Corpus of Contemporary Lithuanian Language – the Standardised Way

Erika RIMKUTĖ, Jolanta KOVALEVSKAITĖ,
Vida MELNINKAITĖ, Andrius UTKA,
Daiva VITKUTĖ-ADŽGAUSKIENĖ

Vytautas Magnus University
Centre of Computational Linguistics
Kaunas, 2010

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Presentation plan

- Introduction: development of Corpus of Contemporary Lithuanian Language (CCLL)
- Why TEI P5?
- Overall architecture of CCLL in TEI P5 format
- Annotation at the document metadata level
- Annotation at the text structure level
- Morphosyntactic annotation
- Supporting tools
- Conclusions

Introduction: development of Corpus of Contemporary Lithuanian Language (CCLL)

- CCLL has been started 16 years ago at the Centre of Computational Linguistics at Vytautas Magnus University

- Currently CCLL is:
  - a 160m word corpus
  - newspaper texts – 46%, non-fiction books – 32%, fiction books – 13%, documents – 3%, spoken language texts – 7%
  - morphologically annotated
  - freely searchable on-line

- CCLL has become a representative and authoritative source of information for the usage of real Lithuanian language

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Need for standardisation

- Main drivers:
  - considering possibilities for simultaneous use of several national corpora (e.g. for machine translation tasks),
  - participation in large-scale national and international projects
  - use of open-source and other available tools for corpus analysis, annotation, search, sharing, etc.
  - considering the future possibilities to join large national and international infrastructures, such as CLARIN

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Why TEI P5?

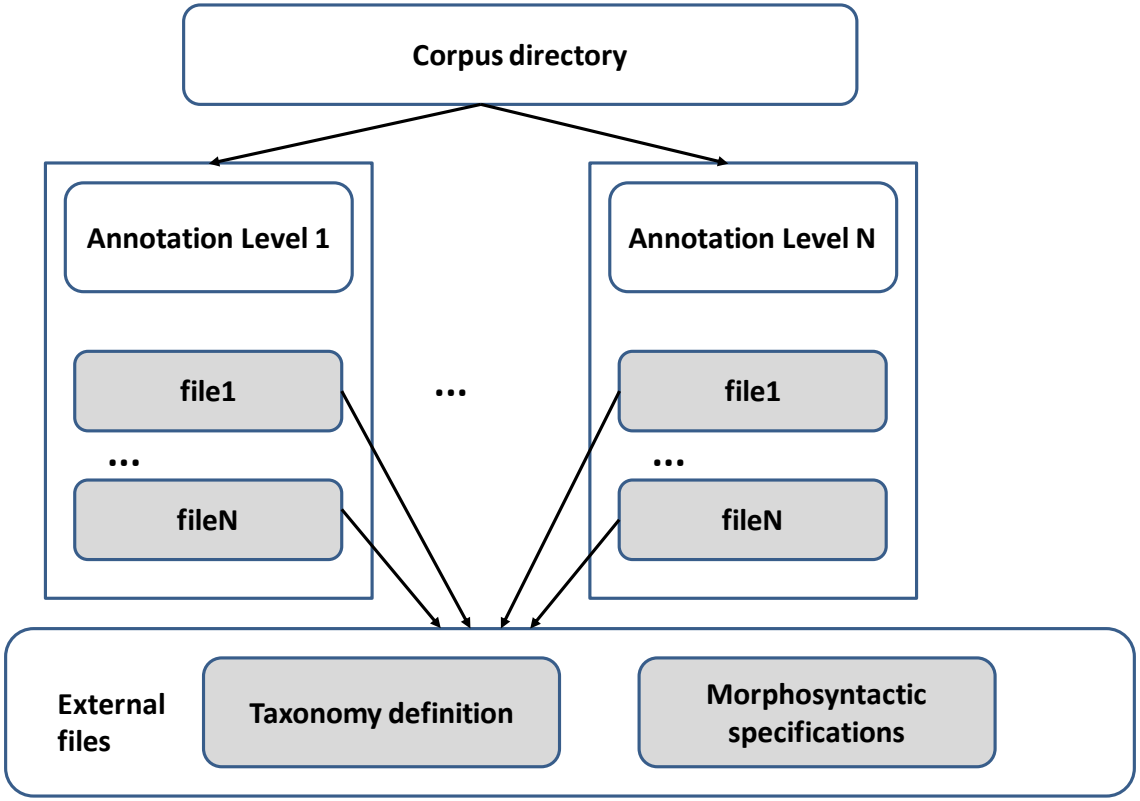
- Choice between the three main alternatives named in the CLARIN short guide:
  - standards developed by International Standards Organization Technical Committee 37 Subcommittee 4 (ISO/ TC37/SC4),
  - XCES (XML Corpus Encoding Standard),
  - TEI P5 (Text Encoding Initiative)

- ISO/ TC37/SC4 family of standards far from being stable
- XCES - still not TEI P5 compatible, poorly documented, also rather limited in annotation levels

- TEI P5:
  - a universal standard for text representation in a digital form, and, thus, a much more complex one,
  - rather flexible in defining different annotation levels,
  - has well-defined semantics and rich documentation,
  - can be easily adapted to various corpus encoding needs.

