University of
HUDDERSFIELD

Corpus Mark-up and Annotation

1. Introduction

As well as raw or plain text, corpora may also include:

__Mark-up__: data about data, e.g. a description of what the text *is*; or data about features of the text formatting and structure.

__Annotation__: data resulting from linguistic analysis of the corpus (linguistic metadata).

> **WARNING*! Markup* and *annotation* are often used interchangeably in the literature. Also, "tagging" can mean *any* addition of extra information to a corpus**

2. Mark-up

2.1 Marking features of the original text (beyond just the words).

- This is typically done by adding codes to indicate features of the original layout / structure of the text, such as: paragraph/sentence/chapter start/end points; page breaks; headings (as opposed to 'normal' paragraphs).
- Mark-up can be used in corpus searches. For instance, you could search for a word together with a mark-up code. This would allow you to find instances of words when they occur, for example, in headings or at the start of a paragraph.

2.2 Marking extra information about the text.

- Corpus files can also contain data about the data, such as where the text was published or recorded, who wrote it or spoke it, and so on. This is usually at the start of each corpus file, and is called the *header*, although some data, for example speaker sex, might be embedded in the body of the file, depending on how the data is structured.
- We can use the metadata about a corpus to limit our searches to particular sorts of texts. For example, in BNCweb, it is possible to restrict a search to just written or just spoken data, and the software uses the metadata in the headers of the BNC texts to do this.

3. Annotation

Annotation is the addition of extra data that results from some sort of linguistic analysis – linguistic metadata. Depending on the type of linguistic analysis that needs to be carried out to add this data, this sort of annotation can be added manually or automatically by computer taggers.

3.3 Types of annotation

- *Discourse annotation*: e.g. tagging anaphora – tagging pronouns to indicate what they refer back to
- *Pragmatic annotation* e.g. tagging utterance for their speech act; for their level of politeness; for their use of indirectness; etc.
- *Stylistic annotation* e.g. tagging different kinds of speech and thought representation (Semino and Short 2004)
- *Parts of speech annotation,* and *semantic annotation* – more about these below.

3.1 Part-of-speech (POS) tagging

The *CLAWS* (Garside 1987) tagger developed at Lancaster University is a part of speech tagging, which annotates each word with a label to show its grammatical category. It uses one of 3 tagsets:

- The **C5** tagset: used in the BNC, 62 tags, simplified a bit for non-linguists
- The **C7** tagset is used most of the time. It has 137 categories, and thus allows more detailed linguistic distinctions to be made.
- The **C8** tagset is a more detailed version of C7
- You can try it out here on up to 100,000 words: http://ucrel.lancs.ac.uk/claws/trial.html

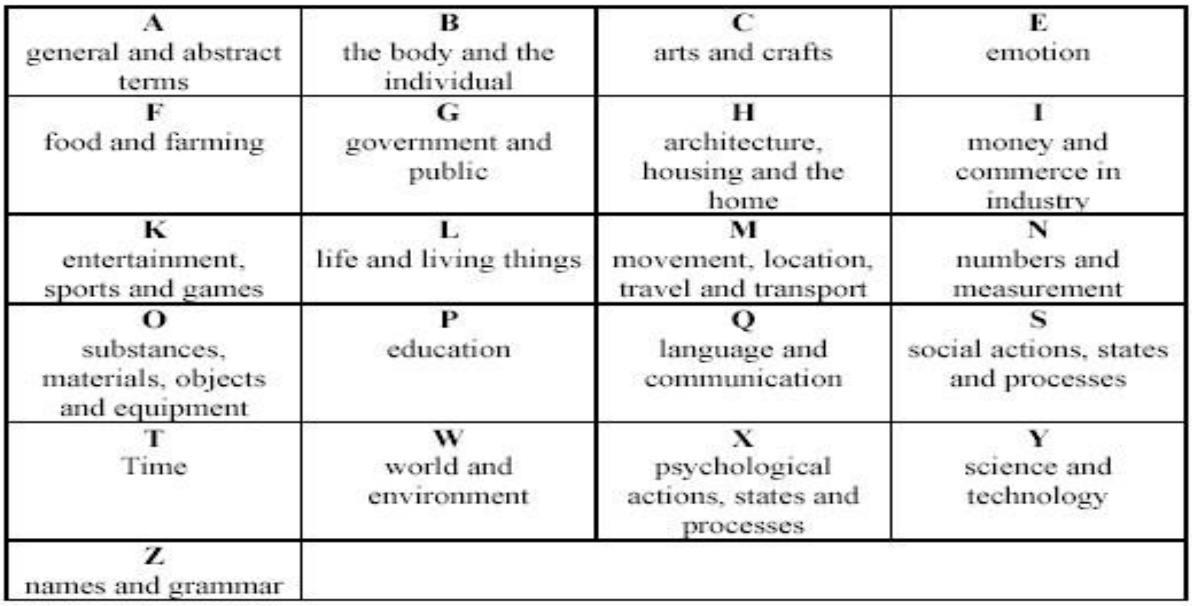
Example - a short extract from a novel tagged using the C7 tagset.

