <q> to <mentioned>** option will typically be used in examples such as the following:

Original elements with {q} tags	Tagged elements
Table 3.b displays the most frequent clusters matching the structure {q}“difference(s) + preposition”/q} in learner and expert writing.	Table 3.b displays the most frequent clusters matching the structure · {start:vespa_mentioned}“difference(s) + preposition” {end:vespa_mentioned} in learner and expert writing.
In this paper I will look at phraseological units in {q}“The Catcher in the Rye”/q}.	In this paper I will look at phraseological units in · {start:vespa_mentioned}“The Catcher in the Rye” {end:vespa_mentioned}.

Table 14. Mentioned items: Changing <q> to <mentioned>

If the mentioned item is not in English, it should be marked as “foreign” (Step 4):

Original elements	Tagged elements
Whereas the German advert claims that <i>Schönheit ist Freiheit</i> , the French version features the slogan <i>La beauté est liberté</i> .	Whereas the German advert claims that · {start:vespa_mentioned-for} · <i>Schönheit ist Freiheit</i> {end:vespa_mentioned-for}, the French version features the slogan · {start:vespa_mentioned-for} · <i>La beauté est liberté</i> · {end:vespa_mentioned-for}.

Table 15. Tagged 'foreign' mentioned items

Please note that the <mentioned> tag should not be used to tag the terminology of the discipline. If a term is placed between quotation marks and receives {q} pseudo-tags, you should delete the {q} tags; if it is highlighted (italics, bold, etc.) it will automatically receive a <hi>-tag that should be kept.

Tagged elements	Corrected
In most of the grammars, this phenomenon is also known as {q}“rankshifting”/q}	In most of the grammars, this phenomenon is also known as “rankshifting”.

Table 16. Correcting {q} pseudo-tags

The tag <mentioned> will probably be mostly used in student texts in linguistics and language studies. It should be used to tag sentences or parts of sentences that were not produced by the students but which the students analyze or use to illustrate an argument, a definition, or an explanation.

Original elements	Tagged elements
The verbs, adjectives and adverbs used provide more insight into the way the effects of this crisis are described in newspapers: <ul style="list-style-type: none"> <li>- The economic crisis hits so hard.</li> <li>- The economic crisis will constrain all initiatives.</li> <li>- Thousands of women and children are dying as a direct consequence of the current economic crisis.</li> </ul>	The verbs, adjectives and adverbs used provide more insight into the way the effects of this crisis are described in newspapers: ¶ <ul style="list-style-type: none"> <li>- {start:vespa_mentioned}The economic crisis hits so hard {end:vespa_mentioned}. ¶</li> <li>- {start:vespa_mentioned}The economic crisis will constrain all initiatives {end:vespa_mentioned}. ¶</li> <li>- {start:vespa_mentioned}Thousands of women and children are dying as a direct consequence of the current economic crisis {end:vespa_mentioned}. ¶</li> </ul>

Table 17. Tagged linguistic examples

### Optional use of the <mentioned> tag:

The <mentioned> tag can also be used to tag words and phrases that are not used in their 'usual' senses but are referred to in student papers that talk about or describe language.

We greatly encourage VESPA partners to tag this type of mentioned words and phrases if they analyze texts produced in linguistics or language studies. However, as this will require much time and effort, we leave it up to them to decide whether they want to tag what could be referred to as "metalinguistic" use of words and phrases.

Original elements	Tagged elements
The connector however is more frequent than nevertheless and nonetheless in academic writing.	The connector {start:vespa_mentioned}however {end:vespa_mentioned} is more frequent than {start:vespa_mentioned}nevertheless {end:vespa_mentioned} and {start:vespa_mentioned}nonetheless {end:vespa_mentioned} in academic writing.
In this paper, I analyze the phraseology of the high-frequency verbs give and take.	In this paper, I analyze the phraseology of the high-frequency verbs {start:vespa_mentioned}give {end:vespa_mentioned} and {start:vespa_mentioned}take {end:vespa_mentioned}.
The adverb however is overused by EFL learners.	The adverb {start:vespa_mentioned}however {end:vespa_mentioned} is overused by EFL learners.
The verbs, adjectives and adverbs used around 'economic crisis' provide more insight into the way the effects of this crisis are described in newspapers.	The verbs, adjectives and adverbs used around {start:vespa_mentioned}economic crisis {end:vespa_mentioned} provide more insight into the way the effects of this crisis are described in newspapers.

