ParaConc: Concordance software for multilingual parallel corpora

Abstract: Parallel concordance software provides a general purpose tool that permits a wide range of investigations of translated texts, from the analysis of bilingual terminology and phraseology to the study of alternative translations of a single text. The software is not tied to particular languages and so can be used with English-Chinese texts, French-Italian texts, and so on. This paper describes the main features of a Windows concordancer, ParaConc, focussing on alignment of parallel (translated) texts, general search procedures, identification of translation equivalents, and the furnishing of basic frequency information. ParaConc accepts up to four parallel texts, which might be four different languages or an original text plus three different translations. A semi-automatic alignment utility is included in the program to prepare texts that are not already pre-aligned. Simple text searches for words or phrases can be performed and the resulting concordance lines can be sorted according to the alphabetical order of the words surrounding the searchword. More complex searches are also possible, including context searches, searches based on regular expressions, and word/part-of-speech searches (assuming that the corpus is tagged for POS). Corpus frequency and collocate frequency information can be obtained.

ParaConc is a new tool designed for researchers who wish to work with translated texts. The software is not tied to particular languages and it can also be used by translators, linguists, teachers, lexicographers and others interested in the analysis of multilingual texts.

1. Alignment

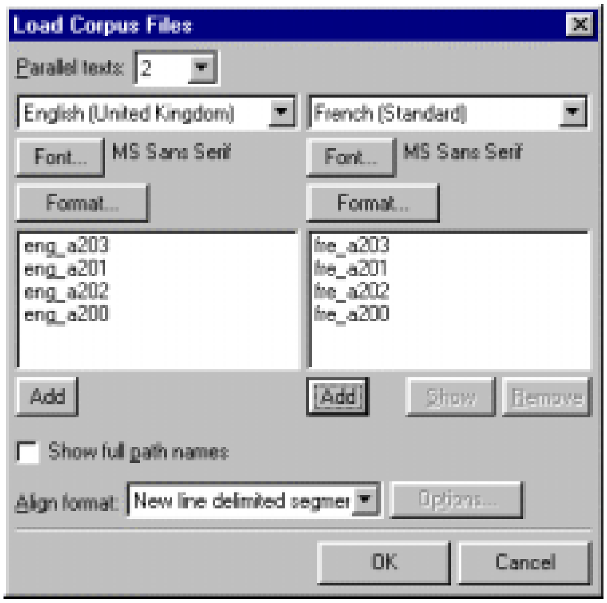
The successful searching and analysis of parallel texts depends on the presence of aligned text segments in each language text. The alignment, an indication of equivalent text segments in the two languages, typically uses the sentence unit as the basic alignment segment, although naturally such an alignment is not one in which each sentence of Language A is always aligned with a sentence of Language B throughout the texts, since occasionally a sentence in Language A may, for example, be equivalent to two sentences in Language B, or perhaps absent from Language B altogether. The size of the aligned segments is not set by the software, however. It would be possible to work with paragraphs as the basic alignment unit, but then the results of a search will be more cumbersome because the translation of a word or phrase will be embedded within a large amount of text, which is especially difficult in cases in which the language is not well-known.

The alignment process is crucial for the successful operation of *ParaConc*. When the program searches through the source text, the only information the program has about the links between the different languages is the alignment. No use is made of bilingual dictionaries or of any kind of language-particular information.

If the parallel texts are pre-aligned, then it will simply be necessary to indicate the way in which the alignment is marked. If the translated texts are not aligned, then the alignment will be carried out by finding equivalent segments first at the level of headings (perhaps with just a single heading for the whole file), then at the paragraph level, and finally at the sentence level. The program allows the user to specify common patterns of marking for these different levels. Sentence level alignment is based on the Gale-Church algorithm. There are further details related to alignment, but we will not pursue them in this paper.

2. Loading a parallel corpus

When the LOAD CORPUS FILE(S) command is given, a dialogue box appears, enabling particular parallel files to be loaded, as shown in Figure 1.



