Vertical and registry files

1) Формат registry-файла

Registry-файл содержит информацию о корпусе, его структуре и атрибутах данных.

При указании информации о корпусе необходимо использовать имеющийся перечень возможных атрибутов, таких как NAME (имя корпуса), INFOHREF (ссылка на информацию о корпусе для пользователей), PATH (путь к папке, где лежит корпус), VERTICAL (путь к папке, где лежит vertical-файл), ENCODING (кодировка корпуса), RIGHTTOLEFT (информации о направлении письма) и множества других[1].

Пример:

```
NAME      "Russian National Corpus"
INFO      "Russian National Corpus (6 mln. tokens)"
INFOHREF    "http://hsecompling.wikispaces.com/visual"
TAGSETDOC   "http://nl.ijs.si/ME/V4/msd/html/msd-ru.html"
PATH      /corpora/rnc_d
VERTICAL    /corpora/rnc_d/vert/rnc_d.vert
LANGUAGE    "Russian"
ENCODING    utf-8
LOCALE          ru_RU.UTF-8
MAXCONTEXT  0
MAXDETAIL   0
```

Далее задаются атрибуты, т.е. параметры для самого токена, леммы, pos-тега и любых других имеющихся в корпусе тегов. Обязательным является указание параметров одного атрибута -- токена. На данном этапе также можно определить, может ли атрибут иметь несколько значений (например, две лемм или pos-тегов в случае омонимии) с помощью параметра MULTIVALUE (поставить значения “yes”) и параметра MULTISEP, который определяет разделитель возможных вариантов значения атрибута (например, "|" или ";").

Имеется возможность задавать динамические атрибуты, которые значительно увеличивают скорость поиска по корпусу. Чаще всего используются динамические атрибуты “lc” и “lemma_lc”, который представляют собой приведенные к нижнему регистру токен и лемму соответственно[2].

---

[1] Полный перечень атрибутов см. по ссылке https://www.sketchengine.co.uk/corpus-configuration-file-all-features/#Attributes

[2] Побробнее о динамических атрибутах см. по ссылкам https://www.sketchengine.co.uk/corpus-configuration-file-all-features/#Dynamicattributes, https://www.sketchengine.co.uk/dynamic-attributes/

Пример определения атрибутов и их параметров:

```
ATTRIBUTE word {
   TYPE "FD_FGD"
}
ATTRIBUTE lemma {
   TYPE "FD_FGD"
   MULTIVALUE yes
   MULTISEP "|"
}
ATTRIBUTE tag {
   TYPE "FD_FGD"
   MULTIVALUE yes
   MULTISEP "|"
}
ATTRIBUTE lc {
   LABEL "word (lowercase)"
   DYNAMIC utf8lowercase
   DYNLIB internal
   ARG1 "C"
   FUNTYPE s
   FROMATTR word
   TYPE index
   TRANSQUERY yes
}
ATTRIBUTE lemma_lc {
   LABEL "lemma (lowercase)"
   DYNAMIC utf8lowercase
   DYNLIB internal
   ARG1 "C"
   FUNTYPE s
   FROMATTR lemma
   TYPE index
   TRANSQUERY yes
}
```

Последнее, что указывается в registry-файле, это структурные сегменты корпуса (границы предложения, заголовков, курсивное выделение и др.) и их параметры, а также метаинформация каждого документа, если таковая имеется (минимальным является указание номера или любого id документа). Имена поля структуры (doc, s, p, g и др.) соответствуют именам тегов vertical-файла. Минимальный набор структур