- TEI P5 also chosen as the encoding standard by National Corpus of Polish, British National Corpus, Bulgarian National Corpus, Croatian Language Corpus, etc.

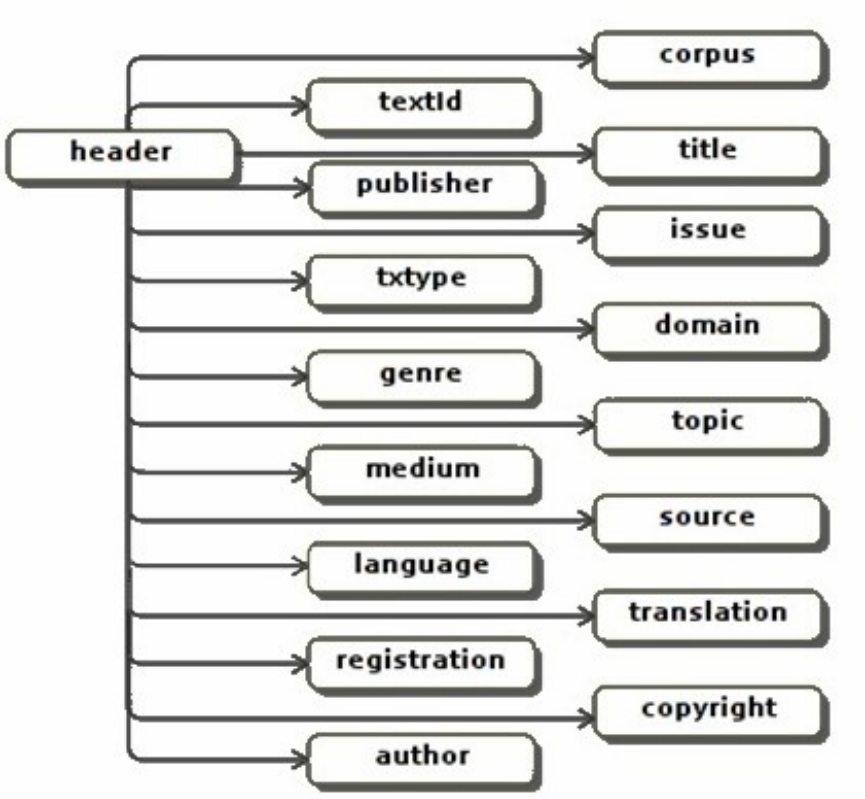
Overall architecture of CCLL



- CCLL is not stored as a single TEI conformant file,
- It is a collection of XML files, s representing separate corpus texts at different annotation levels,
- Each document has its header (<*teiHeader*>), containing document metadata
- Corpus browsing is facilitated using a special directory file for the whole corpus

Annotation at the document metadata level – former status

- Structure of the proprietary <header> element (used before):