Parallel texts: 2

English (United Kingdom) | French (Standard)

Font... MS Sans Serif | Font... MS Sans Serif

```
eng_a203        fre_a203
eng_a201        fre_a201
eng_a202        fre_a202
eng_a200        fre_a200
```

□ Show full path names

Align format: New line delimited segmen

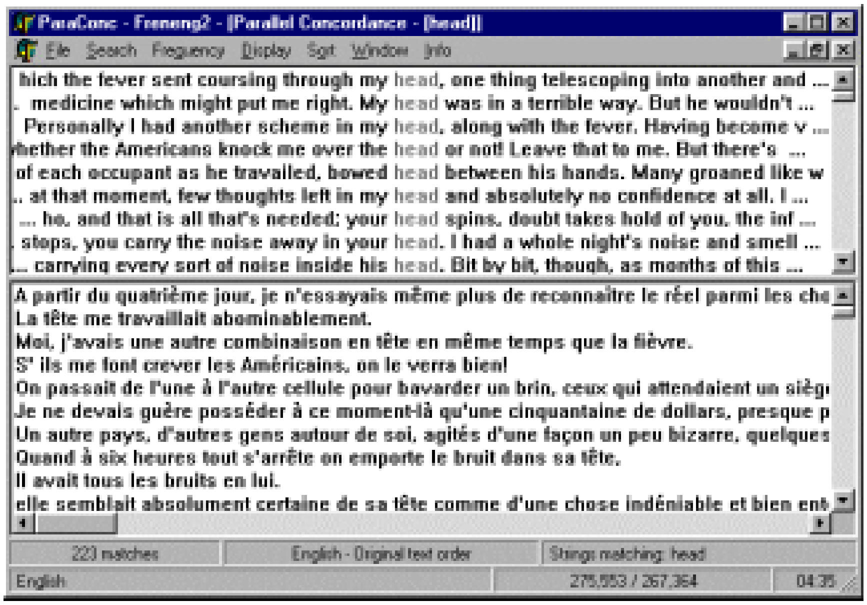
Figure 1. Loading corpus files

The heading PARALLEL TEXTS at the top of the dialogue box is followed by a number in the range 2-4 (i.e, two to four different languages).

3. Searching and analysing the parallel texts

Once a corpus is loaded, some new menu items related to the analysis and display of the text appear on the menu bar. These are FILE, SEARCH, FREQUENCY, and INFO. In addition we can obtain information in the lower left corner relating to the number of the files loaded and in the lower right corner we can find a word count for the two corpora.

To initiate the search, we select SEARCH from the SEARCH menu, or enter CTRL-S. Once the search query has been entered, the program starts to work though the loaded files looking for the search string. Below the results of a search for *head* are illustrated.



```
hich the fever sent coursing through my head, one thing telescoping into another and ...
. medicine which might put me right. My head was in a terrible way. But he wouldn't ...
 Personally I had another scheme in my head, along with the fever. Having become v ...
whether the Americans knock me over the head or not! Leave that to me. But there's ...
of each occupant as he travailed, bowed head between his hands. Many groaned like w
.. at that moment, few thoughts left in my head and absolutely no confidence at all. I ...
... ho, and that is all that's needed; your head spins, doubt takes hold of you, the inf ...
. stops, you carry the noise away in your head. I had a whole night's noise and smell ...
... carrying every sort of noise inside his head. Bit by bit, though, as months of this ...
```

```
A partir du quatrième jour, je n'essayais même plus de reconnaître le réel parmi les che
La tête me travaillait abominablement.
Moi, j'avais une autre combinaison en tête en même temps que la fièvre.
S' ils me font crever les Américains, on le verra bien!
On passait de l'une à l'autre cellule pour bavarder un brin, ceux qui attendaient un siège
Je ne devais guère posséder à ce moment-là qu'une cinquantaine de dollars, presque p
Un autre pays, d'autres gens autour de soi, agités d'une façon un peu bizarre, quelques
Quand à six heures tout s'arrête on emporte le bruit dans sa tête.
Il avait tous les bruits en lui.
elle semblait absolument certaine de sa tête comme d'une chose indéniable et bien ent
```

```
223 matches | English - Original text order | Strings matching: head
English | 275,553 / 267,364 | 04:35
```