I_PPIS1 liked_VVD him_PPHO1 ,_, and_CC he_PPHS1 was_VBDZ different_JJ from_II other_JJ boys_NN2 ,_, not_XX at_RR21 all_RR22 pushy_JJ ,_, except_CS pushy_JJ to_TO please_VVI I_PPIS1 suppose_VV0 ,_, but_CCB even_RR that_DD1 was_VBDZ sweet_JJ in_II a_AT1 way_NN1

Notice that each word in the text is followed by an underscore and a code which tags what part of speech each word is. For example, liked_VVD = past-tense lexical verb.

3.2 Semantic tagging

Lancaster University has also developed a system of semantic tagging called *USAS*. In semantic annotation, tags are assigned that refer to categories of **meaning**



A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

The USAS tagset. Each major tag is a broad area of meaning, with many subdivisions

Example- The same extract from a novel, but this time tagged by *USAS*.

I_Z8mf liked_E2+ him_Z8m ,_PUNC and_Z5 he_Z8m was_A3+ different_A6.1- from_Z5 other_A6.1- boys_S2.2m ,_PUNC not_Z6[i163.3.1 at_Z6[i163.3.2 all_Z6[i163.3.3 pushy_S1.2.3+ ,_PUNC except_Z5 pushy_S1.2.3+ to_Z5 please_E4.2+ I_Z8mf suppose_X2.1 ,_PUNC but_Z5 even_A13.1 that_Z8 was_A3+ sweet_X3.1 in_A13.4[i165.3.1 a_A13.4[i165.3.2 way_A13.4[i165.3.3

Notice that, again, each word is followed by an underscore and a code. These tags narrow each word down to a semantic group. For example, liked_E2+ = EMOTIONAL ACTIONS, STATES AND PROCESSES: liking. The plus sign indicates a positive position on the semantic scale.

4. Encoding mark-up and annotation in a corpus.

Various different ways of coding mark-up, metadata and annotations into corpus texts have been tried. The aim of each is to clearly distinguish the original text from what has been added.

As you may have noticed, both CLAWS and USAS use a code that is attached to each word in a corpus with an underscore. Another widely used format for annotation is **XML.**

4.1 XML: eXtensible Mark-up Language

- XML tags or ***elements*** consist of the element name (which you are free to choose) and angled brackets:
  <element> . ....*text*.... . </element> .
- XML tags can also contain extra information called ***attributes***, in the form:
  <element attribute="value"> ..... *text* .... </element>
- XML tags can be embedded within other tags:
  <p><s>This is a sentence.</s> <s>This is another sentence.</s> <s>The two sentences are inside a paragraph.</s></p>
- You can design whatever XML tags you need for your research – there is no fixed set.
- XML tags can be used for mark-up or annotation.