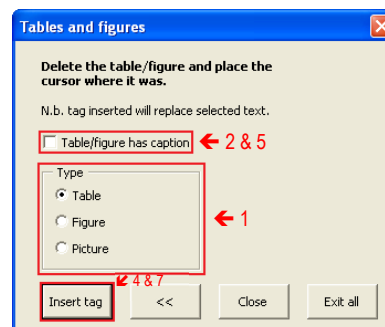
Table 18. The use of the <mentioned> tag to identify words that are referred to by students

### → Step 7: when all elements have been tagged:

- 7a: click on << to return to the previous window;
- or 7b: click on **Close** to move to the next selected feature (if no further option has been selected, the macro dialog box will be closed);
- or 7c: click on **Exit all** to close the macro dialog box.

### 2.3.7. Tables and figures

- Step 1: under *Type*, select either **Table**, **Figure** or **Picture**;
- Step 2: check/uncheck **Table/figure has caption** as necessary;
- Step 3: select the table, figure or picture to tag in the document;
- Step 4: click on **Insert tag**. The selected element will be replaced by the {emptyTag:vespa\_table}, {emptyTag:vespa\_figure} or {emptyTag:vespa\_figure-pic} pseudo-tags (see Table 15.);
- Step 5 (optional): check **Table/figure has caption** if you want to tag a table/figure caption
- Step 6 (optional): select the caption;





→ Step 7 (optional): click on **Insert tag**.

Repeat if required.

→ Step 8:

- 8a: click on << to return to the previous window;
- or 8b: click on **Close** to move to the next selected feature (if no further option has been selected, the macro dialog box will be closed);
- or 8c: click on **Exit all** to close the macro dialog box.

Original elements				Tagged elements	
Table 1.				{start:vespa_table}¶	
Type	of	Nb	Percentage	Table 1.¶	
suffixed words (types)		138	14.5%	{end:vespa_table}¶	
suffixed words (tokens)		208	21.9%		
words without a suffix		740	78%		
Total nr of words in the text		948	100%		
Table 1.				{emptyTag:vespa_table}¶	
Type	of	Nb	Percentage		
suffixed words (types)		138	14.5%		
suffixed words (tokens)		208	21.9%		
words without a suffix		740	78%		
Total nr of words in the text		948	100%		

Table 19. Tagged table with and without caption

A picture is treated as a figure (figure\_pic). When tagging a picture, the first step is to delete the picture from the file, then follow the steps above (ignoring step 3).

### 2.3.8. Formulae

A formula is a series of numbers or letters that represent a mathematical or scientific rule (example 1).

Example of formula
According to Halliday (1965: 35), the sentence “I told you that I thought that he was not there” could be represented by the following formulae: $\alpha^{\beta^{\gamma}}$ .

Formulae include algebraic expressions, logical expressions, chemical formulae, computer code, phonetic transcriptions, etc. and should be distinguished from abbreviations such as CHO (‘carbohydrate’) in the following example.

Example of a “false” formula
Sugar is represented by the simple formula CHO.

Note that formulae will be deleted from the text and only the tag will be left.

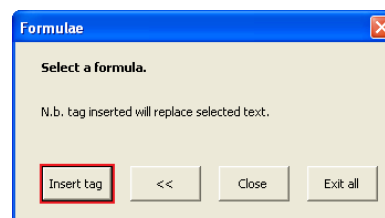
→ Step 1: select the formula

→ Step 2: click on **Insert tag**.

Repeat if required;

→ Step 3:

- 3a: click on << to return to the previous window;
- or 3b: click on **Close** or on **Exit all** to close the macro dialog box.



