Filip Kaas: Corpus of Orkhon runic inscriptions (Korpus orchonských runových textů)

Diplomová magisterská práce
Posudek školitele

Předkládaná práce je výsledkem samostatného úsilí diplomanta. Jejím cílem je vytvoření elektronického korpusu orchonských textů na základě sekundárních pramenů, ovšem s ohledem na principy zpracování korpusů starých jazyků, tedy především s ohledem na jedinečný charakter zdrojových textů a s tím související důraz na vazbu originálu a elektronického zpracování. Dalším ohledem bylo nabídnout lingvistickou anotaci, která usnadní práci s nově vzniklým korpusem. Součástí je rovněž ucelené softwarové řešení korpusu, a především diskuse vhodnosti zvoleného řešení.

Práce sestává z několika částí. V úvodní kapitole autor seznamuje čtenáře s historií orchonských nápisů a jejich tvůrců, charakterizuje jazyk těchto nápisů (z tohoto hlediska je nejdůležitější pasáž věnovaná rekonstruované podobě jazyka a formě zaznamenané runiformním písmem, což se dále promítá do částí věnovaných vztahu originálního písma v korpusu a nabídnutou transkripcí). Dále diskutuje stav zpracování dosavadních elektronických korpusů těchto nápisů. V nejdůležitější, 4. kapitole, pak rozebírá hlavní body budování jeho korpusu, tedy výběr nápisů (pro pilotní verzi, v konečné by měly samozřejmě být obsaženy všechny dochované nápisy), diskutuje datovou strukturu a formu zpracování, od vztahu originálu a elektronické verze, k použité transkripci a také k části analytické, která připravuje korpus pro vhodné další vytěžování. Jedná se především o záležitosti segmentace (v zásadě oddělení ortografických slov, stanovení větných hranic, oddělení významových a funkčních částí slov) a také další formy anotace, tedy jak tzv. glosy, tedy idealizované významy jednotlivých slov (idealizované proto, že obecně jsou beze vztahu ke konkrétnímu umístění slova v kontextu) a také další lingvistická anotace (především tzv. POS anotace). V poslední části pak autor diskutuje doprovodné softwarové funkce ke korpusu, především jde o formy vyhledávání v připraveném korpusu.

Vzhledem ke svému zaměření se v následujícím soustředím především na otázky čistě lingvistické a strukturní, problematika orchonských nápisů, vhodnosti výběru a dalších specifických oborových otázek leží mimo mou kompetenci. V této oblasti mohu jen konstatovat, že autorovy pasáže na mne působí dojmem informovaného pisatele.

V části věnované datové struktuře autor diskutuje především dva typy přístupů, a to záznam pomocí XML a do tzv. „spreadsheetů“, tedy do formy známé např. z programu MS Excel. V obou případech jde v zásadě o rovinné a pravoúhlé uspořádání, a dá se konstatovat, že jako procesní volba představují oba formáty možné řešení, s tím, že v případě XML jde o řešení náročnější jak na velikost souborů, tak i na vyšší složitost softwarového řešení. Proto považuji autorovu volbu spreadsheetu jakožto cesty k pilotnímu korpusu za opodstatněné, i když v budoucnu bych doporučil prozkoumat i další formy datových struktur. V tomto ohledu by bylo dobré vyjít z předem připraveného datového modelu, tedy ze základní představy informací, které je třeba do korpusu zahrnout, a jejich vztahů, a na tomto základě pak volit výslednou datovou strukturu.

Segmentace jazyka podle mého názoru nepředstavuje zásadní problém, jde typologicky o aglutinační jazyk, proto se informace ve znakovém řetězci řadí lineárně od začátku řetězce (srov. např. tab. 9, s. 33). Stejná tabulka nabízí i přehled dalších typů anotace, které autor v dosavadním zpracování připravil, tedy jak „glossing“, tak i další anotaci. Z hlediska současné datové struktury je víceméně nepodstatné řazení, protože jde o samostatné sloupce. Ovšem samotný sloupec „Further annotation“ se čtyřmi podsloupci už vyvolává koncepční otázky. Řadí se zde typy informací, které zřejmě bude vhodné nějakým způsobem

jinak rozdělit, protože v současné formě sdružují několik typů referencí, jejichž kumulace v jednom sloupci může z hlediska dalšího vyhledávání způsobovat jisté problémy (především podsloupec „sp.morph“). Otázka POS anotace (podsloupec „POS“) je samozřejmě do budoucna otevřená, pokud by se ukázala potřeba jemnější klasifikace. Věc je diskutována i v následujících částech (od s. 45, s autorovým řešením na s. 47), pozornost ještě bude potřeba věnovat důkladnější klasifikaci sufixů, která by mohla přinést další zajímavé podněty. Podsloupec „sp.sem“ by mohl být stejně dobře přiřazen do části „Glossing“. Všechny tyto poznámky ukazují na důležitost další diskuse organizace datových struktur.

Otázka vztahu originální podoby nápisu a její elektronické reprezentace se odehrává v zásadě v otázce míry poškození jednotlivých grafémů a víceméně odpovídá možnostem, které autor má. V této věci je závislý na předchozím zpracování, které v dostupných zobrazeních originálů mají své limity a do doby než budou dostupny v odpovídající kvalitě, nebude možné se vyhnout interpretacím dalších badatelů, kteří již tuto věc zpracovali, ovšem, jak dokládá i příslušná pasáž v předložené práci (s. 37-40), s různými výsledky, což přirozeně limituje možnosti dalšího zpracování.

Kromě výše uvedených diskusních oblastí považuji za podstatnou ještě důkladnější diskusi k transkripci. Tato část je zásadní z hlediska kvalitního zpřístupnění textů a může výrazně ovlivnit další interpretaci, proto by tato část měla být zpracována velmi transparentně.

Z formálního hlediska autor v několika částech nedostatečně strukturuje text. V některých místech se můžeme potkat s tím, že autor pojednává najednou o výsledcích dosavadních prací jiných autorů s formou praktických informací (např. instrukce k instalování fontu) až k tomu, co autor sám připravil pro uživatele jeho korpusu (např. rozložení klávesnice). Z technických záležitostí ještě upozorňuji na to, že v obsahu jsou chybně uvedeny stránky jednotlivých kapitol (v PDF ovšem odkazování fungovalo správně).

Závěrem si dovolím konstatovat, že autor připravil v zásadě funkční pilotní korpus orchonských nápisů, jeho přístup působí důvěryhodně a informovaně. Jde bezpochyby o samostatnou autorskou práci, která bude pro obor přínosná. Neubránil se ovšem některým nedostatkům, které bude v příští verzi korpusu nutno ještě propracovat. Podle mého názoru autorova práce vyhovuje požadavkům kladeným na diplomovou práci a mohu ji doporučit k obhajobě, ovšem obsahuje některé nedostatky, kvůli kterým si dovolím navrhnout klasifikaci stupněm „velmi dobře“.

V Praze dne 28. srpna 2017 Petr Zemánek