

Academic research and standards: a discussion on standards for multi-lingual language resources

Pernilla Danielsson and Andrius Utka
{pernilla, andrius}@ccl.bham.ac.uk
Centre for Corpus Linguistics
Department of English
University of Birmingham

1. Introduction

This paper should be read as part of a discussion as to the feasibility of standards in parallel texts and corpus linguistics. Based on experience from several projects working with parallel texts, we would like to argue that whilst recommendations for working practice may be useful to some, the universal need for strict standards is highly questionable. We believe that academic research and standards are not compatible as academic progress in itself implies breaking standards and rules. On the other hand one needs to agree that it would be very convenient, within a research environment, to be able to quickly use, share and test all resources on any computer, without having to spend too much time on changing format.

This paper will begin with a discussion of standards in general, followed by a discussion on more specific matters, such as mark-up and linguistic annotation. The paper benefits from project experience in developing translation bases between English – French, English –Chinese and English-Swedish. Working with such different language pairs, we would argue that standards are best set at the very lowest of levels.

2. Standardising standards

New movements to standardise corpus linguistics seem to surface every few years. EAGLES, CES, TEI are only few examples (Listeri 1996, Ide and Veronis 1995). It is difficult to place a value on this work from its period of activity, however, history reveals that very little impact prevails for future researchers. The question remains as to whether it is worth attempting to once more standardise the field?

Discussions on standardisation usually divide people in two halves; one half argues that standards are limiting to research and progress, the other argues that without standards we cannot exchange experiences and resources. Whilst both halves can successfully argue for their part, the second half appears to lose ground after a few years.

Sometimes standards make life easier. Anyone travelling in Europe using a “European plug” expects to be able to use their electronic equipment in any hotel on the European mainland (including Scandinavia), while a trip to Britain forces you to buy an adaptor. Why then does Britain not wish to follow the standards? Cynics may argue that much of the reason can be found in national pride, others may say “*why fix what is not broken?*”. Most Britons, however, would confidently argue that the three-pinned, fused, plug is a more secure option, i.e. by avoiding the standards Britain has found a safer alternative.

How does this apply to corpus linguistics? In exactly the same way, we believe. Whilst it makes life easier to simply plug a corpus into any software around the world (or at least European mainland), it would also restrict flexibility. The difference lies in that where as electric plugs have one common goal, to provide electricity to their equipment, the use of corpora is not as homogeneous. Users turn to corpora for a vast number of reasons. Thus, before labelling any standard as a “*preferred solution*” many questions should be considered, such as: “*does it cover all interests, all languages and all platforms*”? Our view is that standards are always restrictive and they can never predict future academic requirements. They are therefore bound to be short-lived.

3. In reference to mark-up

In the Call for Papers for this workshop, the following question was asked:

“At present, the preferred solution [to handling parallel texts in tools] is to use XML at all stages and on all layers. But is this really practicable?”

Although we doubt the description “preferred solution”, the question points to an interesting idea that we would all be using XML, but that there are better solutions available. XML is mostly preferred by people that use, or develop, XML-friendly software. However, this is not generally true of people working with parallel corpora. While XML has its advantages, it also suffers from many disadvantages. The following passages attempt to list some of the advantages and disadvantages of using XML according to their relevancy to corpus linguistics. For further discussion on XML see, for example, at <http://www.zapthink.com/reports/proscons-view.html>.

The main advantages of XML are that it:

- allows complex and focused searching procedures;
- supports most languages;
- separates the process from the content;
- is platform and system independent: it can be used on any computer;
- is web-friendly.

On the other hand, the main disadvantages of using XML is that:

- it imposes complicated and often restrictive high-level structures upon language;
- not all XML metadata is always necessary and so often it becomes a burden rather than a help;
- it is not always reversible to the original text, in other words, not all metadata can be easily removed;
- it limits the choice of software: only newer software is XML-friendly;
- it is time consuming and expensive to convert the existing information into XML;
- it requires a lot of expertise (at present more than 400 XML standards are available);
- XML specifications are not complete yet, it is still an experimental standard.