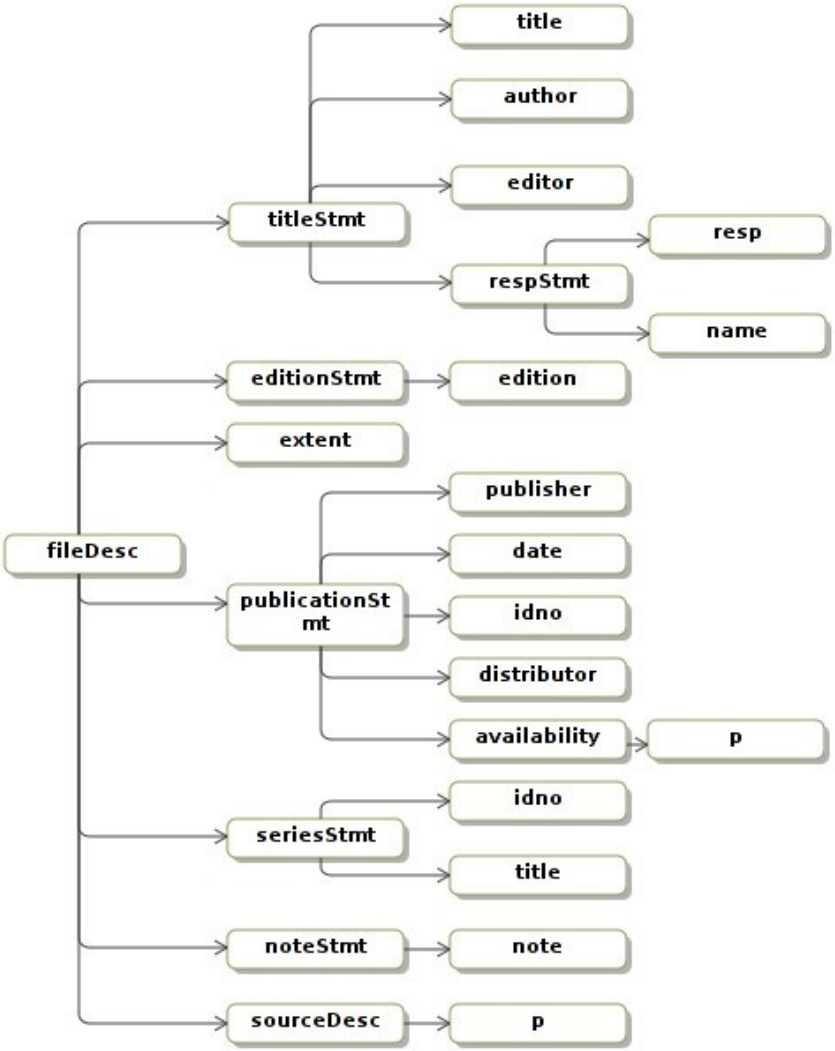


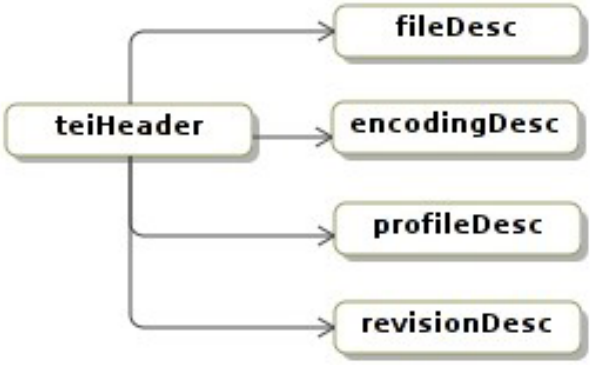
VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

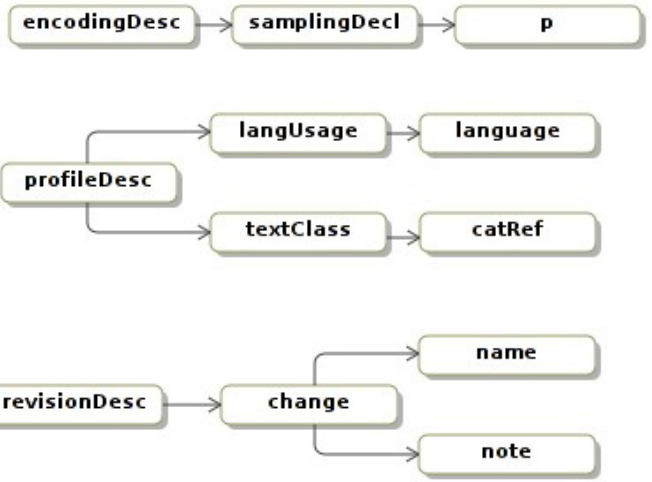
Annotation at the document metadata level – main issues

- Design of the TEI P5 conformant header (<teiHeader>) structure, answering CCLL needs
  - The main constituent parts of a TEI-conformant header (*<fileDesc>*, *<encodingDesc>*, *<profileDesc>* and *<revisionDesc>*) flexible enough to cover all the necessary elements for presenting bibliographical and non-bibliographical description of an electronic text, relationship between the electronic text and its source and the file revision history
  - Quite some of the elements could be described in several alternative ways according to TEI P5
  - Where needed, additional description elements were added to the TEI document header part.
- Design of an automatic conversion tool for the old proprietary CCLL format
- Semi-manual procedure for entering new <teiHeader> fields
- Text taxonomy redesigned according to TEI P5 classification declaration recommendations

