Data capture
and corpus markup

Kron

Data to be collected

- Like other decisions in corpus creation (e.g. balance, representativeness, size), the kind of data to be collected also depends on your research questions
  - If you wish to compare British English and American English, you will need to collect spoken and / or written data produced by native speakers of the two regional varieties of English
  - If you are interested in how Chinese speakers acquire English as a second language, you will then need to collect the English data produced by Chinese learners to create a learner corpus
  - If you are interested in how the English language has evolved over centuries, you will need to collect samples of English produced in different historical periods to build a historical or diachronic corpus

Data capture

- Having developed an understanding of the type of data you need to collect, and having made sure that no ready-made corpus of such material exists, you’ll need to capture the data
- Data digitalisation
  - Machine-readability is a *de facto* feature of a modern corpus

Data capture

- Text must be rendered machine-readable
  - Keyboarding
  - OCR (Optical Character Recognition) scanning
  - Transcribing audio/video recording
- Existing electronic data is preferred over paper-based materials
  - The Web as an important source of machine-readable data for many languages
  - Converting other file format such as HTML, Word, PDF into plain text format
- The World-Wide-Web (WWW) is an important source of electronic text archives

Some useful data source

- Oxford Text Archive
  - http://ota.ahds.ac.uk/
  - Oldest text archive - thousands of texts (and many well-known corpora) in more than 25 different languages
- Project Gutenberg
  - http://www.gutenberg.org/catalog/
  - First producer of free electronic books – 2,8000 e-books!
- Digital collections of university libraries e.g.
  - http://www.digitalcurationservices.org/digital-stewardship-services/etext-projects/
  - http://onlinebooks.library.upenn.edu/
- Corpus4u electronic text archives
  - http://www.corpus4u.org/forumdisplay.php?f=21

Copyright in corpus creation

- A corpus consisting entirely of copyright-free old texts is not useful in study of contemporary language
- Copyright is a major issue in data collection if you are **to publish or make your corpus publicly available**
- The samples taken under the convention of ‘fair dealing’ in copyright law are so small as to jeopardize any claim of balance or representativeness
- There is as yet no satisfactory solution to the issue of copyright in corpus

Copyright in corpus creation

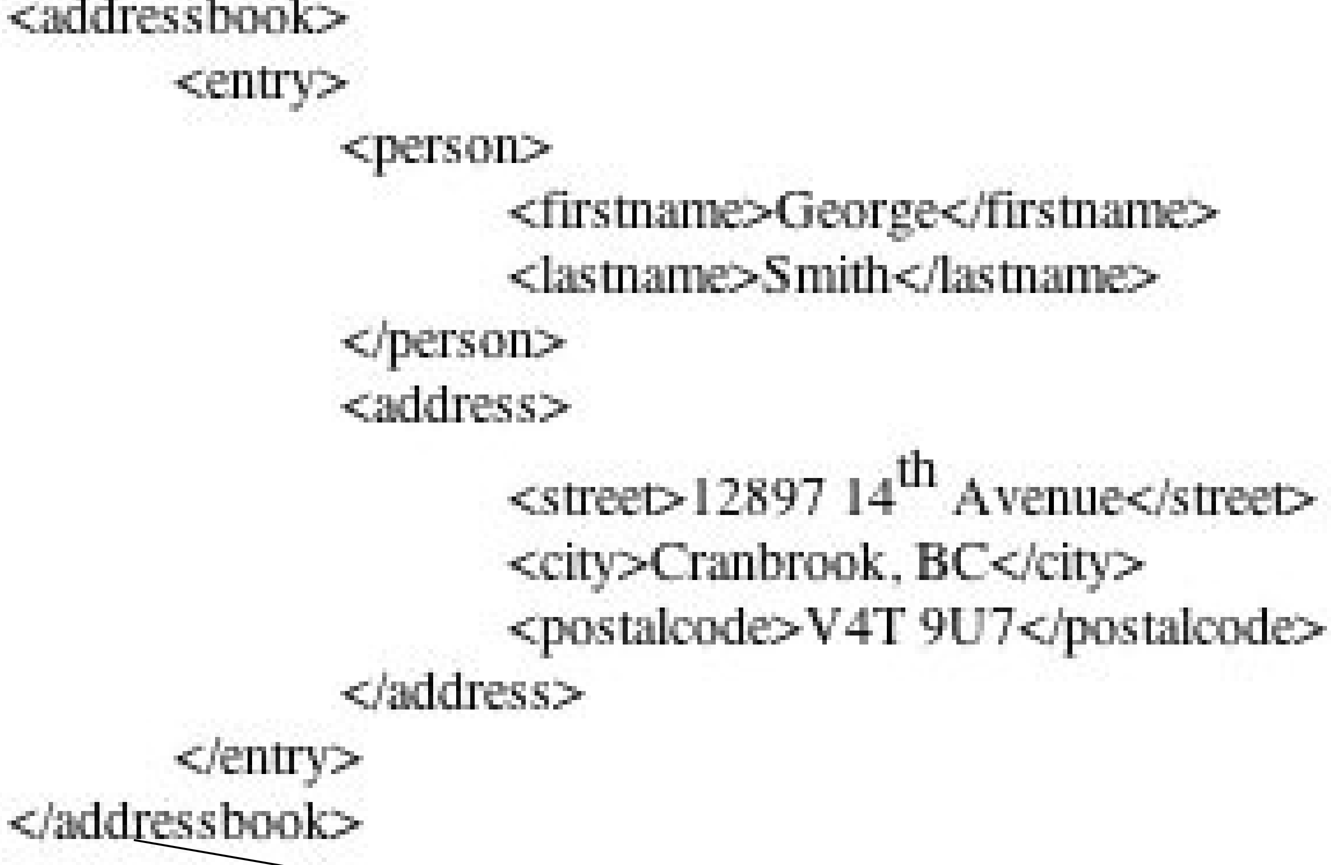
- Tips for copyright issues
  - Usually easier to obtain permission for samples than for full texts
  - Easier for smaller samples than for larger ones
  - If you show that you are acting in good faith, and only small samples will be used in non-profit-making research, copyright holders are typically pleased to grant you permission
  - You don’t need to worry about copyright if you build a corpus **for your private use!**

Corpus markup

- System of standard codes inserted into a document stored in electronic form to provide information **about** the text itself and govern formatting, printing and other processes
  - Describing the document (“metadata” like source, name, author, date, etc)
  - Marking boundaries for paragraphs, sentences, and words, omissions etc
  - Displaying markup (font, font size, positioning)

Example of markup

start tag



```
<addressbook>
        <entry>
                <person>
                        <firstname>George</firstname>
                        <lastname>Smith</lastname>
                </person>
                <address>
                        <street>12897 14th Avenue</street>
                        <city>Cranbrook, BC</city>
                        <postalcode>V4T 9U7</postalcode>
                </address>
        </entry>
</addressbook>
```

end tag

Why markup?