Original elements	Tagged elements
According to Halliday (1965: 35), the sentence “I told you that I thought that he was not there” could be represented by the following formulae: $\alpha^{\beta^{\gamma}}$ .	According to Halliday (1965: 35), the sentence {q}“I told you that I thought that he was not there”{q} could be represented by the following formulae: {emptyTag:vespa formula}¶

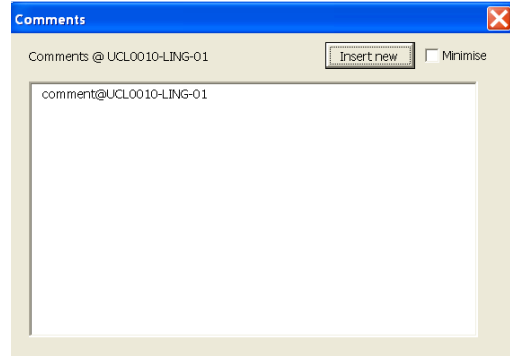
Table 20. Tagged formula

Take into account that a formula **cannot be separated from the main text** by indentation or a new line character

### 2.3.9. Comments

During the tagging process, it is possible to write comments in a separate window which is **automatically** activated by the macro:

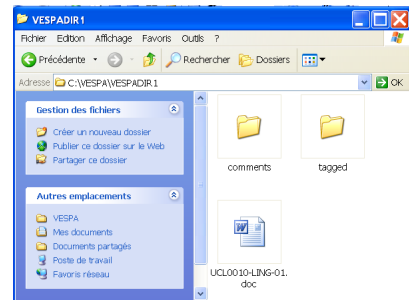
- Step 1: open the *Comments* window by un-ticking the '**Minimise**' button;
- Step 2: click on **Insert new** and write a comment. Repeat this step for each new entry;
- Step 3: tick **Minimise** to hide (or show) comments.



A comment may be something like “the student uses italics to indicate quotes throughout the text”.

### 2.4. Saving, closing and modifying the document

- At the end of the tagging process (**when all VESPA windows are closed**), the tagged document and any additional comments are automatically saved. For example:
  - The **original** (non-tagged) document is **UCL0010-LING-01.doc**.
  - The **tagged** folder (automatically created) contains the **saved tagged** document **UCL0010-LING-01.tagged.doc**;
  - The **comments** folder contains the **additional comments** saved in the document **UCL0010-LING-01.comments.txt**;

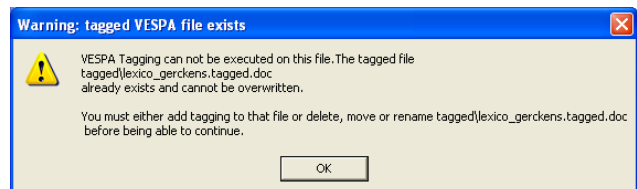


**File names:** It is very important that you do not change the names of the original doc file and the different files that are created by the VESPA macros.

- To resume tagging (or modify what you've previously done), open the saved tagged document in the **tagged** folder.

- As long as there is a tagged file with the same name in the **tagged** folder, it is not possible to tag the original document again.

The tagged file (e.g. *lexico\_gerckens.tagged.doc*) must be moved, renamed or deleted from the **tagged** folder if the original document (*lexico\_gerckens.doc*) has to be tagged again.



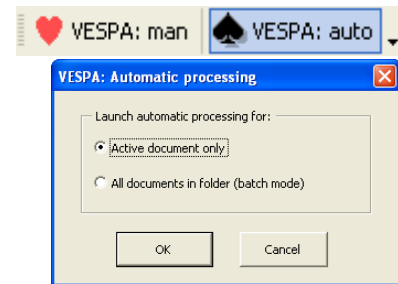
### 3. Using the VESPA macro: “Automatic tagging”

#### 3.1. Run the macro

→ Step 1: click on *Add-Ins* and on ♠ **VESPA: auto** to run the macro.  
!!! !!! The macro does not work if the *Find and Replace* window is active.

→ Step 2: select **Active document only** to tag the active document;

→ Step 3: click on **OK**.

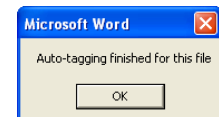


#### 3.2. Start automatic tagging

The macro will run automatically. No further step is required, except, potentially, with a **non-English** version of *Microsoft Word* (cf. 1.5).

#### 3.3. Saving and closing the document

→ A success message is displayed at the end of the tagging process;



→ The tagged xml document is automatically saved in a folder named **tagged2**.

## 4. Post-processing: Perl script

When a batch of VESPA texts has been tagged with the 2 Word macros the files have to go through a post-processing stage (by running a Perl script).