<teiHeader> structure for CCLL

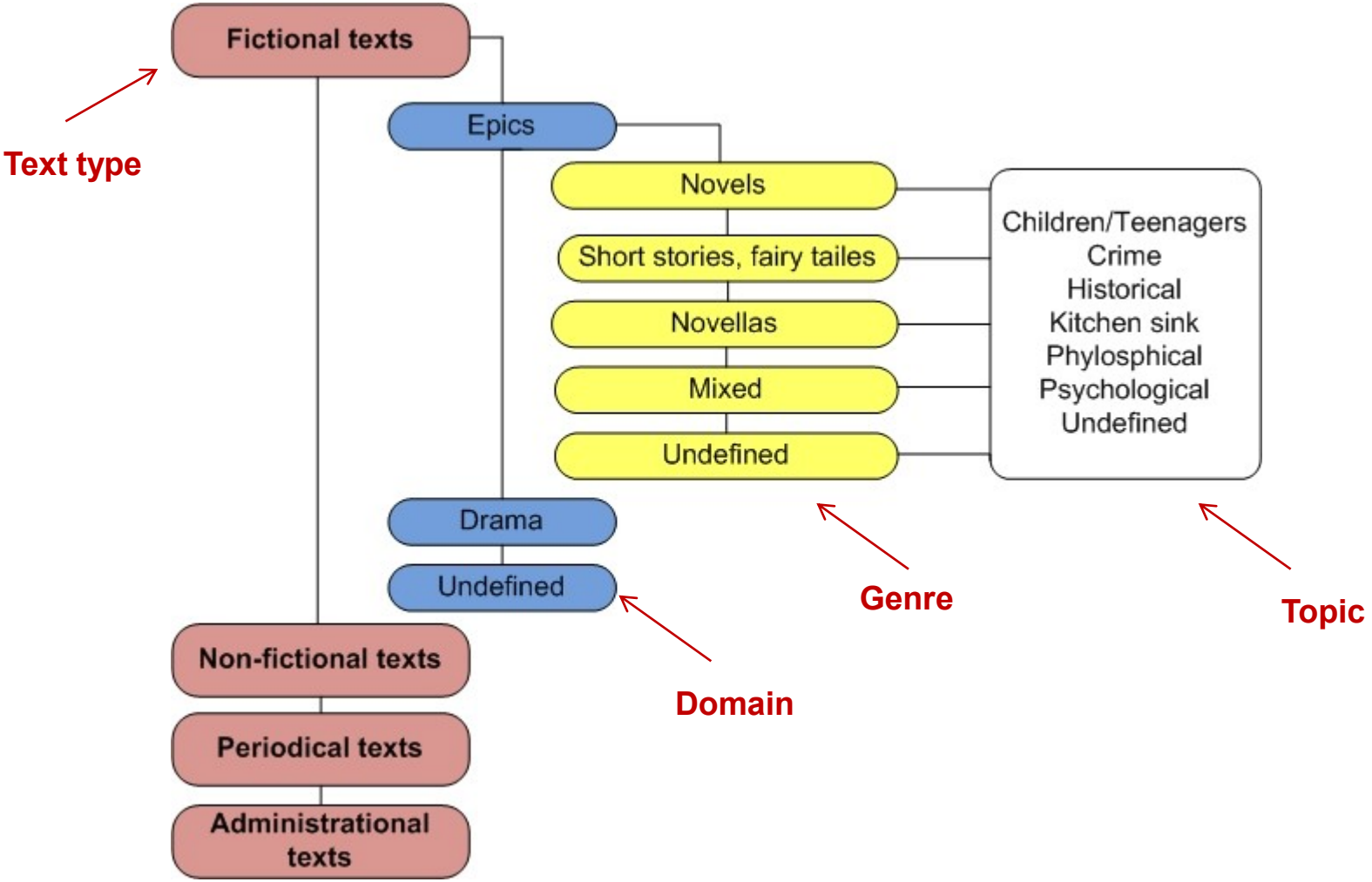






VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Text Taxonomy used by CCLL

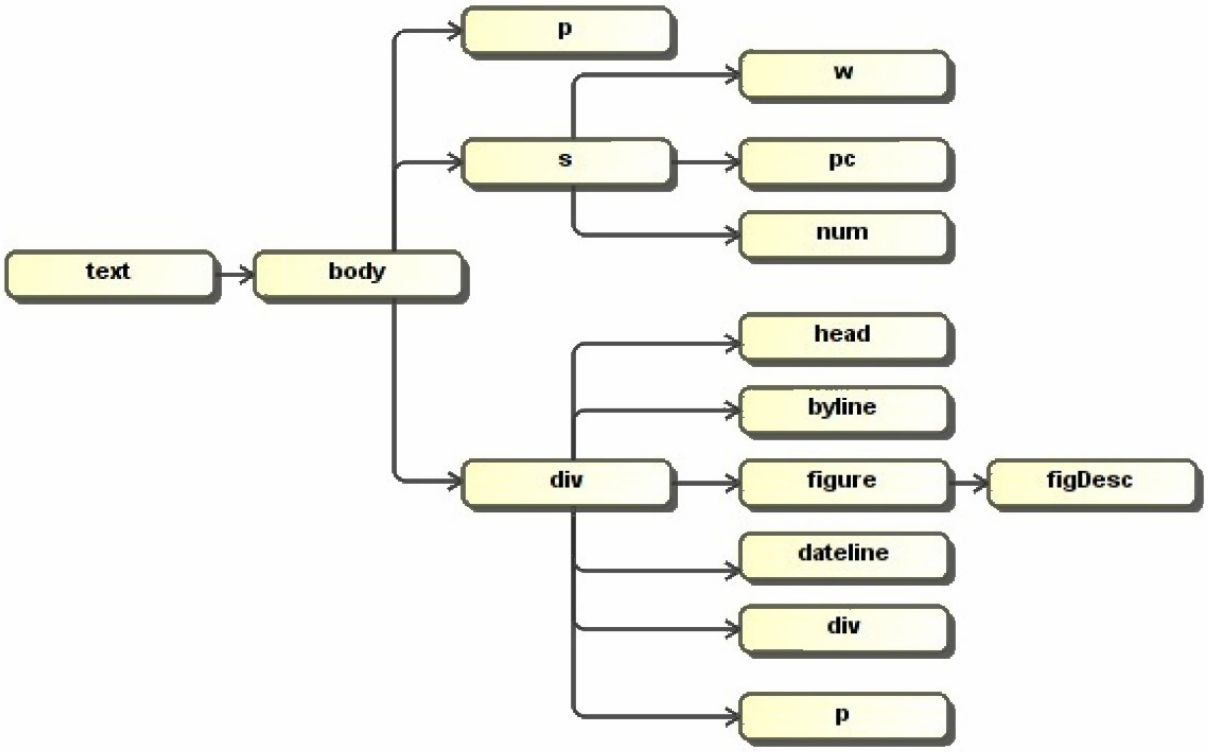


Annotation at the text structure level (1)