- Markup recovers contextual information of sampled texts which are taken out of context
- Markup allows for a broader range of research questions to be addressed by providing extra information such as text types, sociolinguistic variables, structural organization
- Markup allows corpus builders to insert editorial comments during the corpus building process
- Pre-processing written texts (e.g. tables and graphs), and particularly transcribing spoken data, also involves markup (e.g. pause, paralinguistic features etc)

Markup schemes

- The extra markup information must be kept separate from the textual data in a corpus
- Markup schemes
  - COCOA
  - TEI (Text Encoding Initiative)
  - CES (Corpus Encoding Standard)

COCOA reference

- One of the earliest markup schemes
- Consisting of a set of attribute names and values enclosed in angled brackets
  - e.g. <A WILLIAM SHAKESPEARE>
    - *attribute name* = A (author)
    - *attribute value* = WILLIAM SHAKESPEARE
- Only encoding a limited set of features such as authors, titles and dates
- Giving way to more modern schemes

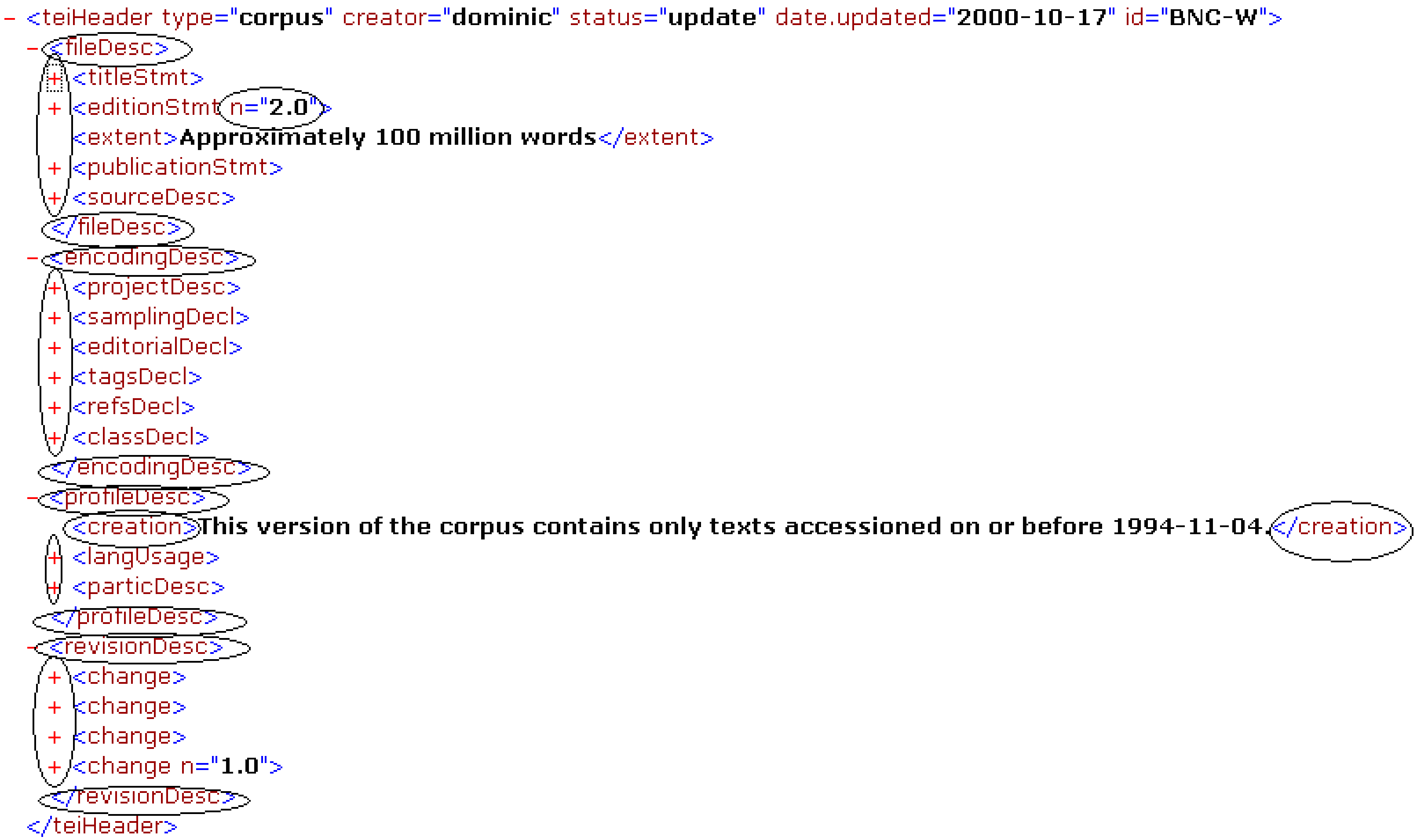
TEI guidelines

- The Text encoding Initiative: sponsored by three major academic associations concerned with humanities computing
  - The Association for Computational Linguistics (ACL)
  - The Association for Literary and Linguistic Computing (ALLC)
  - The Association for Computers and the Humanities (ACH)
- Aiming to facilitate data exchange by standardizing the markup or encoding of information stored in electronic form

TEI guidelines

- Each individual text is a ***document*** consisting in a ***header*** and a ***body***, which are in turn composed of different ***elements***
- TEI corpus header has 4 principal elements
  - A *file description* (<fileDesc>): a full bibliographic description
  - An *encoding description* (<encodingDesc>): relationship between an electronic text and its source or sources (e.g. spelling standardization)
  - A *text profile* (<profileDesc>): a detailed description of non-bibliographic aspects of a text
  - A *revision history* (<revisionDesc>): a record of changes to a file
- Only <fileDesc> is required to be TEI-compliant
  - The other three elements are optional
- Tags can be nested, i.e. an element can appear inside another element

The BNC header



```
- <teiHeader type="corpus" creator="dominic" status="update" date.updated="2000-10-17" id="BNC-W">
  - <fileDesc>
    + <titleStmt>
    + <editionStmt n="2.0">
      <extent>Approximately 100 million words</extent>
    + <publicationStmt>
    + <sourceDesc>
    </fileDesc>
  - <encodingDesc>
    + <projectDesc>
    + <samplingDecl>
    + <editorialDecl>
    + <tagsDecl>
    + <refsDecl>
    + <classDecl>
    </encodingDesc>
  - <profileDesc>
      <creation>This version of the corpus contains only texts accessioned on or before 1994-11-04.</creation>
    + <langUsage>
    + <particDesc>
    </profileDesc>
  - <revisionDesc>
    + <change>
    + <change>
    + <change>
    + <change n="1.0">
    </revisionDesc>
  </teiHeader>
```

