Áàçû äàííûõ è êîðïóñû òåêñòîâ ñðåäíåâåêîâîãî ôðàíöóçñêîãî ÿçûêà: ïîäõîäû, ïðîåêòû, òåõíîëî

Àâòîð Àëåêñåé Ìèõàéëîâè÷ Ëàâðåíòüåâ
18.07.2008 ã.
Ïîñëåäíÿÿ ïðîâåðêà 20.07.2008 ã.

Òåêñò ïå÷àòíîãî èçäàíèÿ â ôîðìàòå PDF

Èñòîðèÿ òåêñòîâûõ áàç äàííûõ íà ôðàíöóçñêîì ÿçûêå âîñõîäèò ê êîíöó 1950 õ ãã., êîãäà áûëà íà÷àòà ðàáîòà íàä ïîäãîòîâêîé ñëîâ
ñ XVI ïî XX âåê. Òåêñòîâàÿ áàçà, ïîëó÷èâøàÿ íàçâàíèå Frantext, àêòèâíî ðàçâèâàëàñü â 60 å — 80 å ãã., ñ íà÷àëà 90 õ
Frantext ïðèîáðåëà øèðîêóþ èçâåñòíîñòü è ñòàëà àêòèâíî èñïîëüçîâàòüñÿ ëèíãâèñòàìè, ëèòåðàòóðîâåäàìè è èñòîðèêàìè. Â òî a
ïðàêòè÷åñêè åäèíñòâåííûì èñòî÷íèêîì äàííûõ î ÿçûêå è ëèòåðàòóðå ñðåäíåâåêîâüÿ, è èñïîëüçîâàíèå èíôîðìàöèîííûõ òåõíîëîãèé ïðè

Èñòîðèÿ òåêñòîâûõ áàç äàííûõ íà ôðàíöóçñêîì ÿçûêå âîñõîäèò ê êîíöó 1950 õ ãã., êîãäà áûëà íà÷àòà ðàáîòà íàä ïîäãîòîâêîé ñëîâ
ñ XVI ïî XX âåê. Òåêñòîâàÿ áàçà, ïîëó÷èâøàÿ íàçâàíèå Frantext, àêòèâíî ðàçâèâàëàñü â 60 å — 80 å ãã., ñ íà÷àëà 90 õ
Frantext ïðèîáðåëà øèðîêóþ èçâåñòíîñòü è ñòàëà àêòèâíî èñïîëüçîâàòüñÿ ëèíãâèñòàìè, ëèòåðàòóðîâåäàìè è èñòîðèêàìè. Â òî a
ïðàêòè÷åñêè åäèíñòâåííûì èñòî÷íèêîì äàííûõ î ÿçûêå è ëèòåðàòóðå ñðåäíåâåêîâüÿ, è èñïîëüçîâàíèå èíôîðìàöèîííûõ òåõíîëîãèé ïðè

Â êîíöå 70 õ è 80 õ ãîäàõ âîçíèêàþò ïåðâûå ïðîåêòû êîìïüþòåðíîãî èçó÷åíèÿ ñðåäíåâåêîâûõ ôðàíöóçñêèõ òåêñòîâ, îäíàêî ïîëó÷åíí
òåêñòîâûõ áàç äàííûõ íà÷èíàþò îñóùåñòâëÿòüñÿ â 90 å ãîäû.

Ñ ñàìîãî íà÷àëà èññëåäîâàòåëè ñòîëêêíóëèñü ñ äâóìÿ ïðèíöèïèàëüíûìè âîïðîñàìè, êîòîðûå íåîáõîäèìî ðåøèòü ïåðåä ñîçäàíèåì ýëåêò

1. Íà êàêîé èñòî÷íèê îïèðàòüñÿ: ðóêîïèñü èëè èçäàíèå?

2. Ðàáîòàòü ëè ñ öåëûìè òåêñòàìè èëè ñ ôðàãìåíòàìè?