***XML header** – e.g. from the Huddersfield Corpus of EModE News*

```
<File id="J2.1">
<Header>
<Title> The spoyle of Antwerpe </Title>
<Author> George Gascoigne </Author>
<PubDate> 1576 </PubDate>
<Source> EEBO </Source>
<Words> 2112 </Words>
<Comment> to the end </Comment>
</Header>
```

***XML annotation** – e.g. from the Huddersfield EModE discourse presentation Corpus*

```
<dptag cat="N"> Here the book dropt from her hand, and a shower of tears ran down into her bosom. In this situation she had continued a minute, when the door opened, and in came Lord Fellamar. Sophia started from her chair at his entrance; and </dptag>
<dptag cat="NRS"> his lordship advancing forwards, and making a low bow, said, </dptag>
<dptag cat="xDS">"I am afraid, Miss Western, I break in upon you abruptly." </dptag>
<dptag cat="xDS"> "Indeed, my lord," </dptag>
<dptag cat="NRS"> says she, </dptag>
<dptag cat="xDS"> "I must own myself a little surprized at this unexpected visit." </dptag>
```

5. Arguments for and against annotation, and why annotation is useful.

Some linguists (notably, John Sinclair) have argued against the use of corpus annotation because it:

- imposes theoretical preconceptions on the data – so you will never get anything out that you don't put in;
- destroys the integrity of the text;

However, linguists who favour the use of tagging (e.g. Geoff Leech) have argued:

- there are types of searches and analysis that you *cannot do* without annotation;
- the original text is preserved (it is always possible to hide tags or keep an untagged version of the data);
- annotation and mark-up add value to the data.

Practical reasons why annotation is useful:

- **Disambiguation** – telling apart words that have the same spelling but different meanings.
- **Advanced searches** for more fine grained, or complex analyses.
    - e.g. find noun phrases with a single premodifying adjective.

Furthermore, annotation allows you to **preserve your analysis** and make it available to others to use or inspect.

6. Using annotation

- We can use annotation to do **advanced searches** based on tags not words, search for grammatical and other patterns (depending on what tags are added), and do quantitative analysis at **other levels of abstraction** (e.g. semantic analysis).
- Depending on the corpus and how it is tagged, we can, for example, compare language use within different sections of corpora, or compare genres, or sections within genres.

However, certain reservations must be kept in mind:

- it might be risky to use someone's analysis uncritically;
- we need to be aware of the annotation scheme used and any problems it might have;
- the annotation in a corpus might not be perfectly accurate;
- automatic tagging is subject to the limits of the software (% accuracy?).

7. Mark-up versus annotation: a clear distinction?

Is there a clear dividing line between *mark-up* (representation of the text) and *annotation* (analysis of the text; metalinguistic info)?

Any additions to the raw data in a corpus could be said to be acts of interpretation. For example, genre classification, deciding what is a sentence boundary …

8. Conclusion: What is corpus annotation good for?

- It can help to disambiguate elements in our corpus.
- It can make possible searches and frequency counts at higher levels of abstraction than just the word level.
- It can create a common basis of analysis for everyone to use – and to inspect.
- It can open up new possibilities for corpus analysis.

9. Further reading.

You can find more information about corpora, and annotation and mark-up at this website:

http://www.ahds.ac.uk/guides/linguistic-corpora/index.htm

Chapter 2 (by Geoff Leech) and Chapter 3 (by Lou Burnard) are the most relevant.

Also try:

Leech, G (1997) Introducing Corpus Annotation. In Garside, Leech & McEnery (1997: 1-18)

10. References

Garside, R, Leech, G and McEnery, A (eds) (1997) *Corpus annotation*. London: Longman.

Garside, R. (1987). The CLAWS Word-tagging System. In: R. Garside, G. Leech and G. Sampson (eds), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman. http://ucrel.lancs.ac.uk/papers/ClawsWordTaggingSystemRG87.pdf

Semino, E. and Short, M. (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.