Figure 2: The results of a simple search

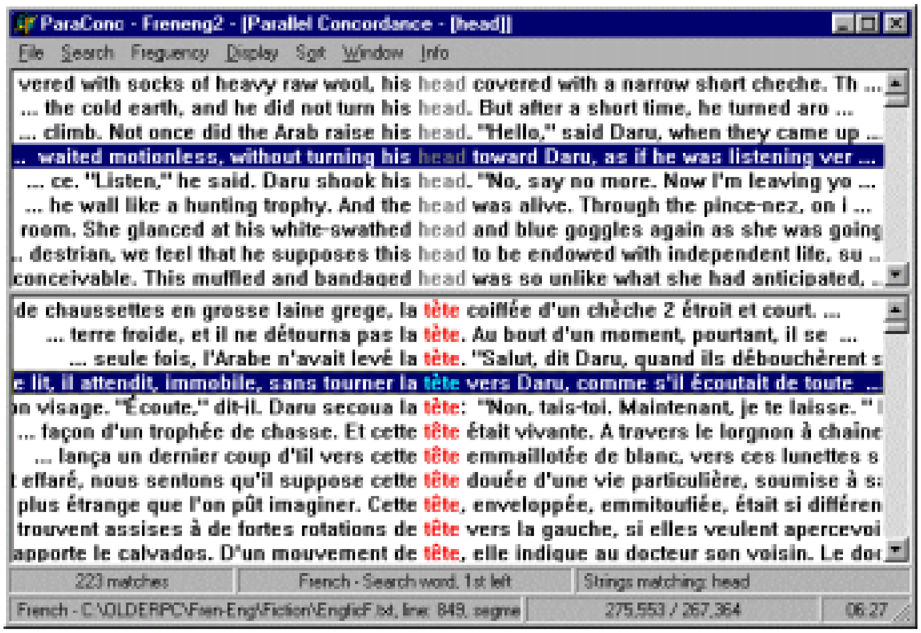
The lower part of the window contains the French sentences (or text segments) that are aligned with the hits displayed in the top window. This display of equivalent units in the two languages is, of course, a consequence of the alignment process. Thus if the first instance of *head* occurred in segment 342 of the English text, then the program simply throws segment 342 of the French text into the lower window, and this process is repeated for all instances of *head*. If you are interested in one particular example of *head* in English, you can click on the appropriate line and both the English and French lines will be highlighted

Let's follow this example further. Once the search is ended, then we can bring to bear the full power of the program to reveal patterns in the results data. You may be interested, for example in different uses (and translations) involving *head: big head, company head, shower head*, etc. One way to find out which words are associated with *head* is to sort the concordance lines so that they are in alphabetical order of the word preceding the search term. The advantage of performing this 'left sort' is that the modifiers (adjectives) of *head* that are the same will occur together. One easy way to achieve this ordering is to select 1ST LEFT, 1ST RIGHT, from the SORT menu.

It can be difficult to locate the position of possible French translations of *head* within each French segment. To alleviate this, we can highlight suggested translations for English *head* by positioning the cursor in the lower French results window and clicking on the right mouse button. A menu pops up and we can select SEARCH QUERY, which gives access to the usual search commands and hence allows us to enter a possible translation of *head* such as *tête*. The program then simply highlights all instances of *tête* in the French results window.

We can now change the context for the French results so that the results in the lower window are transformed into a KWIC layout (at least for those segments containing *tête.)* First, we make sure that the lower window is active. Next we choose CONTEXT TYPE from the DISPLAY menu and select WORDS. Finally, we rearrange the lines to bring those segments containing *tête* together at the top of the French results window. To achieve this, choose SORT and sort the lines by searchword, and 1st left. The sorting procedure will now rearrange the results in lower window since the sort commands are applied to whichever window is active. The two text windows then appear as shown in Figure 3. Naturally, only those words in the French text that have been selected and highlighted can be displayed in this way. By sorting on the searchword, all the KWIC lines are grouped together at the top of the text window; the residue can be found by scrolling through towards the bottom of the window. This is a revealing display, but we have to avoid being misled by the dual KWIC displays. There is no guarantee that for any particular line, the instance of *tête* is the translation of *head*. It could be accidental that *tête* is found in the sentence.