Not all of these points are crucially relevant to the linguistic research community. However, it is the top two points that cause the biggest controversy. They also cause many linguists to compromise. If you want to make complex and focused searches (for example, part-of-speech sensitive limited to the context of titles), you need to impose some sort of structures upon language. So, although one could strongly attack the XML standard on this point, text structuring cannot be altogether ignored, when a more complicated and controlled search is needed.

However, most present-day parallel corpora do not have a complex structure. In fact, parallel corpora are rarely over two million words and apart from occasional grammatical tagging, they are often without linguistic annotation. When using parallel corpora stored as plain texts, the UNIX system command ‘grep’ offers a cheap, but powerful, tool for text retrieval. Grep is still one of the most efficient text search engine available where relevant hits can be retrieved instantaneously. The following illustrations show some example of the Swedish word “*råga*”. In the examples below, *råga* is used in the multi-word unit “*till råga på (allt/elände)*” which translates into English as ‘*to add insult to injury*’ or ‘*on top of all*’.

grep -A1 -B1 'råga' *.eos.al|more

brandbil.eos.al-*** Länk: 1-1 ***

brandbil.eos.al:Till råga på allt var ju detta i stort sett sant. .

brandbil.eos.al-To add insult to injury, that was largely true.

--

kakelbok.eos.al-*** Länk: 1-1 ***

kakelbok.eos.al:Och till råga på eländet hade det tydligen blivit något problem med

kakelsättarna.

kakelbok.eos.al-And on top of all that there'd evidently been some problem with the tilers.

Example 1. Using the UNIX command 'grep' to retrieve information from parallel texts in ASCII format.

Having worked on a number of different projects, we have encountered a variety of standards for monolingual as well as parallel corpora. These range from a simple ASCII format to a highly-structured XML standard (see Table 1 below). All of these standards are recognized and successfully used for a variety of tasks.

ASCII	<i>Source file (English)</i> Yet the euro has been perceived as "weak".	<i>Target file (French)</i> Pourtant, l'euro a été perçu comme une monnaie "faible".
Corpus WorkBench and TreeTagger	<seg id=126 type=1-1> <s id=en.126.2> Yet CC yet the DT the euro NP Euro has VBZ have been VBN be perceived VBN perceive as IN as " " " weak JJ weak " " " . SENT . </s> </seg>	<seg id=126 type=1-1> <s id=fr.126.2> Pourtant ADV pourtant , PUN , l' PRO:PER la le euro PRO:PER <unknown> a VER:aux:pres avoir été VER:ppe être perçu VER:ppe percevoir comme KON comme une DET:ART un monnaie NOM monnaie " PUN:cit " faible ADJ faible " PUN:cit " . SENT . </s> </seg>
"Vanilla Aligner" output 1	*** Link: 1-1 *** Yet the euro has been perceived as "weak". Pourtant, l'euro a été perçu comme une monnaie "faible".	
"Vanilla Aligner" output 2	<align id=126 type=1-1> <s id=en.126.2>Yet the euro has been perceived as "weak". </s> <s id=fr.126.2>Pourtant, l'euro a été perçu comme une monnaie "faible". </s> </align>	
XML/SGML	<i>Source file (English)</i> <seg id=126 type=1-1> <s id=en.126.2><w type="CC" lemma="yet">Yet</w> <w type="DT" lemma="the">the</w> <w type="NP" lemma="Euro">euro</w> <w type="VBZ" lemma="have">has</w> <w type="VBN" lemma="be">been</w> <w type="VBN" lemma="perceive">perceived</w> <w type="IN" lemma="as">as</w> <c type="PUN">"</c> <w type="JJ" lemma="weak">weak</w> <c type="PUN">"</c> <c type="PUN:sep">.</c></s> </seg> <i>Target file (French)</i> <seg id=126 type=1-1> <s id=fr.126.2><w type="ADV" lemma="pourtant">Pourtant</w> <c type="PUN:comma">,</c> <w type="PRO:PER" lemma="la le">l'</w> <w type="PRO:PER" lemma="<unknown>">euro</w> <w type="VER:aux:pres" lemma="avoir">a</w> <w type="VER:ppe" lemma="être">été</w> <w type="VER:ppe" lemma="percevoir">perçu</w> <w type="KON" lemma="comme">comme</w> <w type="DET:ART" lemma="un">une</w> <w type="NOM" lemma="monnaie">monnaie</w> <c type="PUN">"</c> <w type="ADJ" lemma="faible">faible</w> <c type="PUN">"</c> <c type="PUN:sep">.</c></s> </seg>	

Table 1. Different formats for parallel corpora.