TEI guidelines

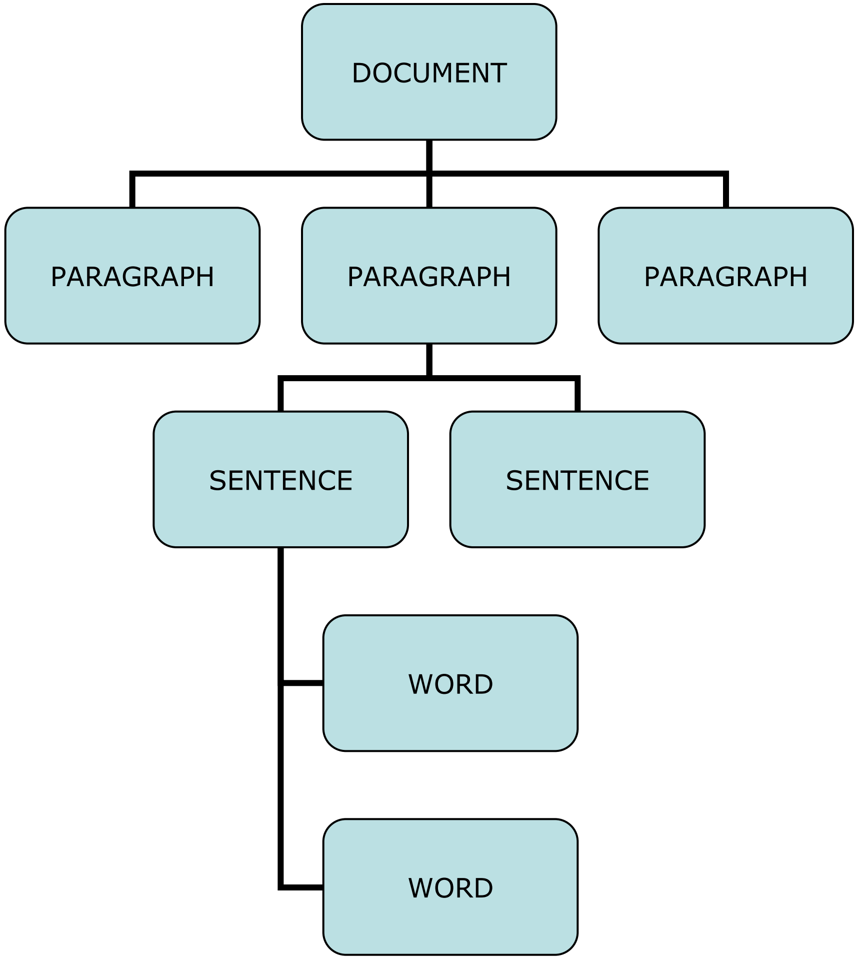
- Markup languages adopted by the TEI
  - SGML (**S**tandard **G**eneralized **M**arkup **L**anguage)
  - XML (e**X**tensible **M**arkup **L**anguage)
- Current version of TEI P5 guidelines (version 2.3.0, published in Jan 2013)
- See the TEI official website for latest updates
  - http://www.tei-c.org/index.xml

HTML, SGML, XML

- HTML (*__H__ypertext __M__arkup __L__anguage*) is based on SGML but with a predefined DTD (Document Type Definition)
  - HTML does not conform to all SGML rules (e.g. tags with no closing counterpart <p> versus <p>…</p>)
    - SGML: Standard Generalized Markup Language
- XML is a simplified subset of SGML intended to make SGML easy enough for use on the Web
  - eliminating some of the more complex DTD constructs
  - introducing **Unicode/multilingual** support
  - (introducing **data types** and **namespaces**)

XML Documents are trees

```
<DOCUMENT>
   <PARAGRAPH>
        <SENTENCE>
                <WORD>
                ...
                </WORD>
        </SENTENCE>
   ...
   </PARAGRAPH>
</DOCUMENT >
```



Metadata in XML

- What properties does a book have?
  - author, ISBN, publisher, number of pages, genre: fiction, etc

```
<BOOK type=“fiction”>
        <AUTHOR gender=“male”>John Smith</AUTHOR>
        <PUBLISHER>CUP</PUBLISHER>
        <TITLE>Lost in translation</TITLE>
        …
</BOOK>
```

- This contains “data” such as *John Smith, CUP, Lost in Translation…*
  - tags can have attributes (e.g. *gender* for author, *type* for book)

- It contains metadata (data about the data) in the form of tags

- Easy for a machine to know which pieces of information are about what

Corpus Encoding Standard (CES)

- Designed specifically for the encoding of language corpora
  - Document-wide mark-up
    - bibliographical description, encoding description, etc
  - Gross structural mark-up
    - volume, chapter, paragraph, footnotes, etc
    - specifying recommended character sets
  - Markup for sub-paragraph structures
    - sentence, quotations, words, MWUs, abbreviations, etc

Corpus Encoding Standard

- CES specifies a minimal encoding level that corpora must achieve to be considered as standardized in terms of descriptive representation as well as general architecture
- 3 levels of standardization designed to achieve the goal of universal document interchange
  - Metalanguage level regulates the form of the “syntactic” rules and the basic mechanisms of markup schemes (e.g. case sensitive, matching start/end tags)
  - Syntactic level specifies precise tag names and “syntactic” rules for using the tags
  - Semantic level ensures the same tag names are interpreted in the same way by the data sender and receiver e.g. <title> vs. <h.title>

Corpus Encoding Standard

- Like the TEI scheme, CES not only applies to corpus markup, it also covers encoding conventions for the linguistic annotation of text and speech
- Available in both SGML and XML
  - The expanded XML version is called XCES
- See the CES official website for latest updates
  - http://www.cs.vassar.edu/CES/