- Encoding of structure in serial composite publications, e.g. texts in newspapers or magazines
- Main issues:
  - Such composite electronic texts contain corresponding hierarchical structures of component elements – textual divisions and subdivisions,
  - Many different electronic sources – a variety of different formats to convert to the defined TEI-conformant text structure
  - requires the selection of a rather universal TEI element subset, capable of covering different structural aspects of serial publications,
  - Corresponding automatic conversion tools have to be designed.

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Annotation at the text structure level (2)



Structure is based on a nested set of elements, usually representing columns (rubrics), articles and paragraphs

Morphosyntactic annotation (1)

- Main issues:
  - morphological analysis of the CCLL is carried out automatically by a morphological annotation tool *(tagger)*,
  - In order to solve the ambiguity problem, 1 m word morphologically annotated corpus has been created for training the tagger,
- Morphological annotation is executed as word-level markup, using context disambiguated lemmas and morphosyntactic definitions (MSDs)
  - e.g., *<w lemma="vyriausybė" ana="#dbmvk">vyriausybės</w>*.
- The morphosyntactic specification, used for the CCLL, has been built in the form, compatible with the MULTEXT-East multilingual dataset for language engineering research and development

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

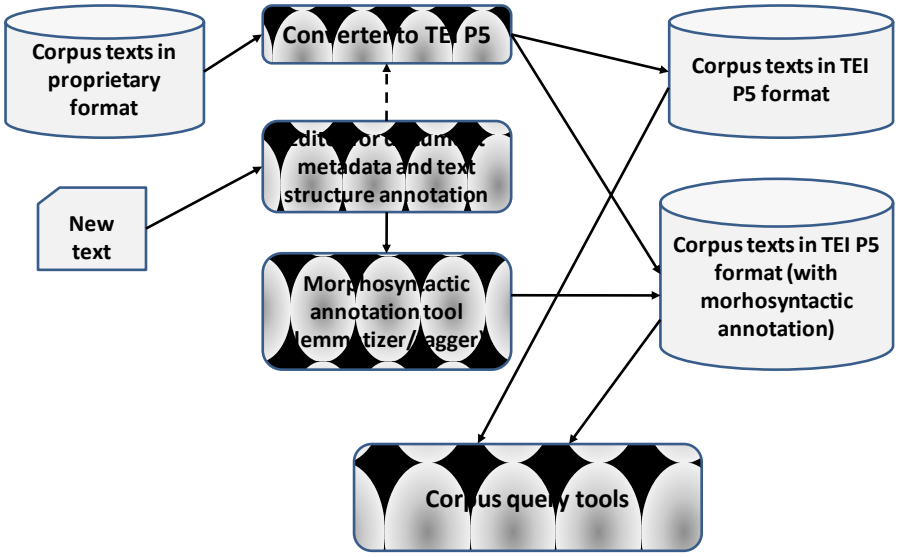
Morphosyntactic annotation (2)

Each MSD is linked to a TEI feature-structure library, which describes the decomposition into morphological features:

