Displaying Bidirectional Text Concordances in KWIC format

Pavel Rychlý, Vojtěch Kovář

Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{pary,xkovar3}@fi.muni.cz

Abstract. The concordance view is one of the main functions of a corpus query system. Usually, it is displayed in the "Key Word(s) In Context" (KWIC) format. It is a table that contains a keyword (or keywords) in its middle column and the left and right context - words that girdle the keyword - in the other two columns. The concordance view enables corpus users to see all possible occurrences of the keyword easily.

Other important task for corpus query system is to deal with as many languages as possible. Documents in corpora can also be bilingual or multilingual, e.g. many documents in any language contain English expressions. By handling languages with right-to-left script, e.g. Arabic or Persian, these bilingual documents are usually bidirectional. Since the data is stored in logical order, there is a problem how to display the correct word order in the concordance view. The current corpus query systems can not handle this task well.

In the paper, we describe the problem of displaying bidirectional texts in the word concordance view and introduce a system that can handle these texts. A few examples of English word sequences in a corpus of Persian are given. We describe display algorithms and corpus input file modifications needed to achieve the correct word order in the concordance view. We also discuss some related problems, e.g. working with neutral characters (like punctuation or numbers) and the recognition of the left-to-right (right-to-left) text boundaries.

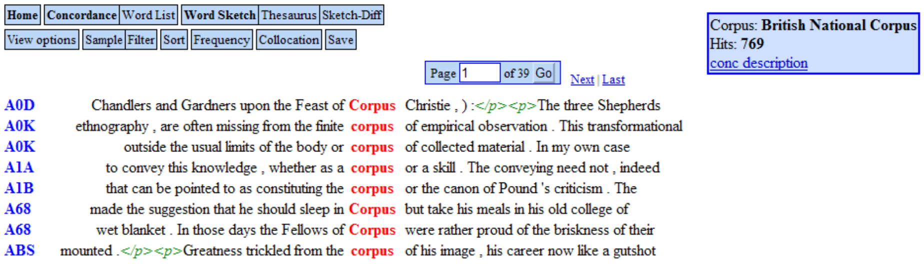
1 Introduction

At the present time, large text corpora form an important source of liguistic information. They are used for a wide variety of tasks, e.g. language learning and teaching, testing of automatic text processing tools, discovering of real words behaviour and many more lingustic research purposes. As the corpus linguistics becomes popular in many countries, the large text corpora are available for more and more languages. A corpus containing two or more languages together is also no exception.

Corpus query systems (or corpus managers) are programs that enable people to work comfortably with large data in text corpora. One of their basic functions is called KWIC (Key Word(s) In Context) view. The KWIC format displays search query result (the keywords) within its close context, one occurence per line, so that it enables users to see all occurences of the keywords in a simple table structure (see Figure 1).

The claims on corpus query systems are growing up, according to the number of languages that can occur in the corpora and to the number of users working with these resources. Nowadays, corpus managers have to deal with many different language phenomena, such as different scripts (English vs. Arabic), tokenization or encodings. At Masaryk University, a corpus manager Manatee/Bonito [1] have been developed, that is able to perform wide variety of tasks including e.g. handling of different scripts or computing word sketches, thesaurus and many more statistical characteristics.

In this paper, we describe a problem of handling texts containing two languages with different scripts, especially in context of displaying these texts in the KWIC format. We will introduce our solution used to display correct word order in these bidirectional texts and we will also discuss some related problems that we had to deal with during the implementation.



Corpus: **British National Corpus**
Hits: **769**
conc description

A0D	Chandlers and Gardners upon the Feast of	Corpus	Christie ,) :</p><p>The three Shepherds
A0K	ethnography , are often missing from the finite	corpus	of empirical observation . This transformational
A0K	outside the usual limits of the body or	corpus	of collected material . In my own case
A1A	to convey this knowledge , whether as a	corpus	or a skill . The conveying need not , indeed
A1B	that can be pointed to as constituting the	corpus	or the canon of Pound 's criticism . The
A68	made the suggestion that he should sleep in	Corpus	but take his meals in his old college of
A68	wet blanket . In those days the Fellows of	Corpus	were rather proud of the briskness of their
ABS	mounted .</p><p>Greatness trickled from the	corpus	of his image , his career now like a gutshot

