Building a digital lexicon, trial and error, the cases of Latin and Avestan

Armin Hoenen
Avestan subsection: Thomas Jügel
Goethe-University Frankfurt

2011 – 10 – 10

The Aim

eLexicon

For a variety of usecases we apply our flexible **eLexicon** data model

- Late Latin Lexicon
- Avestan
- Hugo von Hofmannsthal

The Aim

eLexicon

For a variety of usecases we apply our flexible **eLexicon** data model

- Late Latin Lexicon
- Avestan
- Hugo von Hofmannsthal

The Aim

eLexicon

For a variety of usecases we apply our flexible **eLexicon** data model

- Late Latin Lexicon
- Avestan
- Hugo von Hofmannsthal

The Aim

eLexicon

For a variety of usecases we apply our flexible **eLexicon** data model

- Late Latin Lexicon
- Avestan
- Hugo von Hofmannsthal

The Aim

eLexicon

For a variety of usecases we apply our flexible **eLexicon** data model

- Late Latin Lexicon
- Avestan
- Hugo von Hofmannsthal

data modelling

ANSI/X3/SPARC

eLexicon

- conceptual layer: eLexicon Data Model
- logical layer: Relational Database model, RDF XML model
- physical layer: concrete **MySQL schema**, **RDF schema**

data modelling

ANSI/X3/SPARC

eLexicon

- conceptual layer: eLexicon Data Model
- logical layer: Relational Database model, RDF XML model
- physical layer: concrete **MySQL schema**, **RDF schema**

data modelling

ANSI/X3/SPARC

eLexicon

- conceptual layer: eLexicon Data Model
- logical layer: Relational Database model, RDF XML model
- physical layer: concrete **MySQL schema**, **RDF schema**

data modelling

ANSI/X3/SPARC

eLexicon

- conceptual layer: eLexicon Data Model
- logical layer: Relational Database model, RDF XML model
- physical layer: concrete **MySQL schema**, **RDF schema**

data modelling

ANSI/X3/SPARC

eLexicon

- conceptual layer: eLexicon Data Model
- logical layer: Relational Database model, RDF XML model
- physical layer: concrete **MySQL schema**, **RDF schema**