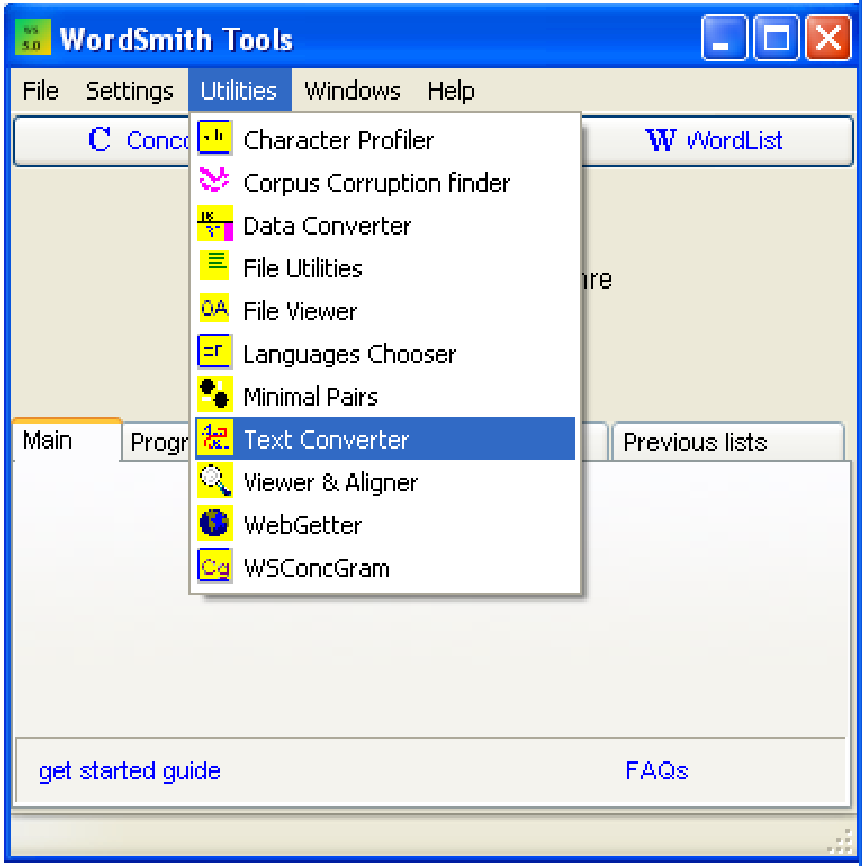
Character encoding

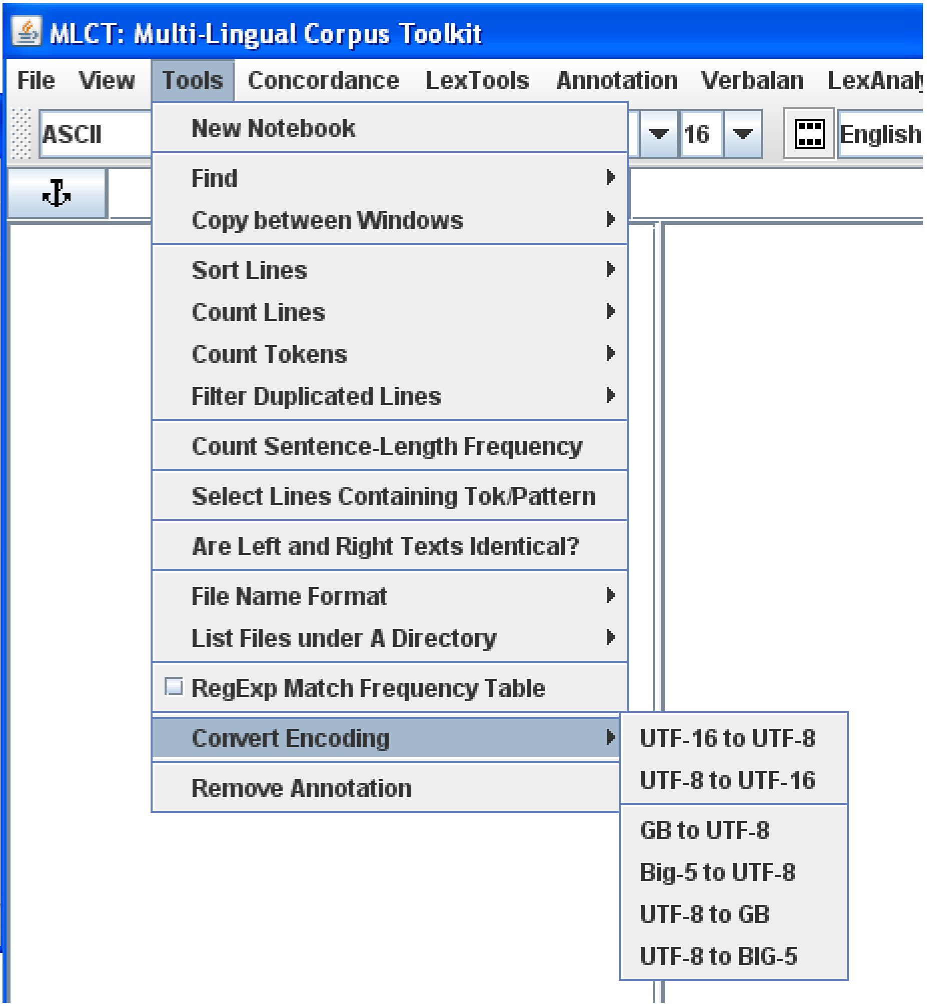
- Rarely an issue for English
  - ASCII (American Standard Code for Information Interchange) – “plain text” (ANSI: American National Standard Institute)
  - Special characters are exceptions, which are represented in SGML version of TEI and CES using *entity references* (included between ampersand and semi-colon)
    - £ = &pound;
    - é = &eacute;
- The ISO-8859 family of 15 members
  - Complementary standardized character codes
- Unicode (Unification Code)
  - Supported in XML
  - UTF-8 (8-bit Unicode transformation format)
  - UTF-16 (16-bit Unicode transformation format)
- See Unicode official website for latest updates
  - http://unicode.org/

Character encoding

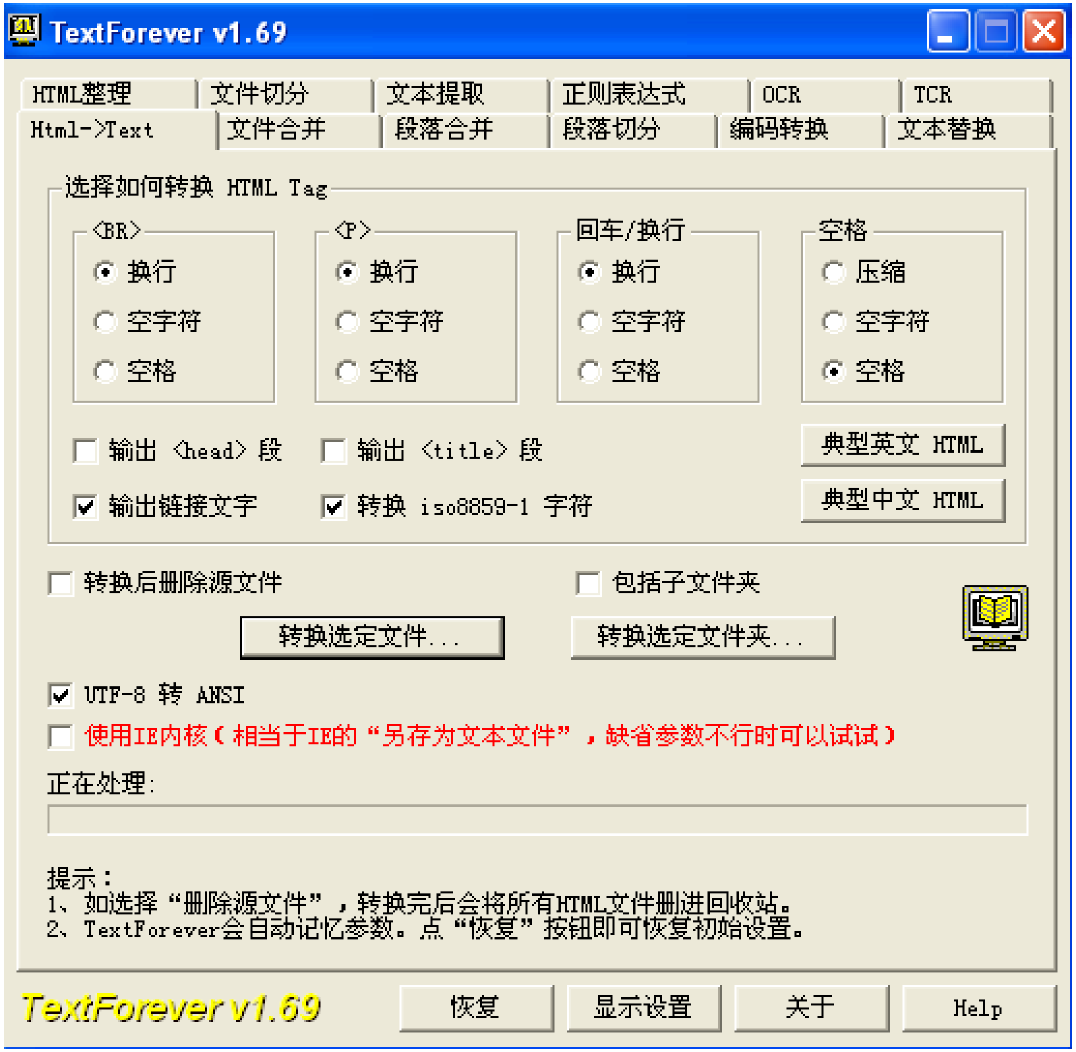
- ASCII (ANSI), GB2312, Big5, UTF8, Unicode (UTF16)
  - For more details see http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter4.htm
- WordSmith 5 is based on Unicode (16-bit)
  - Unless your corpus is all ASCII characters, WST may NOT produce reliable results unless it is converted into Unicode
  - WST Utilities – Text Converter
  - MLCT or Textforever.exe for conversion
- The combination of XML and Unicode is the current standards in corpus building (Xiao et al 2004)

Text conversion





Keep a safe copy of your text before you convert!