Íà èõ ðåøåíèå âëèÿþò íå òîëüêî ìåòîäîëîãè÷åñêèå ïðèíöèïû èññëåäîâàòåëÿ, íî è ôàêòîð âðåìåíè è ðåñóðñîâ, íåîáõîäèìûõ äëÿ ñîçä
îáóñëîâèâøèå â ïðîøëîì òå èëè èíûå ðåøåíèÿ ñîçäàòåëåé êîðïóñîâ.

Ñðåäè ðåàëèçîâàííûõ â ïðîøëîì è ïðîäîëæàþùèõñÿ ïðîåêòîâ òåêñòîâûõ áàç íà ñðåäíåâåêîâîì ôðàíöóçñêîì ìîæíî âñòðåòèòü âñå âàð

Êðàòêî îõàðàêòåðèçîâàâ îñíîâíûå ïðîåêòû êîðïóñîâ ñðåäíåâåêîâûõ òåêñòîâ, ðàçðàáîò÷èêè êîòîðûõ îáúåäèíèëèñü â 2004 ã. â Êîî
âíåøíå ïðîòèâîïîëîæíûå, íî ïî ñóòè âçàèìîäîïîëíÿþùèå ïîäõîäû.

Ïåðâûé ïðîåêò — Áàçà ñðåäíåâåêîâîãî ôðàíöóçñêîãî (BFM) — áûë íà÷àò â 1989 ã. â Âûñøåé íîðìàëüíîé øêîëå ãóìàí
Marchello-Nizia). BFM — ïîëíîòåêñòîâàÿ áàçà äàííûõ, îñíîâàííàÿ íà ñîâðåìåííûõ êðèòè÷åñêèõ èçäàíèÿõ. Öåëüþ ïðîåêòà áû
òåêñòîâûõ äàííûõ.

Èñïîëüçîâàíèå â áàçå ïîëíûõ òåêñòîâ ïðîèçâåäåíèé îáúÿñíÿåòñÿ òåì, ÷òî îíà èçíà÷àëüíî áûëà îðèåíòèðîâàíà íà ìàêñèìàëüíî øèðîêèé
êîðïóñêå ñðîêè ñôîðìèðîâàòü äîñòàòî÷íî êóóïíûé êîðïóñ. Ïðè ýòîì äëÿ âêëþ÷åíèÿ â áàçó îòáèðàëèñü ïðåæäå âñåãî èçäàíèÿ, ïîäãîò
ôàêòè÷åñêêå îøèáêè è îãîâàðêâàÿ âñå èñïðàâëåíèÿ â ñíîñêàõ.

Ïîïîëíåíèå áàçû îñóùåñòâëÿëîñü áëàãîäàðÿ ñêàíèðîâàíèþ òåêñòîâ èçäàíèé, à òàêæå îáìåíàì è áåçâîçìåçäíûì âêëàäàì ñïåöèàëèñòî
Îñíîâáíûì ïåðèîäîì, íà êîòîðûé ïåðâîíà÷àëüíî îðèåíòèðîâàëñÿ ïðîâêò BFM, áûë ñòàðîôðàíöóçñêèé (IX–XIII ââ.), îäíàêî â áàçó à
òåêñòû ñîõðàíÿëèñü â ôîðìàòå «ïðîñòîé òåêñò», â êîäèðîâêå DOS. Ñ 2001 ã. òåêñòû áàçû ïîñòåïåííî ïåðåâîäÿòñÿ â ôîðìàò XML

Â íàñòîÿùåå âðåìÿ îñíîâíîé êîðïóñ òåêñòîâ BFM íàñ÷èòûâàåò 74 òåêñòà îáùèì îáúåìîì îêîëî 3 000 000 ñëîâîóïîòðåáëåíèé. Ýêñïëóà
Ñ. Ýéäåíîì (S. Heiden). Äîñòóï ê áàçå ïðåäîñòàâëÿåòñÿ áåñïëàòíî âñåì ó÷åíûì, ïðåïîäàâàòåëÿì è ñòóäåíòàì, ïðèíèàþùèì óñëîâèÿ