The idea behind this feature of *ParaConc* is to let the user move from English to French and back again, sorting the concordance lines, and inspecting the results to get a sense of the connections between the two languages at whatever level is relevant for a particular analysis.



```
vered with socks of heavy raw wool, his head covered with a narrow short cheche. Th ...
... the cold earth, and he did not turn his head. But after a short time, he turned aro ...
... climb. Not once did the Arab raise his head. "Hello," said Daru, when they came up ...
..  waited motionless, without turning his head toward Daru, as if he was listening ver ...
 ... ce. "Listen," he said. Daru shook his head. "No, say no more. Now I'm leaving yo ...
 ... he wall like a hunting trophy. And the head was alive. Through the pince-nez, on i ...
 room. She glanced at his white-swathed head and blue goggles again as she was going
.. destrian, we feel that he supposes this head to be endowed with independent life, su ..
conceivable. This muffled and bandaged head was so unlike what she had anticipated. ..
```

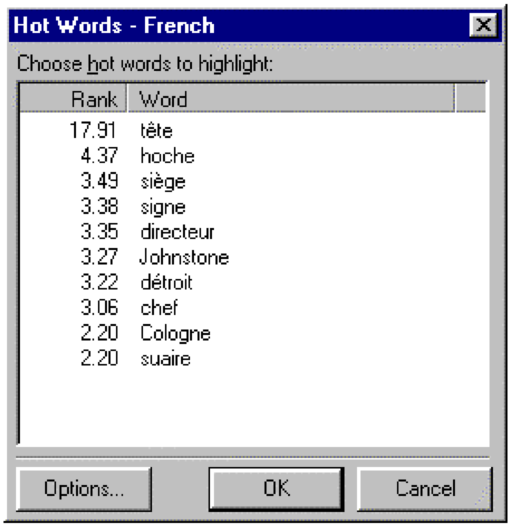
```
de chaussettes en grosse laine grege, la tète coiffée d'un chèche 2 étroit et court. ...
     ... terre froide, et il ne détourna pas la tète. Au bout d'un moment, pourtant, il se  ...
        ... seule fois, l'Arabe n'avait levé la tète. "Salut, dit Daru, quand ils débouchèrent s
e lit, il attendit, immobile, sans tourner la tète vers Daru, comme s'il écoutait de toute  ...
n visage. "Écoute," dit-il. Daru secoua la tète:  "Non, tais-toi. Maintenant, je te laisse. " l
 ... façon d'un trophée de chasse. Et cette tête était vivante. A travers le lorgnon à chaine
   ... lança un dernier coup d'Iil vers cette tête emmaillotée de blanc, vers ces lunettes s
t effaré, nous sentons qu'il suppose cette tête douée d'une vie particulière, soumise à s:
 plus étrange que l'on pût imaginer. Cette tête, enveloppée, emmitouflée, était si différen
 trouvent assises à de fortes rotations de tête vers la gauche, si elles veulent apercevoi
apporte le calvados. D'un mouvement de tête, elle indique au docteur son voisin. Le doc
```

```
223 matches | French - Search word, 1st left | Strings matching: head
French - C:\OLDERPC\Fren-Eng\Fiction\EnglicF.txt, line: 849, segme | 275,553 / 267,364 | 06:27
```

Figure 3: Parallel KWIC displays

4. Hot Words

In the previous section, we described the use of SEARCH QUERY to locate possible translations in the second window. In this section we will look at a utility in which possible translations and other associated words (collocates) are suggested by the program itself. We will call these words: hotwords. First we position the cursor in the lower (French) half of the results window and click using the right mouse button. If we used SEARCH QUERY earlier, we need to select CLEAR SEARCH QUERY and then we can choose HOT WORDS, which brings up a dialogue box containing a ranked list of hot words. The ranked list of candidates for hot words are displayed as shown in Figure 4.



Choose hot words to highlight:

Rank	Word
17.91	tête
4.37	hoche
3.49	siège
3.38	signe
3.35	directeur
3.27	Johnstone
3.22	détroit
3.06	chef
2.20	Cologne
2.20	suaire

Figure 4: Hot Word List

Some or all the words can be selected. When the list of selected words is complete, click on OK. The words will be highlighted in the results and can again be sorted.

5. Frequency information

ParaConc furnishes a variety of frequency statistics, but the two main kinds are corpus frequency and collocate frequency. The command CORPUS FREQUENCY DATA creates a word list for the whole corpus (or corpora). Thus to find out the distribution of words in the entire corpus, choose FREQUENCY ORDER or ALPHABETICAL ORDER from the FREQUENCY menu. Again, it is possible to obtain the frequency information for one particular language corpus or for all the language corpora.

