Tagging and Alignment of Parallel Texts: Current Status of BCP

A. Winarske
ISSCO, Genève*

S. Warwick-Armstrong
ISSCO, Genève†

J. Hajič
Charles University, Prague‡

1 Introduction

Access to on-line corpora is a useful tool for studies in lexicography, linguistics, and translation. Many means of accessing such corpora are available, but few, if any, provide more than a language for matching character strings. As a result, the user is obliged to spend a great deal of time extracting information herself. As more and more texts are put in machine readable format, it becomes increasingly obvious that more specialized, intelligent tools are required to fully exploit the available data. BCP, the Bilingual Concordancy Program under development at ISSCO, is an instance of such a tool.

In previous work done at ISSCO on BCP, a rather oversimplified view of text structure was taken [Warwick *et al.*, 1989]. Attention was focused on the difficulties of alignment and somewhat less so on access questions. Alignment remains a subject of active research, but experience has proven that text marking and morphology are not to be taken so lightly. Indeed, many small difficulties have shown themselves to be insurmountable without the aid of heuristic decision modules. As a result, the initial approach to text tagging and morphology has been thoroughly revised.

2 Brief Overview

The BCP package consists of four submodules: preprocesser, morphology, alignment, and access. The text pre-processor, **bcpmark**, marks paragraph and sentence boundaries, numbers, words, and punctuation. The morphological analyzer, **bcpmorf**, is built around a unification-based parser, and returns feature-structure descriptions in SGML format, although the feature structure itself is in a linear notation only. The alignment module is the subject of much experimentation and currently is running with the Church-Gale alignment algorithm [Gale and Church, 1991]. The access module has been described in previous work [Warwick *et al.*, 1989] and will not be discussed further here. The focus of this abstract will be on **bcpmark** and **bcpmorf**.

*Many thanks to Graham Russell for his invaluable advice on this abstract.

†ISSCO, 54 route des Acacias, Genève 1227, Switzerland

‡Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, 118 00 Praha 1, Czechoslovakia

3 bcpmark: The Pre-Processor

bcpmark is the first step in preparing text for the alignment program. It marks paragraph and sentence boundaries, numbers, words, and punctuation, with the output in SGML notation. **bcpmark** is easily customized to suit a particular text type or language via a user-defined data file. Extensions and alterations to the data are accordingly simple. There are accompanying tools to check number standardization results and sentence boundary marking. Languages currently supported are French, German, Italian, Czech, and English.

3.1 Input Text

bcpmark is intended to be usable on all text types, so that entails a certain amount of flexibility. Regardless, there are two major problems: no interpretation of the input text, and the need to be "parameterized" for different textual conventions.

Problems instantly arise in conjunction with numbers, abbreviations, conflicts with differing punctuation conventions, and capitalization. In particular, German noun capitalization causes great problems to a system which relies heavily on capitalization marking sentence beginnings.

In **bcpmark**, the sentences are marked by either the onset of a paragraph marker or by encountering an end-of-sentence punctuation mark in the appropriate context for a particular language. We define six contexts essential for deliminating sentences:

1. Characters are always considered part of a word.
2. Abbreviations which can never end a sentence even if they are followed by a dot. There may also be contracted abbreviations.
3. Abbreviations which in front of a number cannot end a sentence.
4. Words which followed by a number followed by a period usually signal a sentence boundary.
5. The sequence
   `single-capital-letter. capitalized-word`
   is normally not recognized to be a sentence boundary.
6. Certain words followed by sequences of the form
   `number. capitalized-word`

(especially in German texts) should not be marked as sentence boundaries.

7. Words which probably do not start a new sentence if preceded by a sequence number . This is especially useful for languages like German, which mark ordinal numerals by dots which do not indicate an end of sentence.

4 Morphology

Morphological variations can be classified as inflection, derivation, and compounding [van Gaalen *et al.*, 1991]. An adequate morphology should be able to handle all three. There are several parts to the BCP morphology: the morphology grammar, the regular and irregular dictionaries, and the code. There is also a facility for testing and debugging the morphology grammar. The output format is an SGML-notation version of feature structures, where ambiguous analyses are expressed in tags raher than multiple word-forms in the text.

5 Alignment

The technique originally used for aligning texts was to link regions of texts according to regularity of word co-occurrences across texts [Catizone *et al.*, 1989]. Pairs of words were linked if they have similar distributions in their home texts. This strategy doesn't always work well because in many languages a good writer does not use the exact same word many times in a text. Similarly, a good translator does not always translate a word exactly the same way every time it occurs. Clearly this algorithm is heavily text dependent. For texts with limited vocabularies this might work extremely well, but in "free" text it fails.

Currently we are experimenting with assorted algorithms; a major problem is having good test texts to run them on. So far the best results on reasonable text come from the Gale-Church algorithm [Gale and Church, 1991]. It has been tested on English, German, French, Czech, and Italian parallel texts. The Gale-Church algorithm relies on the length of regions, where the character is the unit of measurement. (For details see their paper.) We have experienced three problems with this method. First, the implementation of the algorithm published in Church-Gale severely limits the size of the input file [Gale and Church, 1991]. This is, however, merely an implementation problem. Second, there is no way to set "anchor points" and align around them. That is, one cannot pick two anchor points, one in each text, and have the program align the corresponding regions above and below the anchor points. (See [Brown *et al.*, 1991] for discussion of an alternative.) This is not necessarily a problem either, and can be worked around. Lastly, it does not give usable results on texts which are not absolutely parallel. That is to say, on texts which do not have exactly the same number of large regions, with the same hierarchical structure. A single extra line of characters in one text will cause a complete failure of the alignment algorithm. This is a major difficulty.

6 Conclusion

We are very happy with our marking program and eagerly anticipating thorough testing of the new morphology, especially with regards to extensive experimentation with German texts. We are satisfied with the current alignment method. We may also end up writing a parser to disambiguate the tagged text and this would fit in well with previous ISSCO work on unification-based grammatical formalisms [Estival, 1990]. Clearly there is room for expansion and improvement.

The modular structure of BCP is a great strength, as it enables independent use of the modules. Similarly, the access module functions to its full capacity on the output of the other three, but can also be used on output of the alignment unit alone. This great flexibility clearly lends itself to ease of integration into other systems.

References

[Brown *et al.*, 1991] Brown, P., J. Lai, and R. Mercer. "Aligning sentences in Parallel Corpora", *Proceedings of the Association for Computational Linguistics.* Berkeley, 1991, 169–176.

[Catizone *et al.*, 1989] Catizone, R, G. Russell, and S. Warwick. "Deriving Translation Data from Bilingual Texts", *Proceedings of the First International Lexical Acquisition Workshop.* Detroit, 1989.

[Estival, 1990] "ELU User Manual.", Technical Report Fondazione Dalle Molle, Geneva, 1990.

[van Gaalen *et al.*, 1991] van Gaalen, M., A. Hugentobler, L. des Tombe, S. Warwick-Armstrong. "Terminology Translation Checking for Company X", internal ISSCO/STT proposal, 1991.

[Gale and Church, 1991] Gale, W., and Church, K. draft version of "A Program for Aligning Sentences in Bilingual Corpora", submitted to ACL 1991. See following entry.

[Gale and Church, 1991] Gale, W., and K. Church. "A Program for Aligning Sentences in Bilingual Corpora", *Proceedings ACL 1991, Berkeley*, 1991, 177–184.

[Warwick *et al.*, 1989] Warwick, S, J. Hajič, and G. Russell. "Deriving Translation Data from Bilingual Texts", Proceedings of the First International Lexical Acquisition Workshop, Detroit, 1989.