Fig. 1. Illustration of the KWIC view in the Manatee/Bonito system

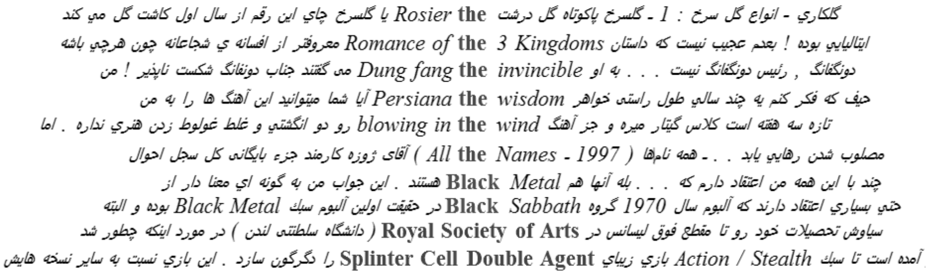
2 The Problem of Displaying Bidirectional Texts

By building a web corpus of Persian, we found out that the resulting text often contains sequences of English words. This fact was not caused by a flaw in data preprocessing, the English sequences were full-value parts of the text (see Figure 2).

Since Persian uses Arabic script, this is a typical example of a bidirectional text. Because words in corpus input files are stored in logical order, we realized very soon that there is a problem how to display the data in KWIC format.

We are using UTF-8 [2] encoding for non-Latin script corpora, UTF-8 is a part of the Unicode standard [3] which provides generic encoding for all known characters. The Unicode standard also defines character properties like 'Latin capital letter', 'Arabic letter' etc., and the bidirectional algorithm [4] for displaying the right word order of a bidirectional text. This algorithm cannot be used directly if a text is divided into several parts.

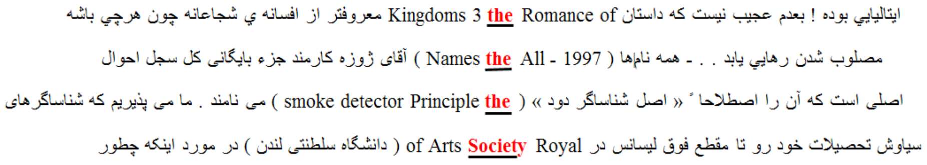
For KWIC view, the displayed text is divided into 3 groups: the keyword(s), its left context and its right context. When KWIC is displayed without modifications, the word



گلكاري - انواع گل سرخ : 1 - گلسرخ پاكوتاه گل درشت Rosier **the** يا گلسرخ چاي اين رقم از سال اول كاشت گل مي كند
ايتاليايي بوده ! بعدم عجيب نيست كه داستان Romance of **the** 3 Kingdoms معروفتر از افسانه ي شجاعانه چون هرچي باشه
دونگفانگ , رئيس دونگفانگ نيست . . . به او Dung fang **the** invincible مى گفتند جناب دونفانگ شكست ناپذير ! من
حيف كه فكر كنم يه چند سالي طول راستى خواهر Persiana **the** wisdom آيا شما ميتوانيد اين آهنگ ها را به من
تازه سه هفته است كلاس گيتار ميره و جز آهنگ blowing in **the** wind رو دو انگشتي و غلط غولوط زدن هنري نداره . اما
مصلوب شدن رهايي يابد . . . - همه نام‌ها (All **the** Names - 1997) آقاى ژوزه كارمند جزء بايگانى كل سجل احوال
چند با اين همه من اعتقاد دارم كه . . . بله آنها هم **Black** Metal هستند . اين جواب من به گونه اي معنا دار از
حتي بسياري اعتقاد دارند كه آلبوم سال 1970 گروه **Black** Sabbath در حقيقت اولين آلبوم سبك Black Metal بوده و البته
سياوش تحصيلات خود رو تا مقطع فوق ليسانس در **Royal Society of Arts** (دانشگاه سلطنتى لندن) در مورد اينكه چطور شد
آمده است تا سبك Action / Stealth بازي زيباي **Splinter Cell Double Agent** را دگرگون سازد . اين بازي نسبت به ساير نسخه هايش

Fig. 2. Examples of English word sequences in Persian web corpus

order is incorrect, as shown in the Figure 3. The reason of this behaviour is the fact that the groups boundaries are defined on the level of logical word order. For displaying the correct word order in the KWIC view, we have to re-arrange the words among the 3 groups, as illustrated in the Figure 4. In the next sections, we describe how to implement this re-arranging to achieve the correct word order display.