Choosing COLLOCATE FREQUENCY DATA from the FREQUENCY menu (or CTRL-F) displays the collocates of the search term ranked in terms of frequency. In *ParaConc,* the collocate frequency calculations are tied to a particular search word and so the frequency menu only appears once a search has been performed. The collocation data produced by the COLLOCATE FREQUENCY DATA command is organised in four columns, with one column for each position surrounding the keyword: 2nd left, 1st left, 1st right and 2nd right. (Thus 1st left refers to the word before the search term and 1st right refers to the word following the search term.) The columns show the collocates in descending order of raw frequency.

Returning to the earlier example of a search for *head* in English (and an associated search for *tête* in French), we can see the words commonly occurring with the search word by scrolling through the concordance lines. Conveniently, we can make the program display the frequency of particular collocates for both *head* and *tête* or either one by selecting COLLOCATE FREQUENCY DATA from the FREQUENCY menu and choosing ALL, ENGLISH or FRENCH. The program then calculates, for the language chosen the frequency of collocates surrounding the search term for a span of the four positions (from 2nd left to 2nd right). One disadvantage of the simple collocate frequency table is that it is not possible to gauge the frequency of collocations consisting of three or more words. To calculate the frequency of three word collocations, it is necessary to choose ADVANCED COLLOCATION from the FREQUENCY menu and select one or more languages. The top part of the dialogue box associated with ADVANCED COLLOCATION allows the user to choose from up to three word positions, for example, SEARCHWORD 1ST RIGHT, 2ND RIGHT.

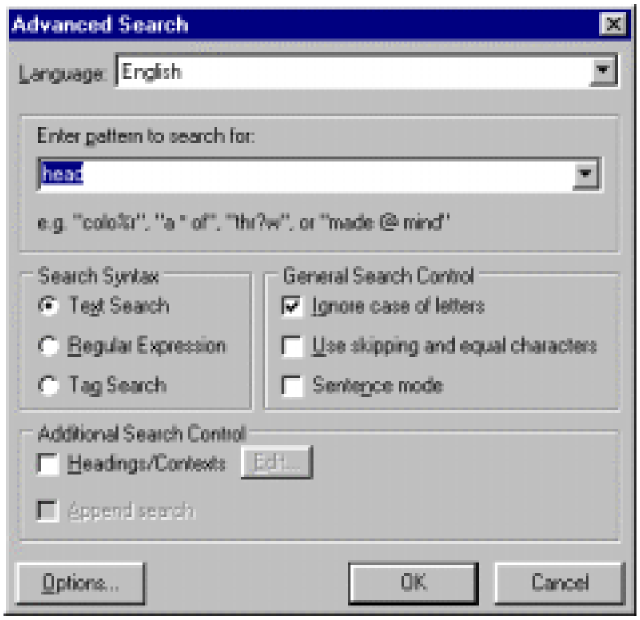
6. Workspace

The loading and processing of a parallel corpus can take some time since the program has to record alignment data and tag data. (The latter is not dealt with in this article.) Since the same sets of corpus files are often loaded each time *ParaConc* is started, it makes sense to freeze the current state of the program, at will, and return to that state at any time. This is the idea behind a workspace. A workspace is saved as a special ParaConc Workspace file (.pws), which can then be opened at any time to restore *ParaConc* to its previous state, with the corpus loaded ready for searching. Searches and frequency data are, however, not included in the saved workspace. (Only the search histories are saved.)

A workspace—the current corpus and settings of *ParaConc*—can be saved at any time by selecting the command SAVE WORKSPACE or SAVE WORKSPACE AS from the FILE menu. The usual dialogue box appears and the name and location of the workspace file can be specified in the normal way. Once a (memorable, descriptive) filename for the saved workspace has been entered, the user is asked to choose some different workspace options. The line/page and the tracked tag info can be saved as part of the workspace. (The saved workspace consists of a saved file and an associated folder of the same name.)

