

CorALit: lietuvių mokslo kalbos tekstynas

Projekto vadovas: prof. Aurelija Usonienė
Projekto trukmė: 2007.09 – 2009.12
Projekto biudžetas:

Projekto tikslas – Lietuvių mokslo kalbos tekstyno sudarymas, siekiant sukaupti didelės apimties, autentišką lietuvių mokslo kalbos duomenų bazę, skirtą naudoti objektyviems, šiuolaikiniams akademinio diskurso kiekybinių ir kokybinių parametru tyrimams, atskleisti tarpdisciplininius ypatumus, žanrų įvairovės charakteristiką, suteikti informacijos ir įrodymų apie esamas ar galimas lietuvių mokslo kalbos savitumo / identiteto formavimosi / formavimo / nykimo tendencijas ir suteikti įrodymų apie veiksnius, įtakančius aptariamus procesus. Praktiniu ir taikomuoju požiūriu, šis tekstynas labai reikalingas akademinio rašymo mokymui ir savarankiškam mokymuisi. Sukaupta faktinė medžiaga nėra sukurta pavienių tyrėjų, kalbininkų, ji atspindi realią kalbinę situaciją, demonstruoja kolektyvinę intuiciją ir parodo tikrąją vartoseną.

Siekiant Lietuvių mokslo kalbos tekstyno reprezentatyvumo, jį sudaro visų Lietuvoje plėtojama mokslo sričių, kryptų ir šakų tekstai, parašyti lietuvių kalba. Buvo atlikta išsami duomenų šaltinių analizė (ištirti visų Lietuvos leidyklų internetiniai puslapiai, atlikta kokybinė spaudinių analizė ir jų atranka pagal mokslinę tematiką, leidžiamų mokslinių publikacijų pobūdį).

Projekto uždaviniai:

- Standartizuotos darbo metodikos sudarymas (kriterijų apibrėžimas (išsami sandaros, tekstų apimties, žanrų, šaltinių analizė).
- Specialiųjų poreikių sistemos funkcionalumui aptarimas ir konsultacijos.
- Programinių priemonių adaptavimas, programinis duomenų apdorojimas, kodavimas.
- Morfolginis ir gramatinis anotavimas, lemavimas.
- Paieškos sistemos sukūrimas.
- „Lietuvių akademinio diskurso tyrinėjimai (Lietuvių mokslo kalbos tekstyno sudarymas ir jo internetinio varianto parengimas)“.

Projekto nuoroda: <http://coralit.lt/>

CorALit: the Corpus of Academic Lithuanian

Aim of the project: Building of the corpus of Academic Lithuanian

The aim of the building of CorALit is to accumulate a large database of authentic academic Lithuanian texts which could be used for objective studies of quantitative and qualitative parameters of academic discourse, would allow to reveal the interdisciplinary peculiarities and characteristics of genre variety, would give information about the existing or possible trends of the development or loss of the Lithuanian academic identity and provide evidence concerning factors influencing the processes mentioned. Considering its practical application, this corpus is indispensable in teaching and learning of academic writing. The factual material accumulated in the corpus is not created by linguists or just a few researchers, it represents the true linguistic situation and demonstrates the collective intuition and shows the real usage.

In order to ensure the representativeness of the corpus of academic Lithuanian it includes texts written in Lithuanian from all the areas, fields and branches of science developed in Lithuania. A detailed analysis of data sources was carried out (internet websites of all Lithuanian publishing houses were investigated, a qualitative analysis of publications and selection according to their research topics and nature was carried out).

Objectives of the project:

- Creation of standardised work methodology (definition of criteria, detailed analysis of corpus structure, texts, genres, sources)
- Adaptation of software, data processing, encoding
- Morphological and grammatical annotation, lemmatisation
- Creation of the search system