ايتاليايي بوده ! بعدم عجيب نيست كه داستان Kingdoms 3 **the** Romance of معروفتر از افسانه ي شجاعانه چون هرچي باشه
مصلوب شدن رهايي يابد . . . - همه نام‌ها (Names **the** All - 1997) آقاى ژوزه كارمند جزء بايگانى كل سجل احوال
اصلى است كه آن را اصطلاحاً « اصل شناساگر دود » (smoke detector Principle **the**) مى نامند . ما مى پذيريم كه شناساگرهاى
سياوش تحصيلات خود رو تا مقطع فوق ليسانس در of Arts **Society** Royal (دانشگاه سلطنتى لندن) در مورد اينكه چطور

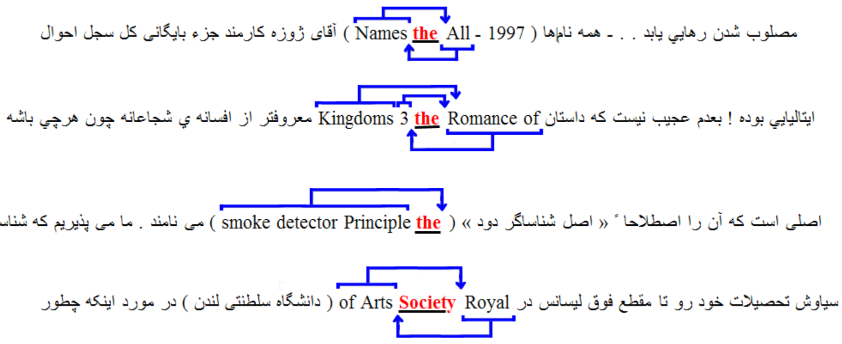
Fig. 3. Incorrect word order in KWIC view of bidirectional texts ...

3 Corpus Input File Modifications

As a first step, we needed to distinguish word sequences in standard corpus script from sequences with the opposite text direction. This task was implemented on the level of corpus input file.

In our system, the corpus input file is in so-called vertical text format. It is a text file formatted one token per line (each line can also contain additional information about relevant token, such as lemma, tag, etc.) enriched by XML-like tagging.

Into this vertical file, we placed additional tags that represented boundaries between different text directions. In case of Persian web corpus, we named these tags as “<ltr>” that stands for “left to right” since default Persian text has the opposite direction. The vertical file modification is illustrated in the Figure 5.



مصلوب شدن رهايي يابد . . - همه نام‌ها (Names the All - 1997) آقای ژوزه کارمند جزء بایگانی کل سجل احوال

ايتاليايي بوده ! بعدم عجيب نيست كه داستان Kingdoms 3 the Romance of معروفتر از افسانه ي شجاعانه چون هرچي باشه

اصلی است که آن را اصطلاحا ً « اصل شناساگر دود » (smoke detector Principle the) می نامند . ما می پذیریم که شناساگرهای

سیاوش تحصیلات خود رو تا مقطع فوق لیسانس در of Arts Society Royal (دانشگاه سلطنتی لندن) در مورد اینکه چطور

Fig. 4. ... and illustration of its re-arranging

By placing the “<ltr>” tags into vertical file automatically, we encountered several complications, described in the Section 5. However, they were satisfactorily resolved so that we could further work with a ”<ltr>-tagged” vetical file as described above.

4 Display Algorithms

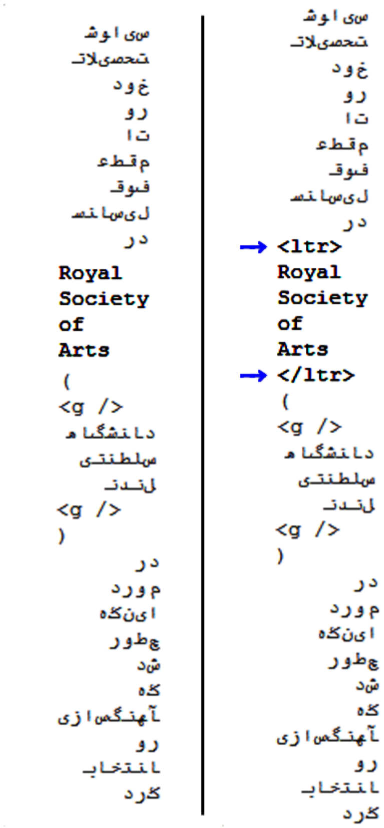
Users use a regular web browser to display a concordance, so a web browser is used to display concordance text snippets. Web browsers have implemented the above mentioned bidirectional algorithm and it works correctly for whole sentences. The only remaining task for the Sketch Engine is to provide the right logical word order of each part of a concordance line and to annotate the right-to-left direction of text in respective text areas.