The Perl script is used to finalize the formatting process:

- Implement TEI XML structure
- Normalize hyphens, dashes, quotes, etc.
- Transform the pseudo-tags that were inserted as character sequences in the Word files into genuine XML TEI-conformant tags.
- Import the contextual information (i.e. learner profiles) from external spreadsheets
- Import the comments created in a separate file during the manual tagging
- Mark-up sentence and paragraph boundaries
- Mark-up footnotes and endnotes
- Numbering of <s> and <p>

The resulting document is then automatically checked for validity against the TEI Vespa.dtd by a parser.

The output of the Perl script is a batch of xml files.

Partners should send a copy of all xml files to the VESPA project co-ordinator (magali.paquot@uclouvain.be).

### 4.1. Installing Perl

It is recommended to install and use *ActivePerl* :

→ Step 1: download and install the **community Edition of ActivePerl** (available at: <http://www.activestate.com/activeperl>). Simply follow the instructions of the Setup program. Preferably do not change the program directory;

→ Step 2: install **AOfP.exe**. Use the directory where Perl is installed (typically **C:\Perl**).

### 4.2. Converting the Excel database

If all learner profiles were recorded in an Excel file, it is necessary to convert it into a text file in order to use it with the *Perl* script:

→ Step 1: open the *Excel* file used to collect students' answers;

→ Step 2: click on the **Office button** (or the **File** tab) and on **Save as**;

→ Step 3: Type **CTXDATA.txt** as *File Name*;

→ Step 4: Select **Text (Tab delimited) (\*.txt)** as *Type* and click on **Save**.

Every time the *Excel* file is updated, a new file conversion (to .txt) is necessary before running the *Perl* script.

### 4.3. Setting up the Perl script

Copy the following two files in the directory containing the tagged2 folder:

- **makeVespaXml.v1.7.pl**
- **CTXDATA.txt**

#### 4.4. Using the Perl script

→ Step 1: click on the Windows **Start** menu (or on the **Windows button**), then on **All Programs, Accessories** and on **Command Prompt**;

→ Step 2: type the following command: **cd [directory containing makeVespaXml.v1.7.pl]** and confirm with the **Enter** key. For example, if *makeVespaXml.v1.7.pl* is found in C:\VESPA:

```
Microsoft Windows [version 6.2.9200]
(c) 2012 Microsoft Corporation. All rights reserved.

C:\Users\Default user>cd C:\VESPA
```

The modified path should be displayed on the next line:

```
Microsoft Windows [version 6.2.9200]
(c) 2012 Microsoft Corporation. All rights reserved.

C:\Users\Default user>cd C:\VESPA

C:\VESPA>
```

→ Step 3: choose one method in order to run the Perl script:

- 3a: type the following command: **makeVespaXml.v1.7.pl** and confirm with the **Enter** key if you want to convert all xml files found in *tagged2*:

```
C:\VESPA>makeVespaXml.v1.7.pl
```

- 3b: type the following command: **makeVespaXml.v1.7.pl [file name]** and confirm with the **Enter** key if you want to convert xml files from *tagged2* containing a specific sequence in their name. For example:

```
C:\VESPA>makeVespaXml.v1.7.pl UCL
* only files containing the sequence UCL will be tagged;
```

```
C:\VESPA>makeVespaXml.v1.7.pl UCL0001-LING-01.tagged2.xml
* only the xml file named UCL0001-LING-01.tagged2.xml will be tagged.
```

The second method is preferable if your computer is not very powerful as it allows converting a limited number of files. It is also very useful when the script throws an error for a particular file.

A new folder named *tagged3* with the new xml files will be automatically generated. No success message will be displayed at the end of the process, but seeing lines starting with **XML::Checker INFO-300**: means that everything is ok:

```
UCL0002-LING-04.tagged2.xml

ent left in txt:      &amp;

parse file: [./tagged3/UCL0002-LING-04.tagged3.xml]
XML::Checker INFO-300: [2] references to ID [foreign]
XML::Checker INFO-300: [0] references to ID [UCL0002-LING-04]
XML::Checker INFO-300: [1] references to ID [vespa_UCL0002-LING-04-ftnote.001]
XML::Checker INFO-300: [4] references to ID [English]
```

## 4.5. Troubleshooting

To run properly, the Perl script requires that the manual tagging be done 100% correctly. Tagging mistakes will create problems and the Perl script will either stop running or generate error warnings accordingly.

