Ïðîáëåìû ëèíãâèñòè÷åñêîé ðàçìåòêè è àíàëèçà ýëåêòðîííûõ êðèòè÷åñêèõ èçäàíèé òåêñòîâ

Àâòîð Àëåêñåé Ìèõàéëîâè÷ Ëàâðåíòüåâ
05.08.2012 ã.
Ïîñëåäíåå îáíîâëåíèå 27.08.2012 ã.

Summary. In this paper we consider some problems of automatic linguistic annotation and analysis of textual heritage documents encoded according to the TEI XML guidelines. TEI XML is a popular standard for encoding electronic editions of textual heritage documents as it allows highly customizable semantically-oriented markup independent of a particular platform or software. TEI is aimed at facilitating data exchange and interoperability. However, rich editorial markup including various readings and interpretations at various levels of linguistic hierarchy may be a serious challenge if one wants to apply NLP (natural language processing) tools to such an edition. Based on the example of the Base de Français Médiéval Old French corpus and on the electronic edition of the Queste del saint Graal, we will discuss the solutions to these problems that are implemented in the TXM platform import modules.

Ñòàíäàðò ýëåêòðîííîé ðàçìåòêè òåêñòîâ â ôîðìàòå XML â ñîîòâåòñòâèè ñ ðåêîìåíäàöèÿìè ìåæäóíàðîäíîé Èíèöèàòèâû ïî Êîäèðîâàíèþ Òåêñòîâ TEI (Text Encoding Initiative, http://www.tei-c.org) ïðèîáðåë â ïîñëåäíåå ãîäû äîñòàòî÷íî øèðîêîå ðàñïðîñòðàíåíèå â îáëàñòè ýëåêòðîííîãî èçäàíèÿ òåêñòîâ, ïðèíàäëåæàùèõ ê ôîíäó ïèñüìåííîãî íàñëåäèÿ ðàçëè÷íûõ ñòðàí è êóëüòóð. 150 ïðîåêòîâ ïðåäñòàâëåíû íà ñàéòå TEI, à â äåéñòâèòåëüíîñòè ýòî ÷èñëî ìîæåò áûòü çíà÷èòåëüíî áîëüøå. Ïðåèìóùåñòâàìè ðàçìåòêè â ñòàíäàðòå TEI ÿâëÿþòñÿ åå îïîðà íà òùàòåëüíî ðàçðàáîòàííóþ òåîðèþ ñòðóêòóðû òåêñòà è äîêóìåíòà, ëåãêîñòü ïåðñîíàëèçàöèè è àäàïòàöèè ê êîíêðåòíîìó ìàòåðèàëó çà ñ÷åò ìîäóëüíîé îðãàíèçàöèè è ñïåöèàëüíîãî ìåõàíèçìà ñïåöèôèêàöèè ODD, à òàêæå íåçàâèñèìîñòü îò êîíêðåòíîé ïëàòôîðìû èëè ïðîãðàììíîãî ïðîäóêòà. Âìåñòå ñ òåì ÷ðåçâû÷àéíàÿ ãèáêîñòü TEI ñîçäàåò îïðåäåëåííûå òðóäíîñòè äëÿ ðàçðàáîòêè ïðîãðàììíûõ ñðåäñòâ îáðàáîòêè, àíàëèçà è ïóáëèêàöèè òåêñòîâ, ðàçìå÷åííûõ â ýòîì ñòàíäàðòå, îñîáåííî åñëè ðå÷ü èäåò î ñðåäñòâàõ «øèðîêîãî ïðîôèëÿ», ïðåäíàçíà÷åííûõ äëÿ èñïîëüçîâàíèÿ áîëåå ðàìîê îòäåëüíîãî âçÿòîãî ïðîåêòà. Â ÷àñòíîñòè, «ãëóáîêóþ» ôèëîëîãè÷åñêóþ ðàçìåòêó, ó÷èòûâàþùóþ ðàçíî÷òåíèÿ è âàðèàíòû èíòåðïðåòàöèè ôðàãìåíòîâ òåêñòà íà ðàçíûõ óðîâíÿõ èåðàðõèè ÿçûêîâûõ ñòðóêòóð, ìîæåò áûòü òðóäíî ñîâìåñòèòü ñ èñïîëüçîâàíèåì èíñòðóìåíòîâ àâòîìàòè÷åñêîé ëèíãâèñòè÷åñêîé ðàçìåòêè (òîêåíèçàöèè, ëåììàòèçàöèè, ìîðôîëîãè÷åñêîé êàòåãîðèçàöèè è ò.ï.).