选择如何转换 HTML Tag

<BR>
● 换行
○ 空字符
○ 空格

<P>
● 换行
○ 空字符
○ 空格

回车/换行
● 换行
○ 空字符
○ 空格

空格
○ 压缩
○ 空字符
● 空格

□ 输出 <head> 段　□ 输出 <title> 段

☑ 输出链接文字　☑ 转换 iso8859-1 字符

□ 转换后删除源文件　□ 包括子文件夹

☑ UTF-8 转 ANSI

□ 使用IE内核（相当于IE的“另存为文本文件”，缺省参数不行时可以试试）

正在处理：

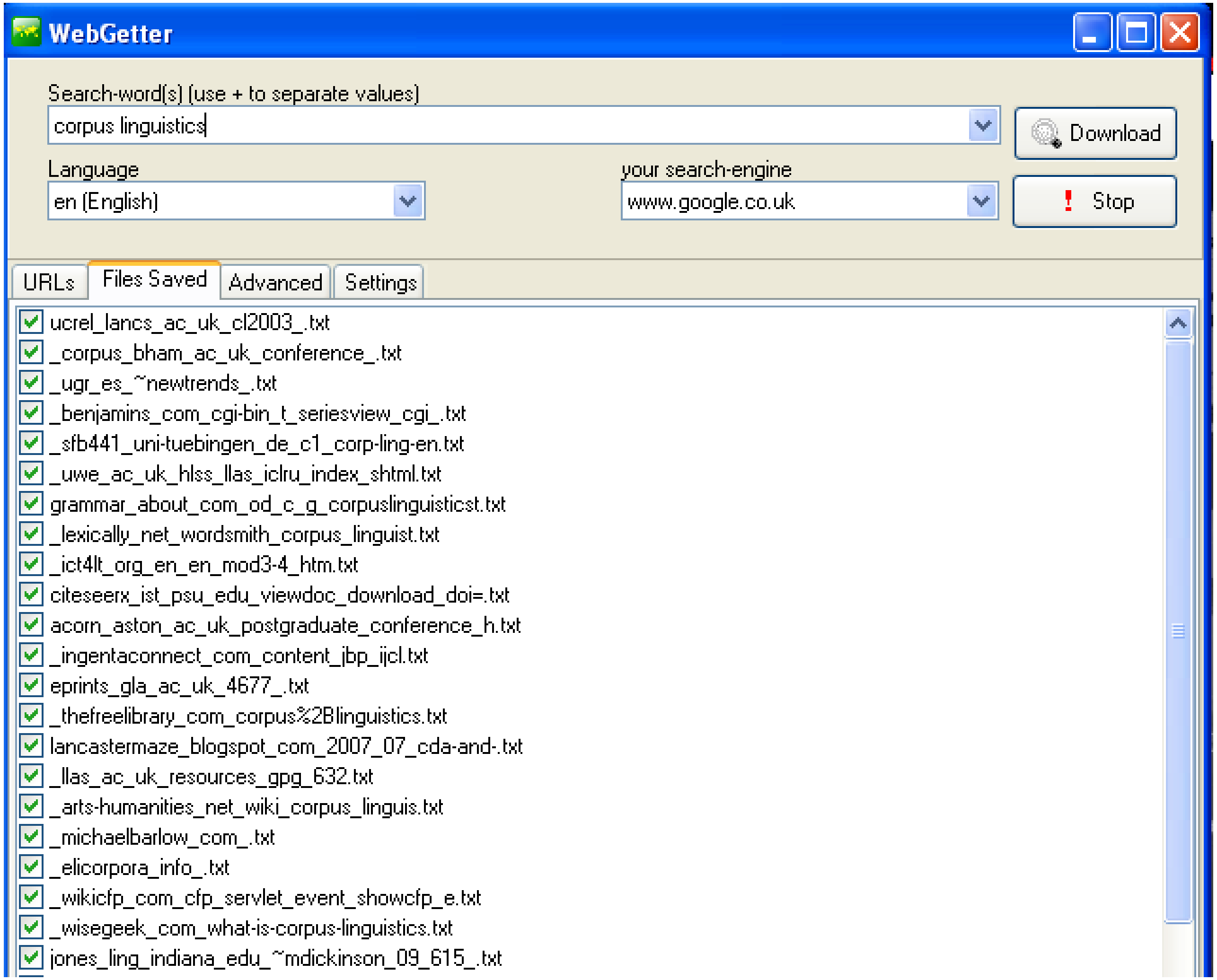
提示：
1、如选择“删除源文件”，转换完后会将所有HTML文件删进回收站。
2、TextForever会自动记忆参数。点“恢复”按钮即可恢复初始设置。

http://download.pchome.net/utility/file/editor/detail-83578.html

Data capture tools

- Freeware tools that help you to download all pages at a selected website at one go
  - Grab-a-Site
    - http://download.cnet.com/Grab-a-Site/3000-2646_4-68934.html
    - HTTrack
    - http://www.httrack.com/
- *Webgetter* in WST 4.0 or 5.0
  - WST menu – Utilities – WebGetter
  - Downloads all the pages containing the specified search word
  - But does not tidy up the data
- Multilingual Corpus Toolkit (MLCT)
  - http://www.ling.lancs.ac.uk/corplang/cbls/zipfiles/MLCT.zip
  - Can download, tidy up and POS tag the selected webpage
  - Can markup textual organization automatically (<p>, <s>)

WST WebGetter



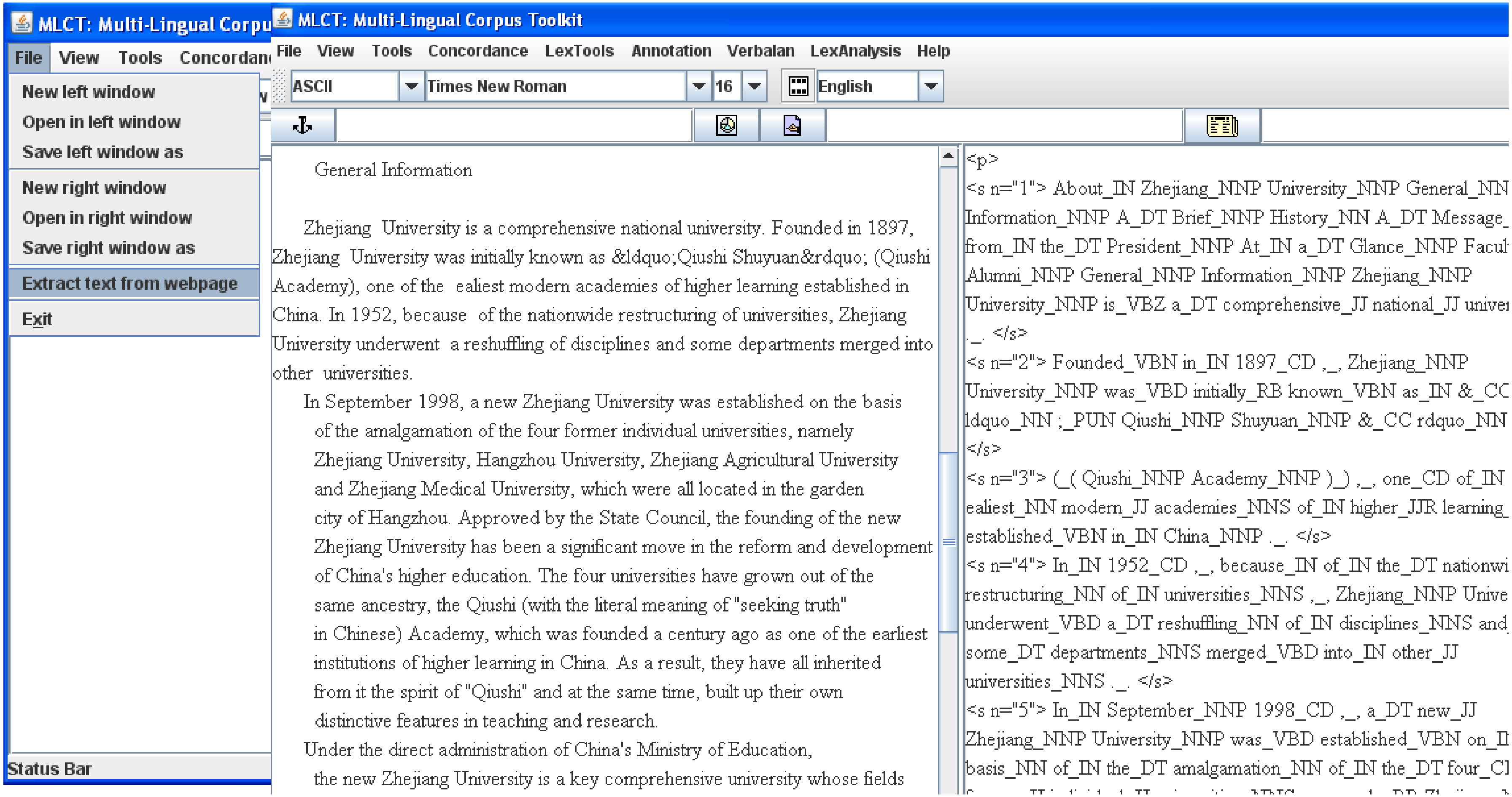
Search-word(s) (use + to separate values): corpus linguistics