the eLexicon

- frater

- object, id: 20

id	name
2	Genitive
...	...
20	frater
21	fratris

object

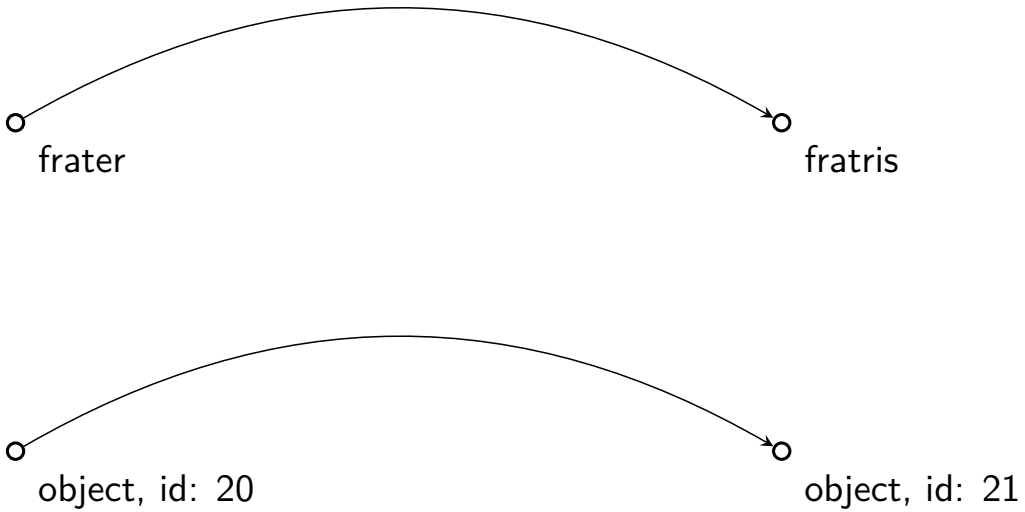
id	mother id	daughter id
0	20	21

relation

relation id	object id
0	2
0	12

dimension

the eLexicon



id	name
2	Genitive
...	...
20	frater
21	fratris

object

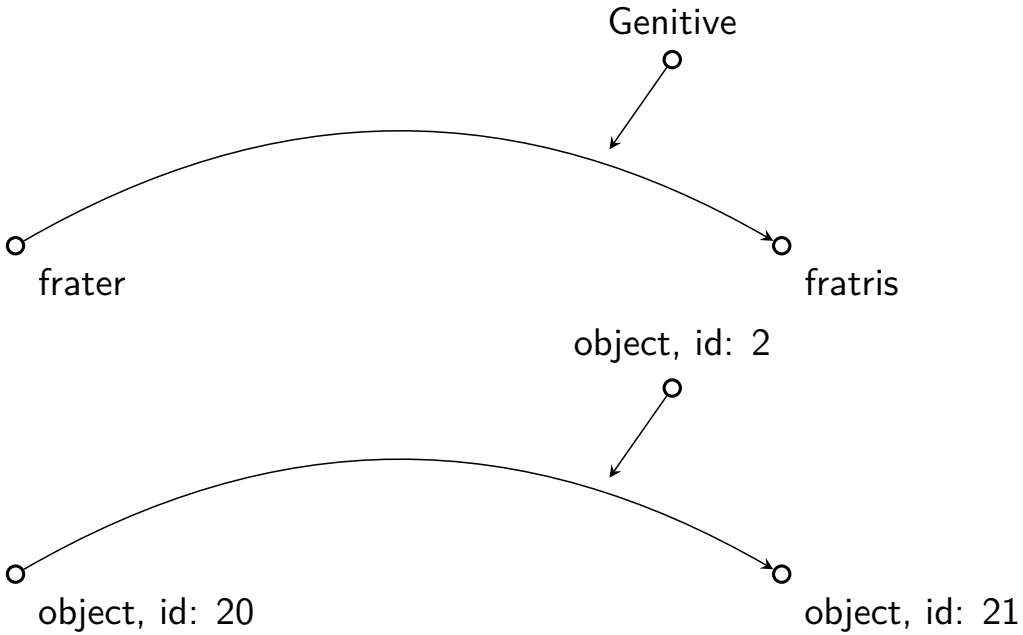
id	mother id	daughter id
0	20	21

relation

relation id	object id
0	2
0	12

dimension

the eLexicon



object

id	name
2	Genitive
…	…
20	frater
21	fratris

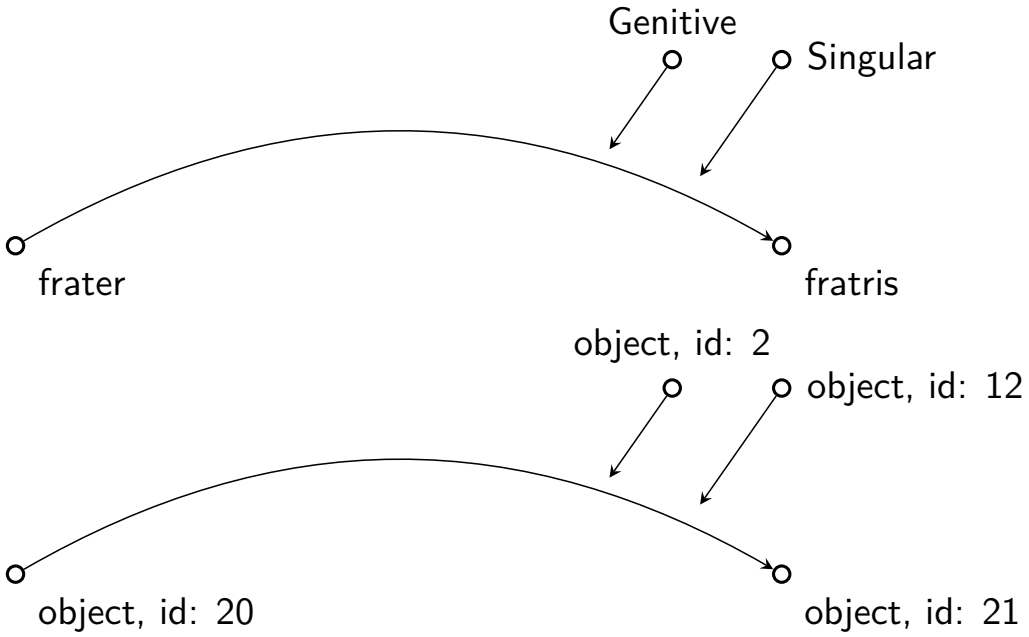