Â íàøåì äîêëàäå ìû ïðåäñòàâèì ìåòîäèêó ïîäãîòîâêè (íîðìàëèçàöèè) ôèëîëîãè÷åñêîé ðàçìåòêè òåêñòîâ Áàçû ñðåäíåâåêîâîãî ôðàíöóçñêîãî ÿçûêà BFM (Base de Français Médiéval, http://bfm.ens-lyon.fr) â ïðîöåññå èõ çàãðóçêè íà ïëàòôîðìó TXM ñ öåëüþ èõ äàëüíåéøåãî ëèíãâèñòè÷åñêîãî àíàëèçà. Ìû òàêæå ðàññìîòðèì ðåçóëüòàòû îïûòà ïî àäàïòàöèè äàííîé ìåòîäèêè ê òåêñòàì äðóãèõ ïðîåêòîâ, ðàçìå÷åííûì íà îñíîâå ðåêîìåíäàöèé TEI, íî ïðèìåíÿþùèì òîëè÷íûå îò BFM ðåøåíèÿ â ðÿäå êëþ÷åâûõ äëÿ ëèíãâèñòè÷åñêîãî àíàëèçà àñïåêòîâ ðàçìåòêè.

Òåêñòîìåòðèÿ (textométrie) âîçíèêëà êàê íàó÷íîå íàïðàâëåíèå âî Ôðàíöèè â 1980-å ãîäû. Â åå ðàìêàõ áûëè ðàçðàáîòàíû ýôôåêòèâíûå ìåòîäèêè àíàëèçà îáúåìíûõ êîðïóñîâ òåêñòîâ. Âñëåä çà ëåêñèêîìåòðèåé è ñòàòèñòè÷åñêèì àíàëèçîì òåêñòà òåêñòîìåòðèÿ ïðåäëàãàåò ñòàòèñòè÷åñêèå îñíîâàííûå ìåòîäû è èíñòðóìåíòû àíàëèçà äëÿ ðàçëè÷íûõ ãóìàíèòàðíûõ íàóê.

TXM – ýòî ìîäóëüíàÿ ïëàòôîðìà ñ îòêðûòûì èñõîäíûì êîäîì, êîòîðàÿ ñî÷åòàåò ôóíêöèè ðàçëè÷íûõ ðàíåå ðàçðàáîòàííûõ ïðîãðàìì òåêñòîìåòðè÷åñêîãî àíàëèçà. Îíà ïðåäñòàâëÿåò íîâîå ïîêîëåíèå òåêñòîìåòðè÷åñêîãî èíñòðóìåíòàðèÿ, èñïîëüçóþùåå ñîâðåìåííûå êîðïóñíûå òåõíîëîãèè (Unicode, XML, TEI, NLP). Ïîäðîáíàÿ èíôîðìàöèÿ î ïëàòôîðìå TXM ïðåäñòàâëåíà â ïóáëèêàöèÿõ [Heiden 2010; Heiden et al. 2010; Pincemin et al. 2010], à òàêæå íà ñàéòå http://textometrie.ens-lyon.fr/?lang=en.

Áàçà ñðåäíåâåêîâîãî ôðàíöóçñêîãî ÿçûêà (BFM) – ýòî êîðïóñ òåêñòîâ ñòàðî- è ñðåäíåôðàíöóçñêîãî ÿçûêà (IX – XV ââ.), â íàñòîÿùåå âðåìÿ ðàçðàáàòûâàþùèéñÿ ëàáîðàòîðèåé ICAR Íàöèîíàëüíîãî öåíòðà íàó÷íûõ èññëåäîâàíèé

Ôðàíöèè (CNRS) è Ëèîíñêîãî óíèâåðñèòåòà. Áàçà íàñ÷èòûâàåò 75 òåêñòîâ îáùèì îáúåìîì áîëåå 3 500 000 òåêñòîôîðì. Èñòî÷íèêàìè BFM â îñíîâíîì ÿâëÿþòñÿ àâòîðèòåòíûå êðèòè÷åñêèå èçäàíèÿ, îäíàêî â ïîñëåäíåå âðåìÿ ðàçâèâàþòñÿ ñîáñòâåííûå èçäàíèÿ, îïèðàþùèåñÿ íà ëèíãâèñòè÷åñêè âûâåðåííûå òðàíñêðèïöèè îðèãèíàëüíûõ ðóêîïèñåé. Â êà÷åñòâå ïðèìåðà ìîæíî ïðèâåñòè èíòåðàêòèâíîå èçäàíèå àíîíèìíîãî ðîìàíà XIII â. «Ïîèñêè Ñâÿòîãî Ãðààëÿ» («La Queste del saint Graal») ïîä ðåäàêöèåé Ê. Ìàðêåëëî-Íèçüÿ [Queste 2011].