Language: en (English)

your search-engine: www.google.co.uk

☑ ucrel_lancs_ac_uk_cl2003_.txt
☑ _corpus_bham_ac_uk_conference_.txt
☑ _ugr_es_~newtrends_.txt
☑ _benjamins_com_cgi-bin_t_seriesview_cgi_.txt
☑ _sfb441_uni-tuebingen_de_c1_corp-ling-en.txt
☑ _uwe_ac_uk_hlss_llas_iclru_index_shtml.txt
☑ grammar_about_com_od_c_g_corpuslinguisticst.txt
☑ _lexically_net_wordsmith_corpus_linguist.txt
☑ _ict4lt_org_en_en_mod3-4_htm.txt
☑ citeseerx_ist_psu_edu_viewdoc_download_doi=.txt
☑ acorn_aston_ac_uk_postgraduate_conference_h.txt
☑ _ingentaconnect_com_content_jbp_ijcl.txt
☑ eprints_gla_ac_uk_4677_.txt
☑ _thefreelibrary_com_corpus%2Blinguistics.txt
☑ lancastermaze_blogspot_com_2007_07_cda-and-.txt
☑ _llas_ac_uk_resources_gpg_632.txt
☑ _arts-humanities_net_wiki_corpus_linguis.txt
☑ _michaelbarlow_com_.txt
☑ _elicorpora_info_.txt
☑ _wikicfp_com_cfp_servlet_event_showcfp_e.txt
☑ _wisegeek_com_what-is-corpus-linguistics.txt
☑ jones_ling_indiana_edu_~mdickinson_09_615_.txt

Using MLCT to capture web text



General Information

Zhejiang University is a comprehensive national university. Founded in 1897, Zhejiang University was initially known as &ldquo;Qiushi Shuyuan&rdquo; (Qiushi Academy), one of the ealiest modern academies of higher learning established in China. In 1952, because of the nationwide restructuring of universities, Zhejiang University underwent a reshuffling of disciplines and some departments merged into other universities.

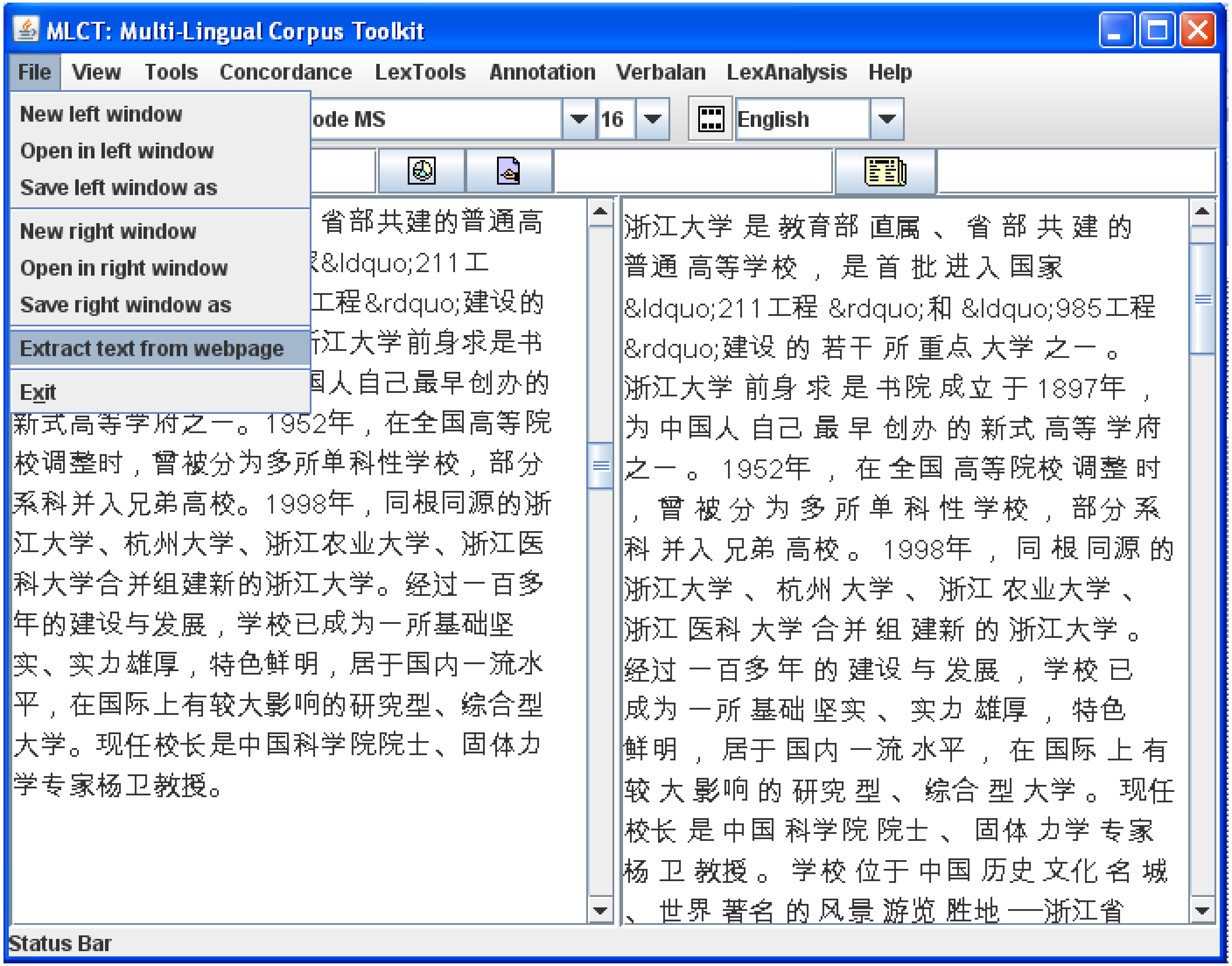
In September 1998, a new Zhejiang University was established on the basis of the amalgamation of the four former individual universities, namely Zhejiang University, Hangzhou University, Zhejiang Agricultural University and Zhejiang Medical University, which were all located in the garden city of Hangzhou. Approved by the State Council, the founding of the new Zhejiang University has been a significant move in the reform and development of China's higher education. The four universities have grown out of the same ancestry, the Qiushi (with the literal meaning of "seeking truth" in Chinese) Academy, which was founded a century ago as one of the earliest institutions of higher learning in China. As a result, they have all inherited from it the spirit of "Qiushi" and at the same time, built up their own distinctive features in teaching and research.