Interestingly the widely-used LOB Corpus the British counterpart of the Brown corpus (Johansson et al 1986) is available in at least three different formats (see Appendix I). All of these formats have been designed to serve different user needs, and nobody of those who did a little bit of text processing would claim that one of these formats is worse or better than the other: all of them are equally good for certain tasks and tools.

The standards we have enumerated are just the tip of the iceberg. The existence of such a variety of standards can be explained by the fact that they are essentially task-driven, and then they also depend on available tools. Any new tasks are creating new requirements for tools as well as for standards. Therefore our recommendation is not to invest too much time and money into new elaborate standards, but rather choose a flexible, cheap and easy convertible option that will satisfy user's minimal requirements.

4. Linguistics annotation

In the Call for Papers for this workshop, the following question was posed:

“What happens, if the units under investigation diverge on the different levels?”

This question becomes very relevant when working with linguistically diverged languages. Trying to compare tagsets between languages such as English and Chinese identifies many discrepancies on how to annotate a language. This section will contain a brief comparison between two such tagsets.

One needs not work with languages of such different nature to find divergences. More closely related languages, such as English, French and Swedish, offer a vast amount of divergence as well.

In our projects, the main concern is to compile a “*TranslationBase*” for each language pair, focusing on multi-word units as the main unit of analysis. Many present-day corpus linguistic studies have shown that in order to extract meaningful units from corpora, we need to focus on maximal units, such as *keep an eye on someone* or *pay your respects* (see for example Sinclair 1997, Teubert 2002, Hunston 2002). However, the study is still in its infancy and there is no free software to support such an approach. Anyone trying to work with multi-word units is most likely to have to write their own tools for retrieval and mark-up, as well as classification (annotation) of the units. Some experiences from on-going projects in Birmingham will be given below.

4.1 Comparing Chinese and English tagset

Appendix II contains a comparison between English and Chinese Part-of-Speech (POS) tags. Two main differences become apparent:

- a) The Penn TreeBank (Santorini 1991) tagset has split the main word classes, such as noun, verb and adjective, into several tags. For example, there is no over-all Noun-tag, we instead find separate tags for Noun singular and Noun plural.
- b) The Chinese tagset has lumped together adjacent words into one and the same tag, we find special tags for idiom and for frequently used fixed expressions.

In general terms we may regard the English tagset as being a splitter, in that it tries to divide everything into as many categories as possible. This is true not only for the Penn TreeBank tagset, but in fact can be said about most English sets. It is further promoted by the fact that many tags are inherited from one tagger to another by allowing the taggers to train on existing tagged corpora. Many English taggers have therefore inherited the tags and categories from the first freely available corpus, the Brown corpus (Kucera and Francis 1967). Along the more surprising sub-tags, we find tags restricted only to be used with one word, such as the very various forms of the verb *be* where *is* has the tag *BEZ* and *was* has the tag *BEDZ*.

On the other hand, the Chinese tagset was lumping words together. The tendency of Chinese tags to tag larger units is perhaps a direct result of having been faced with problems of segmenting texts into meaningful units. Segmentation is performed on Chinese texts in order to clearly delimit each unit. Unlike common misconceptions, one Chinese character does not always correspond to one word. Many words are instead constructed by using two or more characters. In order to make computational tools work on the appropriate unit, most self-respecting Chinese universities have developed their own segmenting tools. This is illustrated in example 2 below, where the text has been segmented and tagged.

裁判官/n 可/d 随时/d 命令/v 将/p 警务处/n 处长/n 根据/p 第/m 2/m 条/q 接管/v 的/u
任何/b 处所/n 发还/v 。 /w

(lit. “The judge can, at any time, order the police to return the a place which has been taken over, according to rule 2”)

Example 2. Example of Chinese segmented and POS tagged text.

