ISJFR ZRC SAZU, portal Fran

Korpus besedil slovenskih protestantskih piscev 16. stoletja

Korpus besedil slovenskih protestantskih piscev 16. stoletja obsega prepise 45 knjižnih del na 12.945 straneh. Zunaj korpusa so ostali samo letak Otrozhia tabla Adama Bohoriča iz leta 1580, večinoma latinsko pisana slovnica Adama Bohoriča iz leta 1584, tujejezična dela s posamičnimi slovenskimi besedami ter oba večjezična slovarja Hieronima Megiserja (1592, 1603).

Korpus zaenkrat ne vsebuje daljših nemških ali latinskih odlomkov, ki se nahajajo znotraj knjig.

Delo na korpusu je potekalo v več fazah. Najprej smo organizirali ročni prepis besedil s pomočjo indijskega podjetja CyberData India po metodi dvojnega prepisa (dva prepisovalca prepišeta enako besedilo, tretji uskladi mesta, kjer je prepis različen). Ta metoda se je izkazala za finančno vzdržnejšo od skeniranja in optične razpoznave besedila, saj je bilo za le malo večji vložek v prepisanih besedilih manj napak, hkrati pa so prepisovalci lahko besedila sproti opremljali s poenostavljenimi oznakami v jeziku xml.

Druga faza dela je obsegala pregledovanje prepisov, ki je potekalo primerjalno z besedilnimi viri. Kombinirali smo študentsko delo s posegi raziskovalcev na težavnejših mestih (poškodovani listi, zabrisana mesta, nejasen tisk, slabši posnetki ipd.).

Zaradi velikega obsega besedil in težavnega branja nekaterih mest so napake v prepisu še vedno možne, tako da bo končna, popolnoma zanesljiva verzija korpusa mogoča šele v nekaj letih, ko bomo sproti z uporabo korpusa zbirali tudi posamične napake.

Neberljiva mesta ali mesta s posebnostmi v tipografiji so zaenkrat označena z znakom #. Evidentnih napak v zapisu nismo popravljali niti kakorkoli drugače posegali v besedilo (npr. zapis *vtimolitvi dershi mo* namesto *vti molitvi dershimo*, zapis *XXTII* namesto *XXVII*). Dvojni *w* zapisan kot v originalih: *vv*. Večina ligatur v tekstu je razvezana (npr. æ > ae, p s tildo > pre), ohranjena pa je tilda, ki označuje m ali n. Napačno številčenje strani smo ohranili kot del besedila.

Tretja faza dela je obsegala pretvorbo preprostih oznak xml v format xml TEI, ki omogoča večjo povezljivost in lažje strojno branje besedil.

Tako pripravljena besedila so objavljena kot korpus na spletni strani, ki ima poleg korpusne tudi izobraževalno funkcijo, saj si lahko na enem mestu pregledno ogledamo tako rekoč celotno slovensko knjižno produkcijo 16. stoletja.

Povezava: https://fran.si/korpus16