Ñ ìàÿ 2012 ãîäà äîñòóï ê BFM îñóùåñòâëÿåòñÿ ïîñðåäñòâîì ïîðòàëà http://txm.bfm-corpus.org/bfm, ïîñòðîåííîãî íà ïëàòôîðìå TXM.

Âñå òåêñòû BFM ðàçìå÷åíû â ôîðìàòå XML íà îñíîâå ðåêîìåíäàöèé TEI, â ñîîòâåòñòâèè ñî ñïåöèôèêàöèåé, ðàçðàáîòàííîé äëÿ íóæä ïðîåêòà [Guillot et al. 2010] ñ ó÷åòîì ïåðñïåêòèâû ëèíãâèñòè÷åñêîãî àíàëèçà. Ïëàòôîðìà TXM íå òîëüêî ñëóæèò äëÿ êîðïóñà BFM ñðåäñòâîì äîñòóïà ïîëüçîâàòåëåé, íî è ïîçâîëÿåò îñóùåñòâëÿòü ðÿä îïåðàöèé àâòîìàòè÷åñêîé è ïîëóàâòîìàòè÷åñêîé ðàçìåòêè (â ÷àñòíîñòè, ìîðôîëîãè÷åñêîé àííîòàöèè, âûÿâëåíèÿ ïðÿìîé ðå÷è).

Îäíîé èç íàèáîëåå ñëîæíûõ çàäà÷ ïðè èñïîëüçîâàíèè ñðåäñòâ àâòîìàòè÷åñêîãî ëèíãâèñòè÷åñêîãî àíàëèçà êîðïóñîâ ïðèìåíèòåëüíî ê òåêñòàì, âêëþ÷àþùèì ãëóáîêóþ ðåäàêòîðñêóþ ðàçìåòêó, ÿâëÿåòñÿ êîððåêòíàÿ èäåíòèôèêàöèÿ ñëîâ (òîêåíèçàöèÿ) è ïðåäëîæåíèé, ê êîòîðûì ýòîò àíàëèç äîëæåí ïðèìåíÿòüñÿ áåç ïîòåðè ñàìîé ðåäàêòîðñêîé ðàçìåòêè.

Ñëåäóþùèé ïðèìåð ðàçìåòêè, ñîäåðæàùèé ïðåäëîæåíûé ðåäàêòîðîì ôðàãìåíò òåêñòà íà ìåñòå ëàêóíû, íà÷èíàþùåéñÿ â êîíöå îäíîãî ñëîâà è çàêàí÷èâàþùåéñÿ íåñêîëüêèìè ñëîâàìè ïîçæå, àáñîëþòíî êîððåêòåí ñ òî÷êè çðåíèÿ ðåêîìåíäàöèé TEI, îäíàêî âåñüìà ñëîæåí äëÿ òîêåíèçàöèè ñ ó÷åòîì îñíîâíîãî ñîáñòâåííîãî ÿçûêà XML (çàïðåò «ïåðåêðåùèâàíèÿ» ýëåìåíòîâ, ðèñê ïîÿâëåíèÿ ïðîáåëîâ ìåæäó òýãàìè è òåêñòîâûìè óçëàìè ïðè îáðàáîòêå):

en<supplied>tra a
cheval en la</supplied> sale une mout bele damoisele

Åùå áîëåå ñëîæíûå ïðîáëåìû âîçíèêàþò ïðè ðàçìåòêå ïðåäëîæåíèé, îñîáåííî â ñòèõîòâîðíûõ òåêñòàõ, ãäå ïåðåêðåùèâàíèå ìåòðè÷åñêîé è ñèíòàêñè÷åñêîé ñòðóêòóðû âñòðå÷àåòñÿ î÷åíü ÷àñòî.

Ðàçóìååòñÿ, ìîæíî «îòôèëüòðîâàòü» âñå ýëåìåíòû, ïåðåêðåùèâàþùèåñÿ ñ îñíîâíûìè ëèíãâèñòè÷åñêèìè ñòðóêòóðàìè (ñëîâàìè è ïðåäëîæåíèÿìè), îäíàêî ýòî ìîæåò ïðèâåñòè ê ïîòåðå ñóùåñòâåííîé èíôîðìàöèè äëÿ çàïðîñà è âèçóàëèçàöèè (íàïðèìåð, î òîì, ïîäâåðãàëàñü ëè ñëîâîôîðìà ðåäàêòîðñêîé ïðàâêå).