European languages could perhaps benefit from copying this approach. Recent studies in linguistics have shown that the unit of analysis in languages, such as English, is not always a single word. Instead, a unit of meaning is often constructed of two or more words, such as *keep an eye on* and *turning a blind eye*, but this has not yet influenced the software we use. Text retrieval software still works on single words as basic units; see for example all indexing software. The only new approach seems to come in the form of suffix trees, or regular expression searches, which allow users to specify their search unit beyond word limits (Munteanu and Marcu 2002). Although mark-up languages, such as XML, do not

explicitly state whether the units have to be single words, the use of *w* as the tag does lead to a direct link *word*. However, they may still be used to identify multi-word units (see below).

Tools aimed at promoting linguistic annotation may also be accused of furthering the single unit approach. The example below comes from the British National Corpus, the BNC (see for example BNC Manual 1998 or Aston and Burnard 1998). The text has been marked-up in SGML, enabling extensive information in the form of a tag around each word. In the BNC, some multi-word units were identified and marked-up in the texts using this method. This approach however, highlights all of the problems that we still have to face before being able to work with multi-word units as a base.

The word-tags in the corpus also contain part-of-speech information, this is a result of the automatic tagging carried out by the CLAWS system (see for example Aston and Burnard (1998) for usage of CLAWS tags). As a consequence of using CLAWS, the tokenisation makes two words out of a unit such as *won't* (<w VMO>wo<w XXO>n't) whilst several (297 to be exact, according to the BNC manual) common multi-word units are treated as one word, such as *a little, now that, at long length, as well as*. While no one will question the presence of *as well as*, many may question the relevance of the multi-word unit *at long length*. The fact that several of these multi-word units have been over-generalized in their use, such as *a little* and *now that* in example 3 and 4 below is perhaps more serious. This over-generalisation is somewhat surprising since the tokeniser, the tool that should segment the text into tokens, is also the tagger, and the information that *a little* when followed by a noun should not be tagged as a multi-word unit seems to be a very straightforward process for a combined tokeniser and tagger.

```
[...] <w VVG>seeing <w DTO>a little <w NN1>girl <w PRP>in <w AT0>a <w  
AJO>tattered <NN1>dress <w CJC>and <w PRP>with <w AJ0>bare <w  
NN2>feet[...]
```

Example 3. Example of over-generalising the use of the multi-word unit *a little*.

```
<W TYPE=NN1>Hen</W><C TYPE=PUN>, </C><W TYPE=CJS>now  
that</W><W TYPE=VBZ>'s </W><W TYPE=PRP>worth </W><W  
TYPE=VHG>having</W><C TYPE=PUN>.</C></S>
```

Example 4. Example including the faulty tokenising around *now that*.

In addition to the English multi-word units, we also find foreign multi-word units that have been treated as single units, such as *belles lettres, et al, in camera, savoir vivre*. It should, however, be pointed out that the BNC manual acknowledges that these multi-word units might not have been recognised consistently and that some of them might be tagged as individual units (BNC Manual, 1998). From the two examples above, it becomes obvious that the BNC project must have been using different standards at different points in the project. This causes the result to have a heterogeneous mark-up in different sections.

Erroneous tagging such as this, further emphasises the importance of studying the characteristics of multi-word units; what is the maximum unit, how can a unit be varied or altered, when is a unit likely to be used in word play? Before understanding the behaviour of maximal units, we are likely to introduce more chaos than order in annotated corpora.

4.2 Retrieving multi-word units.

The problem of marking up relevant multi-word units is directly linked to the difficulties in retrieving them. Present methods usually involve various statistical calculations (such as χ^2 -square used in the example below) to highlight frequently co-occurring patterns.

In our projects, we are currently experimenting with several methods to identify multi-word units. One way is to identify linguistically relevant patterns, such as noun followed by a second noun. Since the word class noun usually contains meaning carrying words, a focus on such linguistically relevant patterns can free the result-list from word combination with little or no meaning attached, such as “*of the*” (described in more detail in Chang et al 2003).