This section lists issues or bugs you may encounter when using the *Perl* script. Finding and solving them can sometimes be time-consuming, depending on their type.

### 4.5.1. Frequent tagging mistakes

#### 4.5.1.1. Mismatched tags

The *Perl* script will stop running and one of the following messages will be displayed:

```
mismatched tag at line 1, column 17055, byte 17055 at C:/Perl/lib/XML/Parser.pm line 187
```

→ The *line*, *column* and *byte* number are not always 100% correct – you may need to look around to identify the tagging mistake (see 4.5.2 for more detail on the debugging procedure).

```
too many p closed: pOp=-1 at C:\VESPA\makeVespaXml.v1.7.pl line 906, <IN> line 3.
```

→ In that case, the *line* number should be correct.

A number of tagging mistakes may generate errors:

- The opening and closing tags are different:

Incorrect elements	Corrected elements
<code>{start:vespa-italic} {start:vespa_mentioned-for} sorte de {end:vespa-italic} {start:vespa-italic} {end:vespa_mentioned-for} {end:vespa-italic} can be found in all NS and NNS corpora.</code>	<code>{start:vespa-italic} {start:vespa_mentioned-for} sorte de {end:vespa_mentioned-for} {end:vespa-italic} can be found in all NS and NNS corpora.</code>

Table 21. Different opening and closing tags

Elements may have nested elements but nested elements should always be closed before the “mother” element is closed. In Table 21, the nested element `{start:vespa_mentioned-for}` must be closed before the italic element is closed.

- There is a missing tag

Incorrect elements	Corrected elements
<code>{start:vespa_mentioned} {start:vespa-italic} Sandwich {end:vespa-italic} (originally the title of the Earl of Sandwich) could be a relevant word.</code>	<code>{start:vespa_mentioned} {start:vespa-italic} Sandwich {end:vespa-italic} {end:vespa_mentioned} (originally the title of the Earl of Sandwich) could be a relevant word.</code>

Table 22. Missing tags

The opening tag `{start:vespa_mentioned}` can be found at the beginning, but not the closing tag `{end:vespa_mentioned}`.

- There is an incorrect tag:

Incorrect elements	Corrected elements
<code>{start:vespa_table} {start:vespa-underlined} Table 3 {end:vespa-underlined}: The major indefinite pronouns according to Quirk (1985:377) {start:vespa_table}</code>	<code>{start:vespa_table} {start:vespa-underlined} Table 3 {end:vespa-underlined}: The major indefinite pronouns according to Quirk (1985:377) {end:vespa_table}</code>

Table 23. Repeated tags

The opening tag `{start:vespa_table}` is correct, but not the closing tag `{start:vespa_table}`. `{end:vespa_table}` is required.

- Some tags cannot be attached to other tags and must stand on their own after a new line character:

Incorrect elements	Corrected elements
<pre>{start:vespa_introduction-1}¶ {start:vespa-underlined}{start:vespa- italic}Introduction{end:vespa-italic}{end:vespa- underlined}{end:vespa_introduction-1}¶</pre>	<pre>{start:vespa_introduction-1} {start:vespa-underlined}{start:vespa- italic}Introduction{end:vespa-italic}{end:vespa-underlined}¶ {end:vespa_introduction-1}¶</pre>

Table 24. Special tags

The closing tags {end:vespa-italic} {end:vespa-underlined} are correctly used, but {end:vespa\_introduction-1} is a division tag and cannot be found next to those.

- There is an error in the tag:

Incorrect elements	Corrected elements
<pre>start:vespa_section-1} {start:vespa-bold}II. Theoretical part{end:vespa-bold} end:vespa_section-1}</pre>	<pre>{start:vespa_section-1} {start:vespa-bold}II. Theoretical part{end:vespa-bold} {end:vespa_section-1}</pre>

Table 25. Incorrect tags

In the above example, there is a missing { in the end tag.