включает в себя следующие теги: doc (границы документа коллекции текстов), s (границы предложения), g (слитное написание двух последовательно идущих тегов)[3].

```
STRUCTURE doc {
   TYPE "map64"
   ATTRIBUTE header {
         TYPE "UNIQUE"
   }
   ATTRIBUTE author
   ATTRIBUTE date
   ATTRIBUTE wordcount
   ATTRIBUTE sex
   ATTRIBUTE sphere
   ATTRIBUTE type
   ATTRIBUTE style
   ATTRIBUTE audience_age
   ATTRIBUTE audience_level
   ATTRIBUTE audience_size
   ATTRIBUTE source
   ATTRIBUTE publication
   ATTRIBUTE publ_year
   ATTRIBUTE medium
   ATTRIBUTE id
}

STRUCTURE font {
         ATTRIBUTE type
}

STRUCTURE p {
   TYPE "map64"
   ATTRIBUTE heading
  ATTRIBUTE langdiff
  ATTRIBUTE pid {
         TYPE "UNIQUE"
  }
   ATTRIBUTE neardupe
   DISPLAYTAG 0
   DISPLAYEND "¶"
}

STRUCTURE s {
   TYPE "map64"
   ATTRIBUTE sid {
```

[3] Подробнее о возможных структурах и их параметрах см. по ссылке https://www.sketchengine.co.uk/corpus-configuration-file-all-features/#Attributes

```
        TYPE "UNIQUE"
  }
  ATTRIBUTE neardupe
}

STRUCTURE g {
   TYPE "map64"
   DISPLAYTAG 0
   DISPLAYBEGIN "_EMPTY_"
}
```

2) Формат vertical-файл

Vertical-файл содержит, во-первых, сам корпус (т.е. последовательность токенов, их лемм и соответствующих грамматических категорий), во-вторых, атрибуты, определяющие метаданные каждого документа, в-третьих, атрибуты, определяющие сегменты и графическое представление корпуса (т.е. теги границ предложений и абзацев, заголовка документа, слитного отображения токенов, курсивного написания и другие).

В vertical-файле каждый документ коллекции начинается с тега “doc”, внутри которого может содержаться любая метаинформация документа: дата создания, имя автора, пол автора, жанр и тип текста, год публикации, год скачивая файла (если имеется интернет-корпус), регион проживания автора и т.д. Метаданные могут иметь и иерархическую структуру[4]. При непосредственном пользовании корпуса будет доступна фильтрации выдачи по указанным метаданным. В конце документа ставится соответствующий закрывающий тег.

Пример doc-тега:

<doc author="Фонвизин Д.И." sex="муж" header="К г. Сочинителю Былей и небылиц" date="1783" wordCount="791" sphere="публицистика" type="письмо открытое" style="нейтральный" audience_age="н-возраст" audience_level="н-уровень" audience_size="средняя" source="Фонвизин Д.И. Собрание сочинений в двух томах. М.; Л., 1959" publication="Фонвизин Д.И. Собрание сочинений: В 2 т. Т. 2" publ_year="1959" medium="книга" id="58902304.22fd3e6c">
...
...
…
</doc>

Внутри doc-тега содержатся собственно текст документы и теги графического представления текста. Последних может и не быть, или часть из них может быть обязательной, а часть нет (всё это определяется в registry-файле). Обычно, как минимум, используют тег “g”, который ставится, если между двумя токенами нет пробела; тег “s”, который обозначает границы предложения.

Тексты представляются следующим образом. Каждый токен записывается на отдельной строке. Если имеются его атрибуты (количество и обязательность которых также определяется в registry-файле), то они в указанной в registry-файле последовательности записываются через знак табуляции на той же строке, что и сам токен. Стандартным является следующее представление:
“token lemma PoS”
Однако количество атрибутов (столбцов) не ограничено, может включать в себя, например, специальные синтаксические, семантические, дискурсивные и иные теги.

[4] См. подробнее о реализации по ссылке https://www.sketchengine.co.uk/text-types-headers-and-subcorpora/

Пример представления содержания doc-тега:

```
<p>
<s>
Как как Cj C 1
известно известно Av R 1
<g/>
, , Zz , 1
смарт-часы смарт-часы Nn Ncmpnn 0
Apple Apple Zz - 1
Watch Watch Zz - 1
оказались оказаться Vb Vmis-p-m-p 1
достаточно достаточно Av R 1
популярными популярный Aj Afpmpif 1
на на Pp Sp-l 1
старте старт Nn Ncmsln 1
продаж продажа Nn Ncfpgn 1
<g/>
. . Zz SENT 1
</s>
…
…
...
</p>
```

* тег “p” обозначает границы абзаца