Chinese

English

χ^2 -score

成人_图书馆	adult_library	68620.5
影子_董事	shadow_director	68469.8
幕_墙	curtain_wall	68469.8
卤味_店	lo_mei	68282.1
橡胶_手套	rubber_glove	68041.9
橡胶_围裙	rubber_apron	67723.5
疾病_津贴	sickness_allowance	67433.1
计算机_软件	computer_software	67281.6

Example 5. Chinese – English multi-word unit pairs.

The example above is only to be viewed as a first attempt. It encompasses several problems, in common for many bilingual approaches using statistics today. For anyone objecting to the idea of implementing their own computational tools, several statistical programs are available. Most of them are built into concordance programs (see for example *WordSmith Tools*, <http://www.lexically.net/wordsmith/>), but several statistical tools are available as well (such as Pedersen's *Ngram Statistical Package* <http://www.d.umn.edu/~tpederse/code.html>). Common within these tools are a few limiting assumptions about language. These tools expect every unit to be of a pre-determined length, primarily bigram, i.e. word pairs, but occasionally also trigram (three word combinations) or tetragrams (four word combinations). The idea that all multi-word units of meaning should be constructed from unbroken strings of word forms is also prevalent. The sequential approach to viewing text is by far the most dominant.

In the example above, multi-word units in Chinese were matched to multi-word units in English. This is only one of the approaches we are currently testing. There is no reason to expect a multi-word unit in one language to correspond to a multi-word unit in another language. Once a multi-word unit has been established it should be treated as one single unit, and any attempt to align it with corresponding units in the target text has to be performed by first extracting units of analysis. As illustrated in the example below, the Swedish multi-word unit “*i stort sett*”, translates into a single word unit in English, *largely*.

Uppgifterna i Expressen hade därvidlag i stort sett varit korrekta, även om dom överdrivit Carls roll i den slutliga uppgörelsen.
The information in Expressen had been largely correct in that regard, although they had exaggerated his role.

De åt räkor och drack vargtass och gjorde i stort sett vad man förväntades göra ombord.
They had shrimps, drank Finnish vodka and juice, and largely did what was expected on board ship.

Example 6. Swedish multi-word unit “*i stort sett*” translates into English single-word unit “*largely*”.

5. Conclusion

In this paper, we have argued the following:

- standards and academic research are not compatible;
- standards are essentially task-driven and therefore short-lived, and thus not worth spending time and money on developing;
- linguistic annotation cannot be expected to carry across the language borders, neither at the level of part-of-speech nor at the level of maximal unit;
- linguistic annotation focuses on a single-word unit, and has not yet been able to follow up on recent findings in corpus linguistic where units of meaning of have proven to be multi-word units.

The focus of the workshop is to discuss possible standards for marking-up and annotating parallel texts. In our experience, every new project and every new language pair has its own peculiarities, whether it is at the level of sentence alignment or attempts to align smaller units. Rather than setting standards at a high level of mark-up, i.e. standards that only a skilled computational linguist can conform to, corpus linguists should aim at finding guidelines to good working practice. It is our belief that standards are not only hard to follow but in fact they may also be damaging to academic research, in that they may

stop progress in new directions. One of these new directions that corpus linguistics has taken lately is the focus on larger units of meaning. At present, freely available software are hard to find attempting to work on larger units than words and smaller than sentences. If standards are set to mark-up and annotate parallel texts at word level, the progress in this area may slow down even further.