Ñîçäàíèå àëãîðèòìà òîêåíèçàöèè, êîòîðûé êîððåêòíî îáðàáàòûâàë áû ëþáîé òåêñò ñ ãëóáîêîé ðåäàêòîðñêîé ðàçìåòêîé â ñòàíäàðòå TEI XML, ïðåäñòàâëÿåòñÿ ïðàêòè÷åñêè íåâîçìîæíûì. Òåì íå ìåíåå, ìîæíî äîáèòüñÿ âïîëíå óäîâëåòâîðèòåëüíûõ ðåçóëüòàòîâ ïðè óñëîâèè, ÷òî ðàçìåòêà èñõîäíîãî äîêóìåíòà îòâå÷àåò ðÿäó ïðîñòûõ ïðàâèë. Íàïðèìåð, «òýãè, ðàñïîëîæåííûå âíóòðè ñëîâà, äîëæíû áûòü ÷åòêî èäåíòèôèöèðîâàíû» èëè «åñëè ðàçìå÷åííûé ñåãìåíò òåêñòà íà÷èíàåòñÿ âíóòðè îäíîãî ñëîâà è çàõâàòûâàåò íåñêîëüêî ïîñëåäóþùèõ, åãî íåîáõîäèìî ðàçäåëèòü».

Òàêæå âîçìîæíî ñîñòàâèòü ñïèñêè òýãîâ TEI â çàâèñèìîñòè îò èõ ïîçèöèè â ëèíãâèñòè÷åñêîé èåðàðõèè òåêñòà â ðàìêàõ îòäåëüíîãî ïðîåêòà. Íàïðèìåð, ïðåäëîæåíèÿ ðàñïîëàãàþòñÿ âíóòðè ýëåìåíòîâ òèïà «àáçàö» <p> èëè «áëîê òåêñòà» <ab>, à íå íàîáîðîò. Íåêîòîðûå òýãè ìîæíî ñ÷èòàòü

ýêâèâàëåíòíûìè ñëîâó (íàïðèìåð, <abbr>, <num>, <pc>), à íåñêîëüêî ýëåìåíòîâ ïî÷òè âñåãäà ðàñïîëàãàþòñÿ âíóòðè ñëîâà (<am>, <c>, <ex>). Ðÿä ýëåìåíòîâ ñîäåðæàò ñåãìåíòû òåêñòà, êîòîðûå íå ñëåäóåò òîêåíèçèðîâàòü, ïîñêîëüêó îíè íå ïðèíàäëåæàò ê ìàòåðèàëó èñòî÷íèêà (íàïðèìåð, ðåäàêòîðñêèå ïðèìå÷àíèÿ è ñíîñêè â òåêñòå êðèòè÷åñêîãî èçäàíèÿ). Â ðàìêàõ êîíêðåòíîãî ïðîåêòà ýòè ñïèñêè ìîãóò áûòü ñóùåñòâåííî ðàñøèðåíû è óòî÷íåíû, â ðåçóëüòàòå ÷åãî ÷èñëî ýëåìåíòîâ, êîòîðûå ìîãóò ïåðåêðûâàòüñÿ ñ ëèíãâèñòè÷åñêèìè ñòðóêòóðàìè, çíà÷èòåëüíî ñîêðàùàåòñÿ.

Ðàçìåòêà òåêñòà BFM ñîîòâåòñòâóåò ÷àñòî ñôîðìóëèðîâàííîé ñïåöèôèêàöèè ïðèìåíåíèÿ ðåêîìåíäàöèé TEI, îòðàæåííîé â äîêóìåíòàöèè ODD [Guillot et al. 2010]. Äàííàÿ ñïåöèôèêàöèÿ âêëþ÷àåò ïðàâèëà èñïîëüçîâàíèÿ âíóòðèñëîâíûõ òýãîâ (íàïðèìåð, èñïðàâëåííûõ ðåäàêòîðîì áóêâ èëè çíàêîâ ïåðåíîñà), à òàêæå ýëåìåíòîâ, êîòîðûå ìîãóò ïåðåêðûâàòüñÿ ïî ñòðóêòóðîé ïðåäëîæåíèé (íàïðèìåð, «ïóñòûå» ýëåìåíòû <lb/> «íîâàÿ ñòðîêà» èñïîëüçóþòñÿ âìåñòî <l> «ñòèõ», à ýëåìåíò öèòèðîâàíèÿ <q> ðàññìàòðèâàåòñÿ êàê ãðàíèöà ïðåäëîæåíèÿ). Ýòî ïîçâîëèëî ðåàëèçîâàòü â ïëàòôîðìå TXM ýôôåêòèâíûé àëãîðèòì òîêåíèçàöèè äëÿ òåêñòîâ BFM.