Under the direct administration of China's Ministry of Education, the new Zhejiang University is a key comprehensive university whose fields

```
<p>
<s n="1"> About_IN Zhejiang_NNP University_NNP General_NN
Information_NNP A_DT Brief_NNP History_NN A_DT Message_
from_IN the_DT President_NNP At_IN a_DT Glance_NNP Facul
Alumni_NNP General_NNP Information_NNP Zhejiang_NNP
University_NNP is_VBZ a_DT comprehensive_JJ national_JJ univer
._. </s>
<s n="2"> Founded_VBN in_IN 1897_CD ,_, Zhejiang_NNP
University_NNP was_VBD initially_RB known_VBN as_IN &_CC
ldquo_NN ;_PUN Qiushi_NNP Shuyuan_NNP &_CC rdquo_NN
</s>
<s n="3"> (_( Qiushi_NNP Academy_NNP )_) ,_, one_CD of_IN
ealiest_NN modern_JJ academies_NNS of_IN higher_JJR learning_
established_VBN in_IN China_NNP ._. </s>
<s n="4"> In_IN 1952_CD ,_, because_IN of_IN the_DT nationwi
restructuring_NN of_IN universities_NNS ,_, Zhejiang_NNP Unive
underwent_VBD a_DT reshuffling_NN of_IN disciplines_NNS and
some_DT departments_NNS merged_VBD into_IN other_JJ
universities_NNS ._. </s>
<s n="5"> In_IN September_NNP 1998_CD ,_, a_DT new_JJ
Zhejiang_NNP University_NNP was_VBD established_VBN on_I
basis_NN of_IN the_DT amalgamation_NN of_IN the_DT four_C
```

http://www.zju.edu.cn/english/about/index.htm

Using MLCT to capture web text



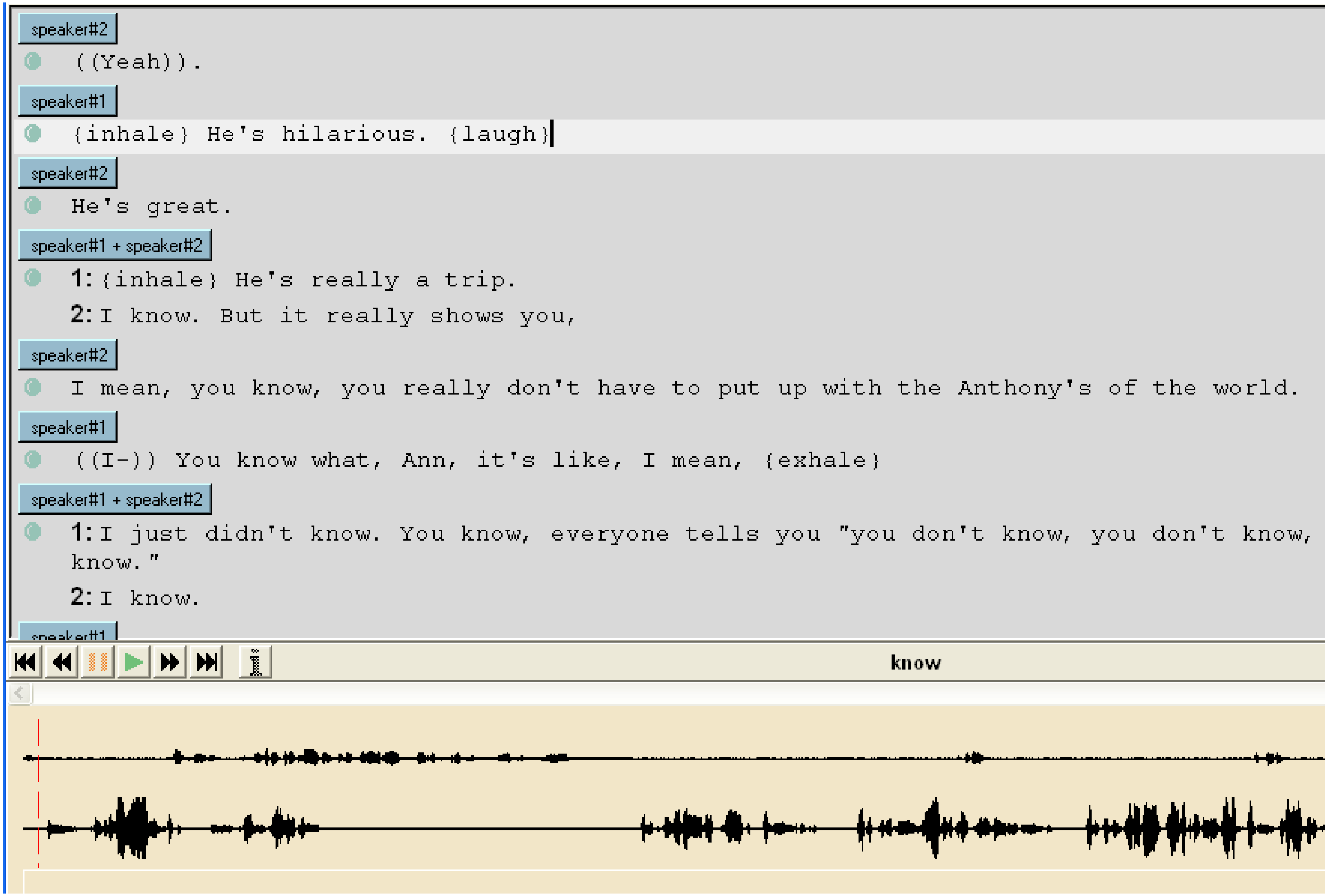
省部共建的普通高
&ldquo;211工
工程&rdquo;建设的
江大学前身求是书
国人自己最早创办的
新式高等学府之一。1952年，在全国高等院
校调整时，曾被分为多所单科性学校，部分
系科并入兄弟高校。1998年，同根同源的浙
江大学、杭州大学、浙江农业大学、浙江医
科大学合并组建新的浙江大学。经过一百多
年的建设与发展，学校已成为一所基础坚
实、实力雄厚，特色鲜明，居于国内一流水
平，在国际上有较大影响的研究型、综合型
大学。现任校长是中国科学院院士、固体力
学专家杨卫教授。

浙江大学 是 教育部 直属 、 省 部 共 建 的
普通 高等学校 ， 是 首 批 进入 国家
&ldquo;211工程 &rdquo;和 &ldquo;985工程
&rdquo;建设 的 若干 所 重点 大学 之一 。
浙江大学 前身 求 是 书院 成立 于 1897年 ，
为 中国人 自己 最 早 创办 的 新式 高等 学府
之一 。 1952年 ， 在 全国 高等院校 调整 时
， 曾 被 分 为 多 所 单 科 性 学校 ， 部分 系
科 并入 兄弟 高校 。 1998年 ， 同 根 同源 的
浙江大学 、 杭州 大学 、 浙江 农业大学 、
浙江 医科 大学 合并 组 建新 的 浙江大学 。
经过 一百多 年 的 建设 与 发展 ， 学校 已
成为 一所 基础 坚实 、 实力 雄厚 ， 特色
鲜明 ， 居于 国内 一流 水平 ， 在 国际 上 有
较 大 影响 的 研究 型 、 综合 型 大学 。 现任
校长 是 中国 科学院 院士 、 固体 力学 专家
杨 卫 教授 。 学校 位于 中国 历史 文化 名 城
、 世界 著名 的 风景 游览 胜地 —浙江省

Transcriber

- A tool for assisting the manual annotation of speech signals
  - Segmenting long duration speech recordings
  - Transcribing audio recordings
  - Labelling speech turns, topic changes and acoustic conditions
- Supporting multiple platforms
  - Windows XP/2k
  - Mac OS X
  - Linux
- Downloading the programme, user manual, annotation guide
  - http://sourceforge.net/projects/trans/

Transcriber



speaker#2

((Yeah)).

speaker#1

{inhale} He's hilarious. {laugh}

speaker#2

He's great.

speaker#1 + speaker#2

1: {inhale} He's really a trip.