Âòîðîé ïðîåêò — êîìïëåêñíîå ìóëüòèìåäèéíîå èçäàíèå ðóêîïèñíîé òðàäèöèè ðîìàíà Êðåòüåíà äå Òðóà «Ðûöàðü òåëåãè» (&
Charrette», îñóùåñòâëåíîå â 1990–2003 ãã. â Ïðèíñòîíñêîì óíèâåðñèòåòå ïîä ðóêîâîäñòâîì ïðîôåññîðà Ê. Þòòè (K. Uitti).

Åãî öåëüþ áûëî ñîçäàòü ïðèíöèïèàëüíî íîâîå èçäàíèå ñðåäíåâåêîâîãî òåêñòà, ñîâìåùàþùåå â ñåáå êðèòè÷åñêè âûâåðåííûé òåêñò
ìîðôîëîãè÷åñêîé è ôèëîëîãè÷åñêîé ðàçìåòêîé è ñíàáæåííîå ìùíîé ïîèñêîâîé ìàøèíîé.

Ïåðâîíà÷àëüíî òðàíñêðèïöèè ðóêîïèñåé â ïðîâêòå Charrette îñóùåñòâëÿëèñü â ôîðìàòå SGML â ñîîòâåòñòâèè ñ ðåêîìåíäàöèÿìè T
îðèãèíàëàìè è âíåñåíèåì äîïîëíèòåëüíîé èíôîðìàöèè. Ïàðàëëåëüíî íà îñíîâå êðèòè÷åñêîãî èçäàíèÿ áûëà ñîçäàíà áàçà äàííûõ ïîýòè÷å

Áàçà äàííûõ ïî ïîýòè÷åñêèì ôèãóðàì, äèëëîìàòè÷åñêèå òðàíñêðèïöèè è ôîòîãðàôèè ðóêîïèñåé äîñòóïíû íà ñàéòå ïðîâêòà.

Â ðàìêàõ Êîíñîðöèóìà CCFM â íàñòîÿùåå âðåìÿ âåäåòñÿ âûðàáîòêà åäèíûõ íîðì êîäèðîâàíèÿ è ïîèñêàíèÿ ñðåäíåâåêîâûõ ôðàíöóçñê
åãî ó÷àñòíèêîâ. Ñîâìåñòíûè óñèëèÿìè ðåøàþòñÿ ïðîáëåìû ìîðôîëîãè÷åñêîé ðàçìåòêè è ëåììàòèçàöèè òåêñòîâ, âêëþòñÿ èññëåäîâ

ïðèîðèòåòíûå íàïðàâëåíèÿ èõ ðàñøèðåíèÿ.

Summary

In this paper we will give a review of the current projects on the Medieval (9th — 15th centuries) French text databases, corpora and electronic editions. We will present: different choices of composition and sources for such databases (full text vs. sample corpora; using critical editions vs. original manuscripts); the principles of markup and the technologies used. Two projects will be considered in more detail: the Base de Français Médiéval (http://bfm.ens-lsh.fr) and the Charrette project (http://lancelot.baylor.edu)

Ñïèñîê ëèòåðàòóðû

Heiden S., Guillot Ñ. Capitalisation des savoirs par le web : une application de la TEI pour l’encodage et l’exploitation des textes de la Base de Français Médiéval // Ancien et moyen français sur le Web, enjeux méthodologiques et analyse du discours / Kunstmann, P. (ðåä.). — Ottawa: Les éditions David, 2002. — Ñ. 77–92.

Heiden, S., Lavrentiev, A. Ressources électroniques pour l’étude des textes médiévaux : approches et outils // Revue française de linguistique appliquée. — 2004. — IX(1). — Ñ. 99–118.

Kunstmann, P. Ancien et moyen français sur le Web: textes et bases de données // Revue de linguistique romane. — 2000. — 253–254. — Ñ. 17–42.