The algorithm works with three lists of tokens (words), one for each group defined in section 2: *keywords*, *left-context*, *right-context*. The word reordering is done in several steps.

1. if *keywords* doesn’t contain left-to-right (Latin) characters, no reordering is needed and only swapping of *left-context* and *right-context* is done
2. find a sequence of left-to-right (Latin) tokens from the end of *left-context* and strip it off from *left-context*
3. find a sequence of left-to-right (Latin) tokens from the beginning of *right-context* and strip it off from *right-context*
4. move stripped sequences to the opposite side (beginning or end) of the opposite *context*.

5 Related Problems

As mentioned in the Section 3, we encountered several complications by placing the “<ltr>” tags into vertical file automatically. The main problem was in neutral charac-



```
سیاوش
تحصیلات
خود
رو
تا
مقطع
فوق
لیسانس
در
Royal
Society
of
Arts
(
<g />
دانشگاه
سلطنتی
لندن
<g />
)
در
مورد
اینکه
چطور
شد
که
آهنگسازی
رو
انتخاب
کرد
```

```
سیاوش
تحصیلات
خود
رو
تا
مقطع
فوق
لیسانس
در
→ <ltr>
Royal
Society
of
Arts
→ </ltr>
(
<g />
دانشگاه
سلطنتی
لندن
<g />
)
در
مورد
اینکه
چطور
شد
که
آهنگسازی
رو
انتخاب
کرد
```

Fig. 5. A part of the vertical file before and after “<ltr>” tags addition

ters handling, i.e. deciding whether the neutral word (consisting of neutral characters) belongs to the left-to-right part of the text or not. The most frequent (and most problematic) neutral characters were digits and punctuations.

We solved this problem by a set of heuristic rules, that showed to be sufficient for a good <ltr>-tagging of the corpus. Firstly, all words were tagged according to characters they contained, neutral words were marked according to the major (default) corpus script. Secondly, a set of modification rules was applied. The rules are quite simple and look similar to the following two examples:

- `if` a group of neutral words lies directly between two <ltr> sections
  `then` mark these neutral words as <ltr> and join the two sections into one
- `if` the left bracket ("(", "[", ...) lies directly in front of <ltr> section
  `then` mark it as <ltr> and join it with the adjacent section

6 Conclusions and Future Directions

In the paper, we have described the problem of displaying bidirectional texts in KWIC format that is used by corpus query systems to provide comfortable work with a huge amount of text data. We have introduced a technique used to achieve correct word order display in these texts. We have discussed the corpus file modifications and display algorithms as well as problems related to the implementation. The described procedures were practically implemented within the scope of the corpus query system Manatee/Bonito.

In the future development, we want to continue in enhancing multilingual features of our system. We believe that the "language flexibility" of such a system should grow up according to the variety of users needs.

Acknowledgements

This work has been partly supported by the Academy of Sciences of Czech Republic under the projects T100300414, T100300419 and by the Ministry of Education of CR within the Center of basic research LC536 and National Research Programme 2C06009.

References

1. Rychlý, P., Smrž, P.: Manatee, Bonito and Word Sketches for Czech. In: Proceedings of the Second International Conference on Corpus Linguisitcs, Saint-Petersburg, Saint-Petersburg State University Press (2004) 124–132
2. Yergeau, F.: RFC2279: UTF-8, a transformation format of ISO 10646. Internet RFCs (1998)
3. Consortium, U.: The Unicode Standard, Version 5.0. Addison-Wesley Professional (2006)
4. Davis, M.: The Bidirectional Algorithm. Standard Annex UAX 9, Unicode Consortium (2006)