

Analysing Parallel Texts with ParaConc

Michael Barlow

Department of Linguistics, Rice University, Houston TX 77005, USA

KEYWORDS: parallel corpora concordance

AFFILIATION: Rice University, USA

E-MAIL: barlow@ruf.rice.edu

FAX NUMBER: 713-523-6543

PHONE NUMBER: 713- 630-8761

Analysing Parallel Texts with ParaConc

Much of the current research on parallel corpora concerns the problem of automatic alignment of two texts that are translations of each other (Gale and Church 1994, Kay and Roscheisen 1994, Johansson and Hofland 1993). This paper, however, focusses on the analysis of aligned parallel corpora rather than on the aligning process itself.

In order to analyse a parallel corpus a suitable text analysis program is needed. ParaConc is a simple parallel text concordance program available in Macintosh and Windows versions, which was created by the author as a tool for linguistic research. This program allows the user to search for a word or phrase, in the way typical of concordance programs. However, the result of the search is displayed in two windows rather than one. The topmost window displays numbered lines containing each instance of the search term in the first language, along with its context. The lower window displays numbered sentences in the second language which correspond to the text displayed in the first language in the upper window. The results of a search can be sorted, printed or saved. To obtain a list of words from each text that correspond, as illustrated below for English *line* and French *ligne*, the results of a search are first saved as a text-only file and then loaded into a word-processor for further formatting.

The use of parallel corpora presents very interesting research opportunities in a variety of disciplines including linguistics, literary studies, translation, and language teaching. While these different areas may be touched upon, the focus of the present paper is on the use of parallel corpora in linguistic analysis. This project is similar in spirit to a variety of parallel corpus projects such as Intersect, Contragram, ENPC, and TRIPTIC, among others.

Taking a language to consist of form-meaning links, what we have in parallel corpora are two sets of form-meaning linkings, one for each language.

And since the two texts are translations, the meaning part can be assumed to be approximately the same in both texts. Thus we are able to see how two different languages encode equivalent meanings. The art of translation is undeniably complex and involves many different kinds of processes, but we can consider three main aspects of translation, namely, language particular encodings of (i) event structure, (ii) discourse structure, and (iii) lexis. Each of these areas can be profitably analysed using parallel corpora.

1. Event Structure

Event structure simply refers to those actions occurring in the world that are of interest to humans, such as a transitive event in which one object acts on another object in some way, which is typically encoded using a transitive clause. Since we can assume that the translations are about the same events, we can use parallel corpora to examine how languages code events in general, in other words, how aspects of an event are expressed grammatically or lexically in different languages. An objection that could be raised here is that the particular choices made by a translator will introduce distortions into the data. It is true that some apparently random choices occur in translations, but the accretion of motivated translation choices allow the general patterns to be perceived using a concordance program.

For example, we can examine the coding of causative events in English and French by searching for the lemma *make* and examining patterns such as "X makes Y do Z" and then observing the patterns used in French to refer to these same events. A concordance search reveals that causative *make* in English covers a wide variety of situations, for instance, causing a change in state (*make something possible*), causing someone or something to perform an action (*make a dog go away*), and causing some kind of transformation expressed as *make* followed by two contiguous noun phrases (*make John the president*). Having searched for English causative constructions involving *make*, we can investigate how these different causative event structures are coded in French. And, in fact, we find a rather different set of patterns for French. For the construction of the type *make John president*, the equivalent occurs in French with *faire* in most cases. However, the corpus data shows that other causative uses are often not translated by *faire* in French. A variety of constructions are used instead, including verbs such as *rendre*, as shown in (1).

- (1) a. The American blockage makes life very difficult for us.
Le blocus rend nos conditions de vie très rudes.

On the other hand, uses of *make* expressing a causative event in which an agent acts on an animate causee to bring about an event are more likely to be translated with *faire*, as exemplified in (2).

- (2) a. It is a behaviour which makes you think of France...
Un comportement qui fait penser à la France ...
- b. ... their parents had made them lose their French nationality.
... leurs parents, ..., leur ont fait perdre la nationalité française.

This example shows how ParaConc can be used to investigate fairly subtle cross-linguistic distinctions in the expression of causative events.

2. Discourse structure

Parallel corpora can also be used to highlight the way in which different languages transform the bare bones of event structure into discourse structures appropriate for each language. There are many interesting questions related to the structuring of discourse in different languages and the use of parallel corpora offers one avenue of research in this area. As an example, we can consider how English and French discourse signals the fact that two events occur concurrently (or are alike in some other way). In English the conjunction *while* is used both to link clauses that refer to events that overlap in time and to indicate that the speaker is contrasting two events. The two types, the temporal use and the contrastive use, are shown in (3) and (4).

- (3) a. That means that while we're shooting one film we can start dreaming about the next.
b. That's the way to get the economy going again while at the same time discouraging looters.
- (4) a. While it never misses an opportunity to blame the Socialists for the worsening job situation, the right appears to be just as helpless in the face of rising unemployment.
b. While saleswomen remain as surly as ever, shop windows have become much more attractive.