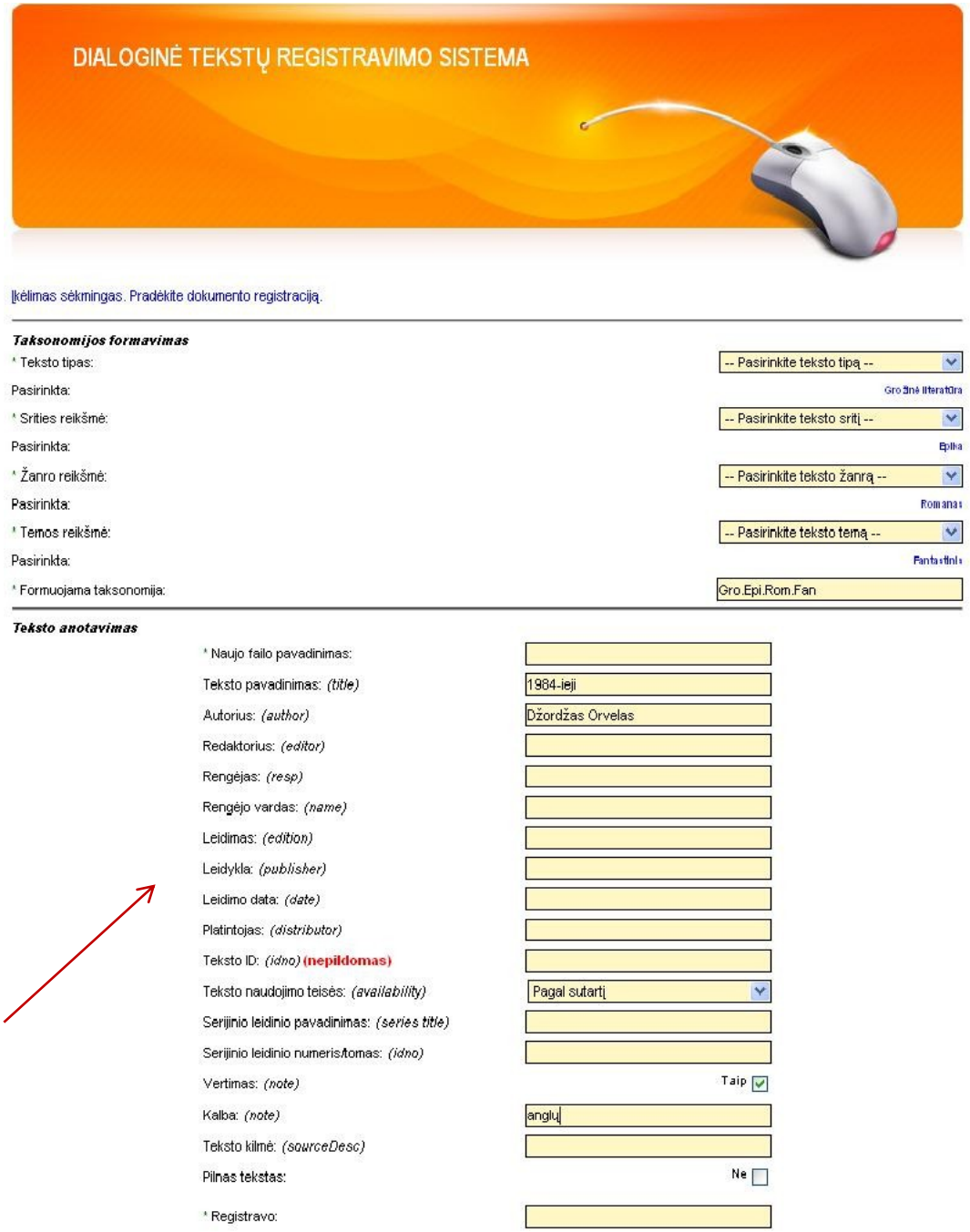
```
<fs xml:id="dbmvk" xml:lang="lt" feats="#P1.1 #P2.2 #P10.1 #P11.1 #P12.2"/>
<f name="POS" xml:id="P1.1" xml:lang="lt"><symbol value="dktv."/></f>
<f name="Voice" xml:id="P2.2" xml:lang="lt"><symbol value="bend."/></f>
<f name="Gender" xml:id="P10.1" xml:lang="lt"><symbol value="mot.g."/></f>
<f name="Number" xml:id="P11.1" xml:lang="lt"><symbol value="vns."/></f>
<f name="Case" xml:id="P12.2" xml:lang="lt"><symbol value="klm."/></f>
....
```

Supporting tools

- The CCLL is equipped with a set of software tools, falling into two main categories:
  - Tools for annotating and managing the CCLL;
  - Tools for the CCLL query and analysis.