7. Advanced Search

The simple searches described in Section 3 will suffice for many purposes and are especially useful for exploratory searches. The basic TEXT SEARCH is also very useful when used in conjunction with a sort-and-delete strategy. Particular sort configurations can be chosen to cluster unwanted examples (words preceded by *a* and *the* perhaps), which can then be selected and deleted. For more complex searches, however, we need to use the ADVANCED SEARCH command. This command brings up a more intricate dialogue box (displayed in Figure 5), which at the top contains the text box in which the search query is entered.



Advanced Search

Language: English

Enter pattern to search for:

head

e.g. "colo%r", "a * of", "thr?w", or "made @ mind"

Search Syntax

● Text Search

○ Regular Expression

○ Tag Search

General Search Control

☑ Ignore case of letters

□ Use skipping and equal characters

□ Sentence mode

Additional Search Control

□ Headings/Contexts

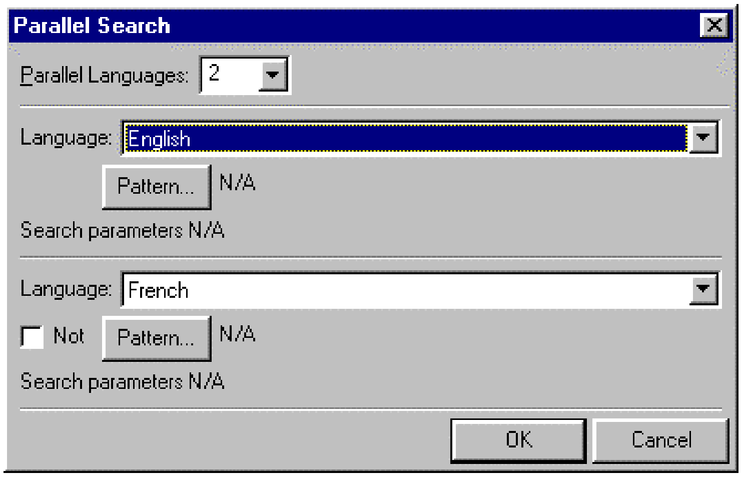
□ Append search

Figure 5: Advanced Search

The most important part of the ADVANCED SEARCH dialogue box is labelled SEARCH SYNTAX,. The three radio buttons allow users to specify the kind of search we wish to perform. The first, TEXT SEARCH, refers to the basic searches described in the section above. The REGULAR EXPRESSION search allows for search queries containing boolean operators (AND, OR and NOT). For example, a regular expression to capture the *speak* lemma might be given as **sp[eo]a?k**. This expression will match the string *sp* followed by *e* or *o*, an optional *a* and finally *k*. The software also supports the expanded set of regex metacharacters: \d, \w, \s, \S, etc. The third option in the advanced search dialogue box is TAG SEARCH, which allows the user to specify a search query consisting of a combination of words and part-of-speech tags, with the special symbol **&** being used to separate words from tags in the search query. This search syntax is used whatever particular tag symbols are used in the corpus. (Thus it is necessary to enter the form of the tags in TAG SETTINGS before a tag search can be performed.) To give an example: the search string **that&DD** finds instances of *that* tagged as a demonstrative pronoun, which may appear in the corpus as *that<w DD>*. Similarly, a tag search for **&JJ of&** will find all instances of adjectives followed by the word *of*.

The dialogue box in Figure 5 contains a variety of other options, which will not be discussed in this brief paper.

Finally, one kind of search tailored for use with parallel texts is a parallel search, which is one of the options within the SEARCH menu. This type of search, shown in Figure 6, allows you to constrain a search based on occurrences in the different parallel texts.



Parallel Search

Parallel Languages: 2

Language: English

Pattern... N/A

Search parameters N/A

Language: French

☐ Not Pattern... N/A

Search parameters N/A

Figure 6: Parallel search

Clicking on the Pattern box under Language: English brings up the normal advanced search dialogue box and a search query can be entered. In this case, the search term **head** has been entered. Moving to Language: French and again clicking on Pattern, it is possible to enter another search string such as **tête**. If we click OK, the search routine will look for examples in which *head* occurs in the English and *tête* is also found in the corresponding French segment. If the NOT box is selected, then the search routine will display *head* only if *tête* does not occur in the equivalent French segment.

8. Summary

This paper has provided a brief overview of a Windows parallel concordance program which can be used by a variety of professionals working on the analysis of multilingual texts for translation or linguistic purposes.