Using parallel texts, it is possible to search for English *while* and investigate how temporal and contrastive structures are represented in French discourse. Searching for *while* produces a variety of equivalent items in French including: *tout en*,

alors que, *tandis que*, *pendant que*, *contre*, and *si*, among others. To provide a complete analysis of these conjunctions it is necessary to examine the results of the search in some detail and also to examine the translations in English of the different expressions: *tout en*, *alors que*, etc. One result that we can identify is that French *si* is used to indicate a contrastive meaning. Thus the equivalent sentences to (4) are those given in (5), which use *si* for *while*.

- (5) a. Si elle ne manque aucune occasion de verser au débit des socialistes la détérioration de la situation de l'emploi, la droite paraît tout aussi désarmée devant la montée du chômage.
b. Si les vendeuses sont toujours aussi peu amènes, les vitrines, en revanche, se font plus alléchantes.

3. Lexis

ParaConc has several uses in investigating the meaning of lexical items and collocations in two languages. The program can take advantage of the information-on-demand aspect of concordance searching and provide equivalences that may be incompletely captured or not captured at all in bilingual dictionaries. For example, a parallel corpus based on computer texts will allow the user to see how modern computer terms such as *information highway*, *email*, and *home shopping* are being translated. Thus a parallel corpus can be used to reveal both the latest usage and also the variation in usage that occurs.

Rather than explore these lexicographic uses of ParaConc, in this final section I will again pursue a linguistic investigation and indicate how ParaConc can reveal metaphorical and other extensions of a concept occurring in two languages. The word *line* in English, for example, has a variety of uses, some of which are based on extensions of the prototypical meaning. Tied in with these extensions is the existence of certain collocations such as *hard line*, *firm line*, etc. Some of these extensions are also present in French; others are not. We find, for instance that the *in line with* uses do not appear to have a French equivalent based on *ligne*.

In (6) a small sample of correspondences is given. (Examples of *ligne* and their equivalents in English are omitted from this abstract for reasons of space.)

- (6)
- | | |
|--------------------|------------------------|
| a line | une réplique |
| communication line | ligne de communication |
| cultural line | ligne culturelle |
| dedicated line | ligne spécialisée |

a democratic line	une ligne démocratique
dividing line	ligne de partage
dividing line	ligne de fracture
drain line	canalisation d'écoulement
took a firm line	apporté un soutien d'une fermeté
the following lines	cette formule
the front line	au front
front line	front
hard-liners	l'intransigeance des
in line with	à l'image de
kept in line	on encadre
not in line with	pas correspondre au
In line with	Comme l'indiquait
in line with	s'inscrit dans
into line with	en accord avec
line	positions
our line of conduct	notre conduite
our line	notre principe
the poverty line	le seuil de pauvreté

Given these sets of data, it is possible to map out how the semantic domain of *line* and *ligne* resemble each other and how they differ in terms of semantic extensions and usages. The undertaking of this kind of investigation can play a part in linguistic investigations of grammaticalisation (Hopper and Traugott 1993) and of the study of general constraints on form-meaning mappings (Barlow and Kemmer 1994).

4. Conclusion

These analyses provide an illustration of how the common content of parallel corpora can be exploited to gain linguistic insights into the structure and function of languages. The technique of investigating pairs of languages is promising for a variety of research areas. One advantage is that a two-way analysis of a domain, from language A to language B, and from language B to language A provides clues to the different meanings/uses of each language form.

In sum: in this paper I describe the analysis of parallel texts using ParaConc, a parallel concordancer, and outline some fruitful areas of corpus-based research that are opened up by the use of such a program.

References

- Barlow, M. To appear. Parallel Texts for Linguistic Analysis. In M. Barlow and S. Kemmer (eds) Usage-Based Models of Language.
- Barlow, M. 1995. A Guide to ParaConc. Athelstan: Houston.
- Barlow, M. and S. Kemmer. 1994. A Schema-based Approach to Grammatical Description. In S. Lima, R. Corrigan and G. Iverson (eds) The Reality of Linguistic Rules. Amsterdam: Benjamins.

- Gale, W. and K. Church. 1994. A program for Aligning Sentences in Bilingual Corpora. In S. Armstrong (ed) Using Large Corpora. MIT Press: Cambridge.
- Hopper, P. and E. Closs Traugott. 1993. Grammaticalization. Cambridge: CUP.
- Johansson, S. and K. Hofland. 1993. Towards an English-Norwegian parallel corpus. Paper from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich, May 19-23, 1993. In U. Fries, G. Tottie, and P. Schneider (eds.), Creating and Using English Language Corpora. Rodopi: Amsterdam.
- Kay, M. and M. Roscheisen. 1994. Text-Translation Alignment. In S. Armstrong (ed) Using Large Corpora. MIT Press: Cambridge.
- Moon, R. 1987. The Analysis of Meaning. In J.M. Sinclair (ed) Looking Up. Collins: London.
- Noel, Jacques. 1992. Collocation and Bilingual Text. In G. Leitner (ed) New Directions in English Language Corpora. Mouton de Gruyter: Berlin.