Tool demo - annotation



Taxonomy

Header

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Tool demo - annotation

XML editor

Pasirinkto dokumento turinys

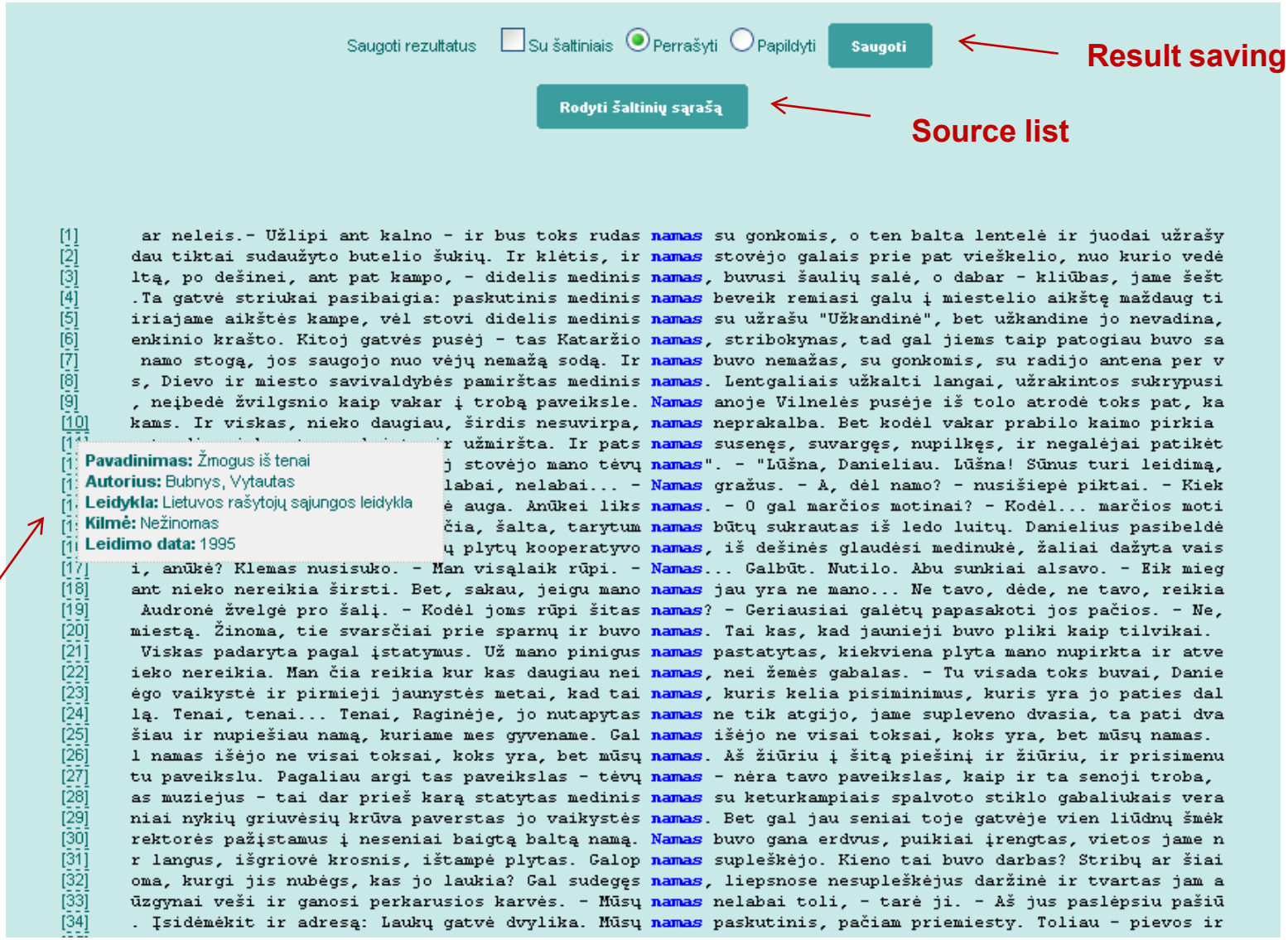
```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Rinkos visuomenė</title>
                <author>Don Slater, Fran Tonkiss</author>
                <editor>Vytautas Rubavičius</editor>
                <respStmt>
                    <resp>vertėja</resp>
                    <name>Rūta Tumėnaitė</name>
                </respStmt>
            </titleStmt>
            <editionStmt>
                <edition>1</edition>
            </editionStmt>
            <extent>65226</extent>
```

Laukų užpildymas

* Naujo failo pavadinimas:	
Teksto pavadinimas: *(title)*	Rinkos visuomenė
Autorius: *(author)*	Don Slater, Fran Tonkiss
Rengėjas: *(editor)*	Vytautas Rubavičius
Vertėjas: *(resp)*	vertėja
Vertėjo vardas: *(name)*	Rūta Tumėnaitė
Leidimas: *(edition)*	1
Leidykla: *(publisher)*	Lietuvos rašytojų sąjungos leidykla
Leidimo data: *(date)*	2004
Organizacija: *(authority)*	Lietuvos rašytojų sąjungos leidykla
Teksto naudojimas: *(availability)*	Pagal sutartį
Iš kokios kalbos verčiamas tekstas: *(note)*	
Teksto kilmė: *(sourceDesc)*	Slater.doc
Ar registruojamas pilnas tekstas: *(samplingDesc)*	Pilnas tekstas / Ne
* Registravo:	
* Ar bus atliekami pakeitimai?	Ne