#### 4.5.1.2. Unexpected elements

Some tags cannot be used within other tags (cf. the 'mentioned' tag, the 'formula' tag). Others require some elements in their context. If an unexpected element is found, a warning is shown but the Perl script should continue running:

A number of unexpected elements may generate warnings:

- A 'mentioned' pseudo-tag is found within another element which does not require it:

```
XML::Parser parsing ERROR
validation error: XML::Checker ERROR-157: unexpected Element [mentioned]
Context: ChildElementIndex 0, line 3, column 30821, byte 30909
```

→ The *line*, *column* and *byte* number are not always 100% correct – you may need to look around to identify the tagging mistake.

Incorrect elements	Corrected elements
<pre>{start:vespa_documentTitle} Project in Corpus Linguistics: The use of {start:vespa_mentioned} indeed{end:vespa_mentioned} by French learners of English compared to natives. {end:vespa_documentTitle}</pre>	<pre>{start:vespa_documentTitle} Project in Corpus Linguistics: The use of indeed by French learners of English compared to natives. {end:vespa_documentTitle}</pre>

Table 26. Mentioned tags as unexpected elements

- A 'formula' pseudo-tag is found on a line of its own:

```
XML::Parser parsing ERROR
validation error: XML::Checker ERROR-157: unexpected Element [formula]
Context: ChildElementIndex 0, line 3, column 64022, byte 64110
```

→ The *line*, *column* and *byte* number are not always 100% correct – you may need to look around to identify the tagging mistake.

Incorrect elements	Corrected elements
<pre>Formation: ¶ {emptyTag:vespa_formula}¶ Back-formation:¶ {emptyTag:vespa_formula}¶</pre>	<pre>Formation: {emptyTag:vespa_formula}¶ Back-formation: {emptyTag:vespa_formula}¶</pre>

Table 27. A formula tag is used as a paragraph

{emptyTag:vespa\_formula} has to be found next to a text and cannot be used alone.

```
A {q} pseudo-tag is found on its own:UCL0045-LING-01.tagged2.xml
Illegal content of root: not an element at C:\... \makeVespaXml.v1.7.pl line 1380, <IN> line
16
```

→ The *line*, *column* and *byte* number are not always 100% correct – you may need to look around to identify the tagging mistake.

A pseudo-tag {q} cannot be used on its own. The error is frequent and is often due to an apostrophe having been automatically tagged as a {q} (see 2.4.1.1.1).

#### 4.5.1.3. CTXDATA-generated errors

The following message may appear when running the Perl script:

```
ERROR IN CONTEXTUAL_DATA SHEET: no entry for text id=[UCL0002-LING-04]
```

The *Perl* script will continue running. The error warning means that the learner profile for the processed text could not be found in CTXDATA. You should check the CTXDATA file and make sure the learner profile is there (do not forget to update the CTXDATA file when you process new files).

#### 4.5.2. Debug process

Debugging is a very tedious process because it is not easy to identify the error. This is why you should be particularly careful when manually tagging your files.

To correct the bugs, VESPA partners have adopted various solutions.

- Step 1: identify the error message in the *Perl* script window;
- Step 2: open the problematic file in the *tagged2* directory (open in Word even though the file extension is xml);
- Step 3: manually check the tagged file (opened in Word) and correct errors
- Step 4: run auto-tag again
- Step 5: run the *Perl* script, using the following command: `makeVespaXml.v1.7.pl [file name]`, until Perl is “happy”

OR

- Step 1: identify the error message in the *Perl* script window;
- Step 2: open the problematic file in the *tagged2* directory (open in Word even though the file extension is xml);
- Step 3: delete a first page/section of the document and save the changes;
- Step 4: run the *Perl* script, using the following command: **`makeVespaXml.v1.7.pl [file name]`**:
  - 4a: repeat steps 1 to 4 if the same message appears;
  - 4b: if the conversion is successful, it means that you have isolated the page/section with the error. Cancel the *last* deletion and check the text manually (do not correct it yet). This search zone can also be reduced by repeating steps 1 to 4, but it is recommended not to make it too small;
- Step 5: when the error is found (remember where it is), cancel all deletions to restore the original document, go back to the problematic section, correct the mistake and save the changes;
- Step 6: run the *Perl* script, using the same command in step 4:
  - 6a: if no error message is displayed, the file has been successfully debugged. No further steps are required;
  - 6b: if another error message is displayed, the error has not been corrected properly or a new problem has been found in the file. Start the whole debug process again. Note that the same type of error can occur several times in a document;



Please get back to us if you have a better solution!