Â ïîñëåäíåå âðåìÿ áûë óñïåøíî ïðîâåäåí ðÿä òåñòîâ ïî àäàïòàöèè äàííîãî àëãîðèòìà ê äîêóìåíòàì TEI XML, ïîäãîòîâëåííûì â äðóãèõ ïðîåêòàõ. Ñðåäè íèõ ìîæíî íàçâàòü «Âèðòóàëüíóþ áèáëèîòåêó ãóìàíèñòîâ» (http://www.bvh.univ-tours.fr) è èçäàíèå ÷åðíîâèêîâ ðîìàíà «Áóâàð è Ïåêþøå» Ãþñòàâà Ôëîáåðà (http://dossiers-flaubert.ish-lyon.cnrs.fr). Àäàïòàöèÿ ñîñòîèò â ïðèìåíåíèè ñïåöèàëüíûõ ôèëüòðîâ XSL íà âõîäå è íà âûõîäå ïðîöåäóðû òîêåíèçàöèè. «Âõîäíîé» ôèëüòð óäàëÿåò òýãè, êîòîðûå íå ïðåäñòàâëÿþò èíòåðåñà äëÿ ýêñïëóàòàöèè ñ ïîìîùüþ TXM, à òàêæå óïðîùàåò è íîðìàëèçóåò íåêîòîðûå ñëîæíûå ñòðóêòóðû ýëåìåíòîâ XML. «Âûõîäíîé» ôèëüòð ïîçâîëÿåò èñïðàâèòü äëÿ îøèáîê, êîòîðûõ ïðàêòè÷åñêè íåâîçìîæíî èçáåæàòü â ïðîöåññå ïåðâè÷íîé òîêåíèçàöèè (íàïðèìåð, äåëåíèå ñëîâà ïðè ïåðåíîñå â êîíöå ñòðàíèöû, êîãäà öåëûé ðÿä òýãîâ ìîæåò ðàñïîëàãàòüñÿ ìåæäó åãî íà÷àëîì è êîíöîì).

Â òåêñòå äîêëàäà ìû ïðèâåäåì êîíêðåòíûå ïðèìåðû ðåàëèçîâàííûõ ðåøåíèé è ïðîäåìîíñòðèðóåì ðåçóëüòàòû, êîòîðûå ìîæíî ïîëó÷èòü ïðè ýêñïëóàòàöèè òåêñòîâ BFM è äðóãèõ ïðîåêòîâ ñ ïîìîùüþ ïëàòôîðìû TXM. Áîëåå ïîäðîáíî ðàçëè÷íûå àëãîðèòìû òîêåíèçàöèè, ðåàëèçîâàííûå íà ïëàòôîðìå TXM, îïèñàíû â [Heiden 2010].

Ñïèñîê ëèòåðàòóðû

Guillot, C., Heiden, S., Lavrentiev, A., Bertrand, L. Manuel d’encodage XML-TEI des textes de la Base de Français Médiéval, Lyon: Équipe BFM, 2010. – Àäðåñ â Èíòåðíåò http://bfm.ens-lyon.fr/IMG/pdf/Manuel_Encodage_TEI.pdf.

Heiden, S. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. 24th // Pacific Asia Conference on Language, Information and Computation / Ed. Kiyoshi Ishikawa Ryo Otoguro. Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010. P. 389-398.

Heiden, S., Magué, J.-P., Pincemin, B. TXM : Une plateforme logicielle open-source pour la

textométrie – conception et développement // Statistical Analysis of Textual Data - Proceedings of 10th International Conference JADT 2010 / Bolasco, S. et al. (Eds.). – Rome: Edizioni Universitarie di Lettere Economia Diritto, 2010. – P. 1021-1032. – Àäðåñ â Èíòåðíåò http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-1021-1032_025-Heiden.pdf.

Pincemin, B., Heiden, S., Lay, M.-H., Leblanc J.-M. and Viprey, J.-M. Fonctionnalités textométriques: Proposition de typologie selon un point de vue utilisateur. // Statistical Analysis of Textual Data - Proceedings of 10th International Conference JADT 2010 / Bolasco, S. et al. (Eds.). – Rome: Edizioni Universitarie di Lettere Economia Diritto, 2010. – P. 341-353. – Àäðåñ â Èíòåðíåò http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-0341-0354_023-Pincemin.pdf.

Queste del saint Graal. Édition numérique intéractive / Ed. Marchello-Nizia, Ch. – Lyon: Équipe de la BFM, 2011. – Àäðåñ â Èíòåðíåò http://txm.bfm-corpus.org/txm.