Registruoti

Tool demo - concordancing



[1] ar neleis.- Užlipi ant kalno - ir bus toks rudas namas su gonkomis, o ten balta lentelė ir juodai užraš
[2] dau tiktai sudaužyto butelio šukių. Ir klėtis, ir namas stovėjo galais prie pat vieškelio, nuo kurio vedė
[3] ltą, po dešinei, ant pat kampo, - didelis medinis namas, buvusi šaulių salė, o dabar - kliūbas, jame šešt
[4] .Ta gatvė striukai pasibaigia: paskutinis medinis namas beveik remiasi galu į miestelio aikštę maždaug ti
[5] iriajame aikštės kampe, vėl stovi didelis medinis namas su užrašu "Užkandinė", bet užkandine jo nevadina,
[6] enkinio krašto. Kitoj gatvės pusėj - tas Kataržio namas, stribokynas, tad gal jiems taip patogiau buvo sa
[7] namo stogą, jos saugojo nuo vėjų nemažą sodą. Ir namas buvo nemažas, su gonkomis, su radijo antena per v
[8] s, Dievo ir miesto savivaldybės pamirštas medinis namas. Lentgaliais užkalti langai, užrakintos sukrypusi
[9] , neįbedė žvilgsnio kaip vakar į trobą paveiksle. Namas anoje Vilnelės pusėje iš tolo atrodė toks pat, ka
[10] kams. Ir viskas, nieko daugiau, širdis nesuvirpa, namas neprakalba. Bet kodėl vakar prabilo kaimo pirkia
[11] užmiršta. Ir pats namas susenęs, suvargęs, nupilkęs, ir negalėjai patikėt
[1] stovėjo mano tėvų namas". - "Lūšna, Danieliau. Lūšna! Sūnus turi leidimą,
[1] labai, nelabai... - Namas gražus. - A, dėl namo? - nusišiepė piktai. - Kiek
[1] auga. Anūkei liks namas. - O gal marčios motinai? - Kodėl... marčios moti
[1] čia, šalta, tarytum namas būtų sukrautas iš ledo luitų. Danielius pasibeldė
[1] plytų kooperatyvo namas, iš dešinės glaudėsi medinukė, žaliai dažyta vais
[17] i, anūkė? Klemas nusisuko. - Man visąlaik rūpi. - Namas... Galbūt. Nutilo. Abu sunkiai alsavo. - Eik mieg
[18] ant nieko nereikia širsti. Bet, sakau, jeigu mano namas jau yra ne mano... Ne tavo, dėde, ne tavo, reikia
[19] Audronė žvelgė pro šalį. - Kodėl joms rūpi šitas namas? - Geriausiai galėtų papasakoti jos pačios. - Ne,
[20] miestą. Žinoma, tie svarsčiai prie sparnų ir buvo namas. Tai kas, kad jaunieji buvo pliki kaip tilvikai.
[21] Viskas padaryta pagal įstatymus. Už mano pinigus namas pastatytas, kiekviena plyta mano nupirkta ir atve
[22] ieko nereikia. Man čia reikia kur kas daugiau nei namas, nei žemės gabalas. - Tu visada toks buvai, Danie
[23] ėgo vaikystė ir pirmieji jaunystės metai, kad tai namas, kuris kelia pisiminimus, kuris yra jo paties dal
[24] lą. Tenai, tenai... Tenai, Raginėje, jo nutapytas namas ne tik atgijo, jame supleveno dvasia, ta pati dva
[25] šiau ir nupiešiau namą, kuriame mes gyvename. Gal namas išėjo ne visai toksai, koks yra, bet mūsų namas.
[26] l namas išėjo ne visai toksai, koks yra, bet mūsų namas. Aš žiūriu į šitą piešinį ir žiūriu, ir prisimenu
[27] tu paveikslu. Pagaliau argi tas paveikslas - tėvų namas - nėra tavo paveikslas, kaip ir ta senoji troba,
[28] as muziejus - tai dar prieš karą statytas medinis namas su keturkampiais spalvoto stiklo gabaliukais vera
[29] niai nykių griuvėsių krūva paverstas jo vaikystės namas. Bet gal jau seniai toje gatvėje vien liūdnų šmėk
[30] rektorės pažįstamus į neseniai baigtą baltą namą. Namas buvo gana erdvus, puikiai įrengtas, vietos jame n
[31] r langus, išgriovė krosnis, ištampė plytas. Galop namas supleškėjo. Kieno tai buvo darbas? Stribų ar šiai
[32] oma, kurgi jis nubėgs, kas jo laukia? Gal sudegęs namas, liepsnose nesupleškėjus daržinė ir tvartas jam a
[33] ūzgynai veši ir ganosi perkarusios karvės. - Mūsų namas nelabai toli, - tarė ji. - Aš jus paslėpsiu pašiū
[34] . Įsidėmėkit ir adresą: Laukų gatvė dvylika. Mūsų namas paskutinis, pačiam priemiesty. Toliau - pievos ir

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Conlusions

- The process of transformation of the CCLL into a new standard has proved to be a complicated, but necessary step in the development of the corpus.
- Whereas this task is rather difficult and time consuming endeavor, it may be noted that selection of an appropriate format from several candidate standards depends not only on functionalities of standards, but also on how well they are documented.
- In this aspect, TEI P5 standard stands out as a very well documented standard.
- Further CCLL development plans to include additional annotation levels, namely syntactic and semantic metadata, mark-up of collocations, named entities and other textual elements, necessary for various corpus-based natural language processing tasks.
- Preliminary investigation has shown, that TEI P5 encoding scheme includes elements necessary for such annotation.

VYTAUTAS MAGNUS UNIVERSITY
MCMXXII

Thank you!

Contacts:

e.rimkute@hmf.vdu.lt,

j.kovalevskaite@hmf.vdu.lt,

a.utka@hmf.vdu.lt,

v.melninkaite@if.vdu.lt,

d.vitkute@if.vdu.lt