## **References**

- Ebeling, S.O. & Heuboeck, A. (2007). Encoding document information in a corpus of student writing: the British Academic Written English Corpus. *Corpora* 2(2): 241-256.
- Heuboeck, A., Holmes, J. & Nesi, H. (2008). *The BAWE Corpus Manual*.  
[http://www.reading.ac.uk/AcaDepts/ll/app\\_ling/internal/bawe/BAWE.documentation.pdf](http://www.reading.ac.uk/AcaDepts/ll/app_ling/internal/bawe/BAWE.documentation.pdf)

**Appendix 1: VESPA learner profile**

## VESPA Learner profile

Text code: (do not fill in)

### GENERAL INFORMATION

Last name:	First name(s):
Year of birth: _ _ _ _	Country of birth:
First language:	

Language(s) spoken at home (if you speak more than one language at home, please list the one(s) you know best first)

- Language 1: \_\_\_\_\_
- Language 2: \_\_\_\_\_
- Language 3: \_\_\_\_\_

Other foreign language(s): please specify any foreign languages (other than English) that you speak, starting with the one(s) you know best

- Foreign language 1: \_\_\_\_\_
- Foreign language 2: \_\_\_\_\_
- Foreign language 3 : \_\_\_\_\_

### EDUCATION

Language of instruction<sup>1</sup>:

- In primary school<sup>2</sup> \_\_\_\_\_
- In secondary school<sup>3</sup> \_\_\_\_\_
- At university<sup>4</sup> (in your field of study) \_\_\_\_\_

Current field of study (e.g. English studies, business, law): \_\_\_\_\_

Name of the university where you are currently studying: \_\_\_\_\_

Number of years at university:

- |                            |                                    |
|----------------------------|------------------------------------|
| <input type="checkbox"/> 1 | <input type="checkbox"/> 4         |
| <input type="checkbox"/> 2 | <input type="checkbox"/> 5         |
| <input type="checkbox"/> 3 | <input type="checkbox"/> 6 or more |

<sup>1</sup> Language used by the teacher(s) for most courses.

<sup>2</sup> A school for children (often between the ages of 5 and 11)

<sup>3</sup> A school for young people (often between 11 and 16 or 18)

<sup>4</sup> Please think of all the languages that are used, e.g. if you study modern languages and some classes are taught in English, others in Spanish and more in French, you should choose answer 3: English and 2 or more other languages + list Spanish and French in the box on the right.

Level of study:

- Bachelor's                       Master's                       PhD

English courses

- Years of English at school (before university): \_\_\_\_\_

- Years of English at university: \_\_\_\_\_

Have you spent some time in an English-speaking country?

- No     4-6 months  
 2-4 weeks                                       7-12 months  
 1-3 months                                       more than one year

## TEXT

Title : \_\_\_\_\_

Name of course for which you wrote the text: \_\_\_\_\_

Type of text:

- term paper                       research report                       Master's dissertation  
 Other (please specify) : \_\_\_\_\_

Did you write the text in the classroom?

- Yes     No

Is this text part of an examination (that is, will your grade be based on it fully or in part)?

- Yes     No

Were you allowed to use reference tools (dictionaries or others) to write the text?

- Yes     No

If yes, please specify what reference tools: \_\_\_\_\_

- dictionary                       grammar  
 scientific articles                       Other (please specify) : \_\_\_\_\_

Did you use your computer's grammar and spell-checker?

- Yes     No

Thank you for taking the time to fill in this questionnaire.

---

I hereby give permission for my essay to be used for research and teaching purposes.

Date :

Signature :

## **Appendix 2: The subset of TEI (P5) used in VESPA**

<back>, <body>, <cell>, <distributor>, <div1>, <div2>, <div3>, <div4>, <div5>, <docTitle>, <encodingDesc>, <extent>, <figure>, <fileDesc>, <formula>, <front>, <head>, <hi>, <item>, <list>, <mentioned>, <name>, <note>, <notesStmt>, <p>, <particDesc>, <person>, <profileDesc>, <publicationStmt>, <q>, <quote>, <row>, <s>, <sourceDesc>, <table>, <TEI>, <teiHeader>, <text>, <title>, <titlePage>, <titlePart>, <titleStmt>

See <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html> for more details on each element