Pre-corpus processing of texts (on the material of Vasil Shklar's work «Marusia»)

Khrystyna Pstrak

Applied Linguistics Department, Lviv Pylytechnic National University, Ukraine, Lviv, Stepana Bandery Street 79013, E-mail: pskhrystyna@gmail.com

Abstract – The approach to pre-corpus processing of texts on the example of works of Ukrainian writers is offered.

Keywords – pre-corpus processing, index, Xpaths, XQL, requests, XML, corpus, text

Introduction

Pre-corpus studies related to the collection, organization, and description of linguistic material. This kind of processing of Ukrainian-language texts is a priority direction. The problem of pre-corpus processing of texts was investigated by Karunesh Arora, Shyam, S. Agrawal [1], Chu Tao Zheng, Cheng Liu, Hau San Wong [2], Jasmeet Singh, Vishal Gupta [3], among domestic researchers - Ihor Kulchytskyi [4].

Main part of the research

An index reflects the occurrence of index terms prior to a location in a corpus, which can be identified in several ways, in addition to an internally-defined location system. This xref scheme is used by the system to indicate the context of cases found by the search program. The cases themselves are precisely arranged according to the scheme of internal arrangement. Although the index contains information about the full location of XPath cases in the corpus, the internal layout scheme is highly optimized and cannot be used to support access through Xpaths or XQL queries.

The reference schema used to identify contexts has the following components:

- single "text" code: it can be derived from a system identifier or specified by the nominated attribute on an element that contains text, or it can be calculated by an indexer behind the XML structure indexing.
- single identifier "scope": this can be derived from the value of the specified attribute on any text element; calculated by an indexer in terms of XML structure; or derived from the physical structure of the input data.
- selectively additional unit labels: these can be derived from the value the specified attribute on any element in the text; calculated an indexer in terms of an XML structure; or derived from a physical line the entrance.

The element from which the text identifier derives also limits one "text" in the corpus. This effectively limits the kinds of values that can be used to identify it: it must be an attribute value or pseudo-value; a content element is not allowed.

The reference specification for the xaira index is assigned a <xairaList type = "refSpec" > containing exactly one <xairaItem type = "textRef">, and then one <xairaItem type = " scopeRef "> and possibly one or more additional <xairaItem type = "unitRef" > elements. Each such <xairaItem> element contains a <valSource> element, as defined above, to indicate where the reference value is to be obtained in the incoming document. It can also contain a <labelGen> element that further defines the parts of the document referenced by the help and its format (Fig. 1)

```
  </xairaItem>
  <xairaItem type="scopeRef">
    <valSource type="attribute" ident="n">
      <nameList>
        <gi>s</gi>
      </nameList>
      <labelGen>%1.%2</labelGen>
    </valSource>
  </xairaItem>
</xairaList>
<xairaList type="refSpec">
  <xairaItem type="textRef">
    <valSource
      type="attribute"
      ident="id"
      ns="http://www.w3.org/XML/1998/namespace">
      <nameList>
        <gi>bncDoc</gi>
      </nameList>
    </valSource>
```

Fig. 1. The reference specification

In BNC, each <bncDoc> starts a new "text" that is identified by the value of its xml: id attribute, and the scope for each query is the complete <s> element identified by its n attribute. the Reference must be formatted with a dot between these two values.

This specification will give a reference to ABC.123 for the <s> element with the n attribute set to 123 found in the <bncDoc> element for which the xml: id attribute is set to ABC.

In addition to the index terms derived from the lexical content of the corpus is the Khaira index. It also provides information about the occurrence of XML start and end tags in the corpus. This information is used to facilitate a number of search parameters: to search for non-lexical features, to search for lexical functions in a particular structural context, to discuss common lexical or non-lexical features, and the like.

By default, a record is created in the index for each event of each tag, both beginning and ending. This entry can also distinguish the start tag events based on the values of the specified attributes that come with them. (Note that this does not depend on using this attribute value when creating index terms, as described in the previous section.)

The contents of each element found in the corpus are indexed by default, as are all labels and all their attributes. You can change this behavior by specifying explicit identification rules for items to which these default rules do not apply.

Within research, I encountered several problems. First of all, it was difficult to arrange the markings so that the different markings do not intersect. So as in epic Vasil shklyara "Marusya" very many replicas direct speech, then I well learned importantly rule: under rozznachenni should remember that no amount benchmarks should not perehreschuvatis, and be only nested one in one:

- correct - < tag1> ... <tag2> ... <tag3> ... < / tag3> ... <tag3> ... < / tag3> ... < / tag2> ... <tag2 > ... <tag3> ... < / tag3 > ... <tag3 > ... < / tag3> ... < / tag2 > ... < / tag1>;

- incorrect - <tag1> ... <tag3> ... <tag2> ... </tag3> ... <tag3> ... </tag2> ... </tag3> ... <tag2> ... <tag3> …

Much attention was paid to the rules of marking epigraphs as in the resulting part of the text they were two.

Some difficulties arose when I started to mark the pages, as in this task you need to be especially careful, checking the text with the original paper version of the book "Marusya". Example of text processing:

<s>Євген Васильович Соколовський сторожував колгоспний садок, коли прибігла дочка Ліза й сказала, що приїхали якісь чужі люди на чорній машині й кличуть його додому.</s>

<s><q>— Кличуть — то треба йти,</q>— зітхнув Євген Васильович.</s><q><s> — А ти побудь тут в апашнику 1.</s><s> Я туди й назад.</s></q>

<s><q>— Ні, тату, я піду з вами, </q> — сказала Ліза так гостро, аж йому пробіг холодок у грудях.</s><s> Донька стала зовсім дорослою після того, як померла бабуся Надя.</s><s> Перед тим стареньку забрали в тюрму, і, хоч тримали її там недовго, повернулась додому такою, що страшно було дивитися.</s><s> Жила вона через дорогу від них у хатині, зробленій із повітки, але Лізі ніде не було так добре, як у бабуні Наді.</s><s> Найдужче вона любила розчісувати їй волосся, бо тоді бабуня розповідала такі історії, від яких перехоплювало дух.</s><s><q>«Скажу тобі, Лізко, по секрету,</q> — однаково починала вона чи не кожну бувальщину, а далі справді говорила таке, чого Ліза не чула більш ні від кого.</s><s><q>— А знаєш, що ти й не Лізою мала бути,</q>— казала бабуня Надя.</s><q><s>— Спершу дали тобі інакше ім’я, а записали отак…</s><s> Але ти будеш щасливою.</s><s> Скажу тобі по секрету.

Conclusion

Thus, the approach of pre-corpus processing of texts was applied. A significant contribution is the processing of Ukrainian-language texts using this approach.

References

[1] Karunesh, A., Shyam, S. A. (2018). Pre-processing English-Hindi Corpus for Statistical Machine Translation. *Computacion y Sistemas*, 21, 725–737.

[2] Chu, T. Z., Cheng, L., Hau, S. W. (2018). Corpus-based topic diffusion for short text clustering. *Neurocomputing*, 275, 2444-2458.

[3] Jasmeet, S., Vishal, G. (2019). A novel unsupervised corpus-based stemming technique using lexicon and corpus statistics. *Knowledge-Based Systems*, 180, 147-162.

[4] Kulchytskyi, Ihor (2018). Tekhnolohichni aspekty dokorpusnoho opratsiuvannia tekstiv. *Mizhnarodna naukova konferentsiia IX Olomoutskyi sympozium ukrainistiv*, 75–80.