2: I know. But it really shows you,

speaker#2

I mean, you know, you really don't have to put up with the Anthony's of the world.

speaker#1

((I-)) You know what, Ann, it's like, I mean, {exhale}

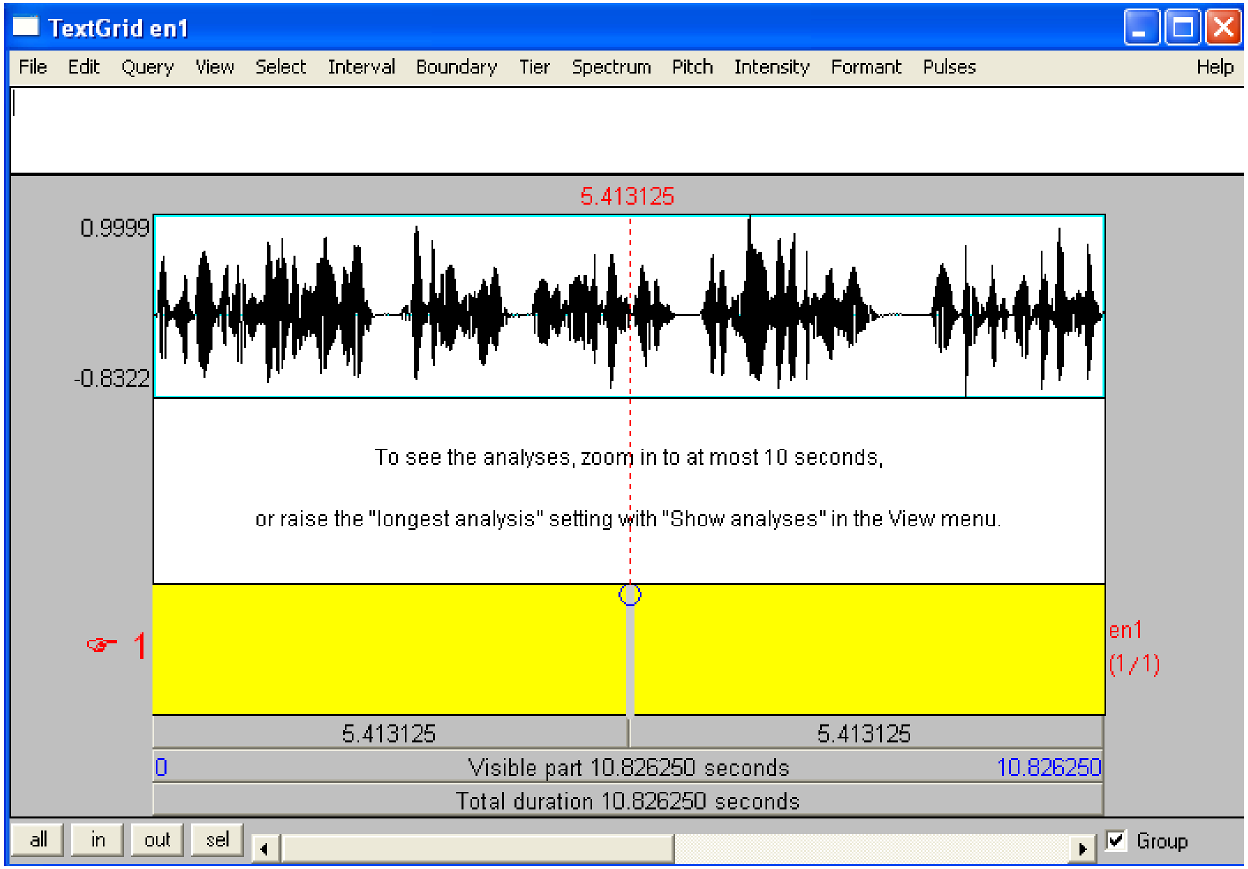
speaker#1 + speaker#2

1: I just didn't know. You know, everyone tells you "you don't know, you don't know, know."

2: I know.

speaker#1

Praat



To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the View menu.

Visible part 10.826250 seconds

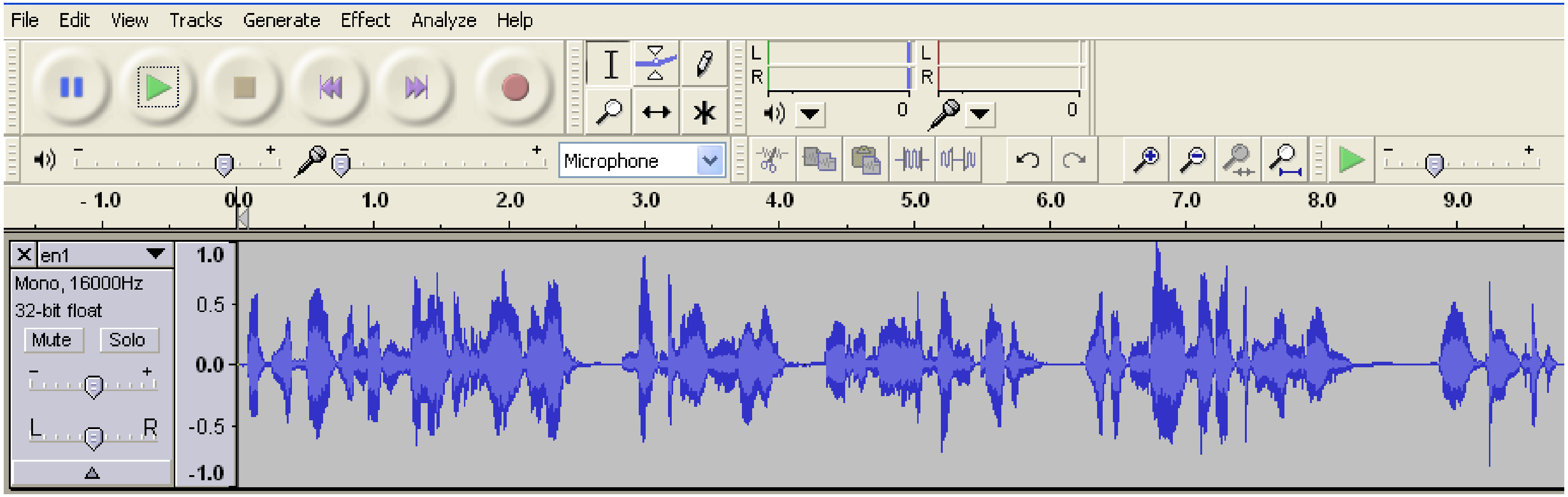
Total duration 10.826250 seconds

Well known and widely used (many online tutorials)

Suitable for acoustic analysis of files that are shorter than 15 minutes

http://www.fon.hum.uva.nl/praat/download_win.html

Audacity



Recording and editing sounds

Can work with large files

Digitalise your cassette tapes
Download at http://audacity.sourceforge.net/

Voice walker: http://www.ruf.rice.edu/~reng/trans/voicewalker.html
F4: http://www.audiotranskription.de/english/f4.htm