References

- Aston G, Burnard L 1998 *The BNC Handbook: Exploring the British national Corpus with SARA*. Edinburgh, Edinburgh university press.
- BNC Manual 1998 *Users Reference Guide for the British National Corpus, Version 1.0, 1998*. Oxford, Oxford University Computing Services.
- Chang B, Danielsson P, Teubert W 2003 Chinese-English Translation Database. In Barnbrook G, Mahlberg M, Danielsson P (eds), *Meaningful Texts: Corpus and Discourse*. Birmingham, Birmingham University Press.
- Danielsson P, Ridings D 1997 Practical Presentation of a 'Vanilla' aligner. In Reyle U, Rohrer C (eds), *The TELRI Workshop on Alignment and Exploitation of Texts*. Ljubljana, Institute Jozef Stefan.
- Johansson S, Atwell E, Garside R, Leech G 1986 *The Tagged LOB Corpus: User's Manual*. Bergen, Norwegian Computing Centre for the Humanities.
- Hunston S, Francis G 1999 *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam, John Benjamins.
- Ide N, Veronis J (eds) 1995 *The Text Encoding Initiative: Background and Context*. Dordrecht, Kluwer Academic Publishers.
- Kucera H, Francis W N 1967 *Computational Analysis of Present-Day American English*. Providence, Brown University Press.
- Llisteri J 1996 EAGLES Text Corpora Working Group, Reading Guide, EAG--TCWG--FR—2, Version of May, 1996, "url = <http://www.ilc.pi.cnr.it/EAGLES96>".
- Munteanu D S, Marcu D 2002 Processing Comparable Corpora With Bilingual Suffix Trees, "url=<http://citeseer.nj.nec.com/537541.html>".
- Santorini B 1991 *Part-of-Speech Tagging Guidelines for the Penn Treebank Project. Technical Report MS-CIS-90-47*. Pennsylvania: Department of Computer and Information Science, University of Pennsylvania.
- Schmied H 1994 Probabilistic Part-Of-Speech Tagging using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, pp 44-49.
- Sinclair J 1997 The Lexical Item. In Weigand, E. (ed.), *Contrastive Lexical Semantics. Current Issues in Linguistic Theory*. Amsterdam, John Benjamins.
- Teubert W 2001 Corpus Linguistics and Lexicography. *International Journal of Corpus Linguistics* 6 (Special Issue): 125-153.
- Teubert W 2002 The role of parallel corpora in translation and multilingual lexicography. In Altenberg B, Granger S (eds), *Lexis in Contrast. Corpus-based Approaches*. Amsterdam, John Benjamins.

Appendix I

Different formats of the LOB Corpus.

LOB Corpus, untagged version, text	A01 2 *<<*7STOP ELECTING LIFE PEERS**>> A01 3 *<<*4By TREVOR WILLIAMS*>>
LOB Corpus, tagged version, horizontal format	A01 2 ^*'_* stop_VB electing_VBG life_NN peers_NNS **'_'_.. A01 3 ^by_IN Trevor_NP Williams_NP . . .
LOB Corpus, tagged version, vertical format	A01 2 001 ----- A01 2 002 *' *' H A01 2 010 VB stop H A01 2 020 VBG electing H A01 2 030 NN life H A01 2 040 NNS peers H A01 2 041 **' **' H A01 2 042 . . H @ A01 3 001 ----- A01 3 010 IN by H A01 3 020 NP Trevor H A01 3 030 NP Williams H A01 3 031 . . H @ A01 4 001 -----

Appendix II

The following table is a comparison between the English and Chinese tagsets used in Birmingham Chinese-English Translation Base. The table is published in its entirety in Chang et al (2003).

The Penn Treebank English tagset		The ICL/PKU Chinese tagset	
JJ	Adjective	a	Adjective
JJR	Adjective, comparative	b	Distinctive
JJS	Adjective, superlative	z	Status
NN	Noun, singular or mass	n	Noun
NNS	Noun, plural	s	Location
NP	Proper noun, singular	r	Pronoun
NPS	Proper noun, plural		
PP	Personal pronoun		
PPS	Possessive pronoun		
WP	Wh-pronoun		
WPS	Possessive wh-pronoun		
VB	Verb, base form	v	Verb
VBD	Verb, past tense		
VBG	Verb, gerund or present participle		
VBN	Verb, past participle		
VBP	Verb, non-3 rd person singular present		
VBZ	Verb, 3 rd person singular present		
MD	Modal		
EX	Existential <i>there</i>		
RP	Particle		
TO	<i>to</i>		
DT	Determiner		
		f	Direction
		o	Onomatopoeia
		i	Idiom
		g	Morpheme
		h	Prefix
		k	Suffix
		l	Frequently used fixed expression