relation

id	mother id	daughter id
0	20	21

dimension

relation id	object id
0	2
0	12

the eLexicon



object

id	name
2	Genitive
...	...
20	frater
21	fratris

relation

id	mother id	daughter id
0	20	21

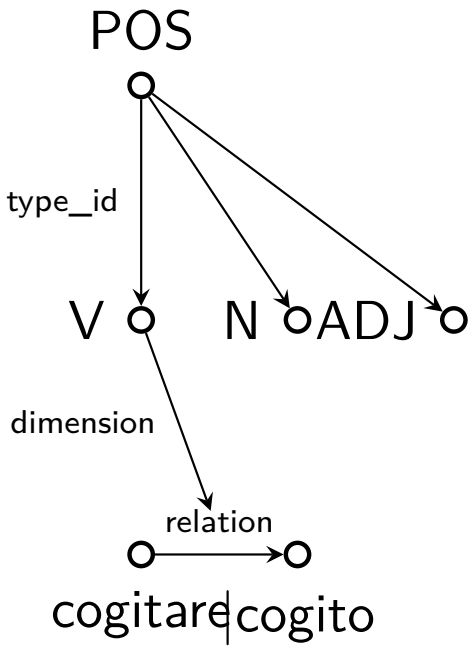
dimension

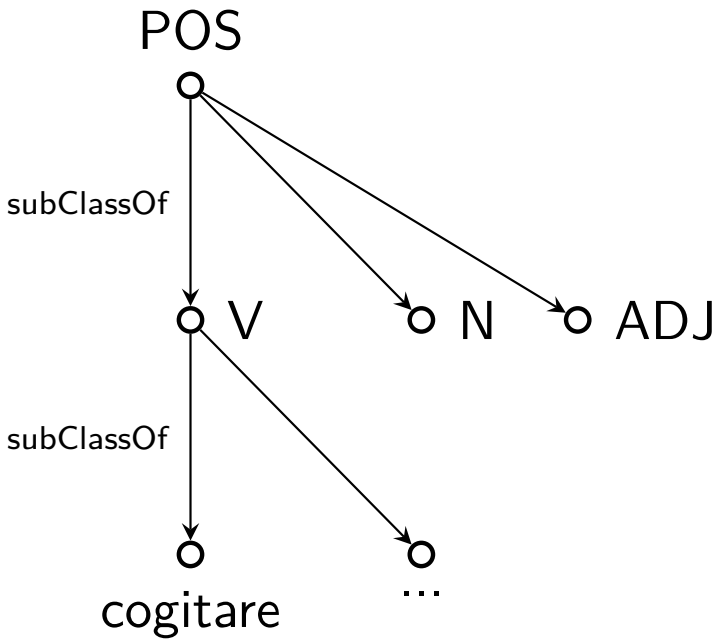
relation id	object id
0	2
0	12

Resource Description Framework

As an export format for scientific exchange, we use RDF.

RDFS





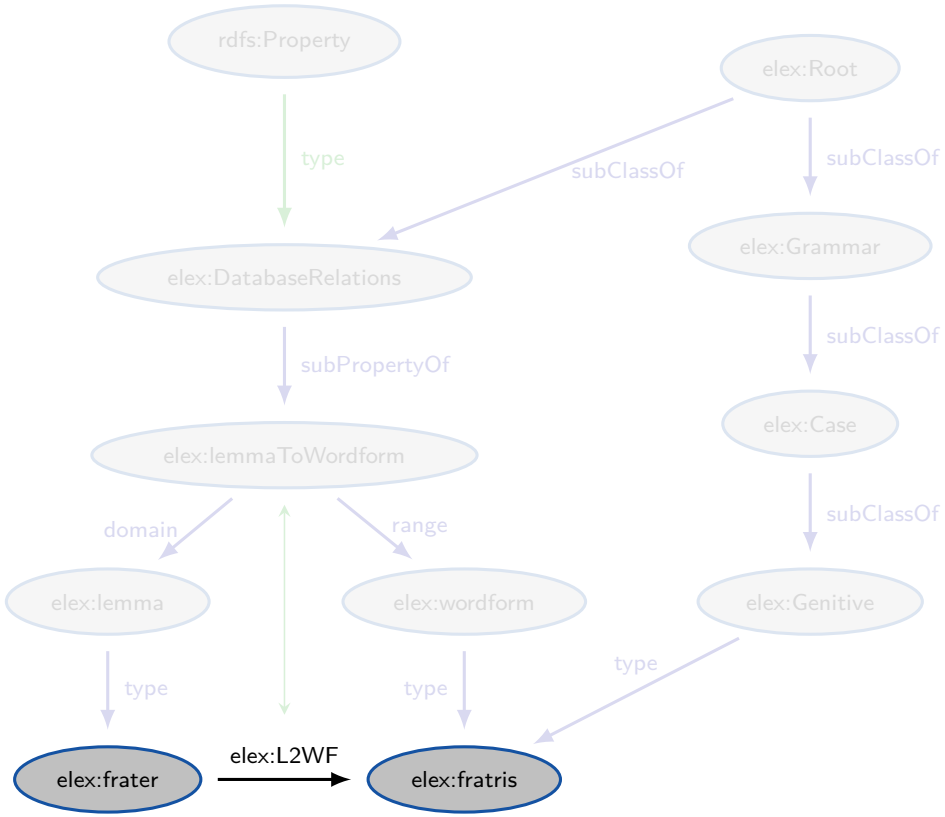
eLexicon»»»»»»»»»»»»»RDF representation

Our RDF model

2 intuitive files

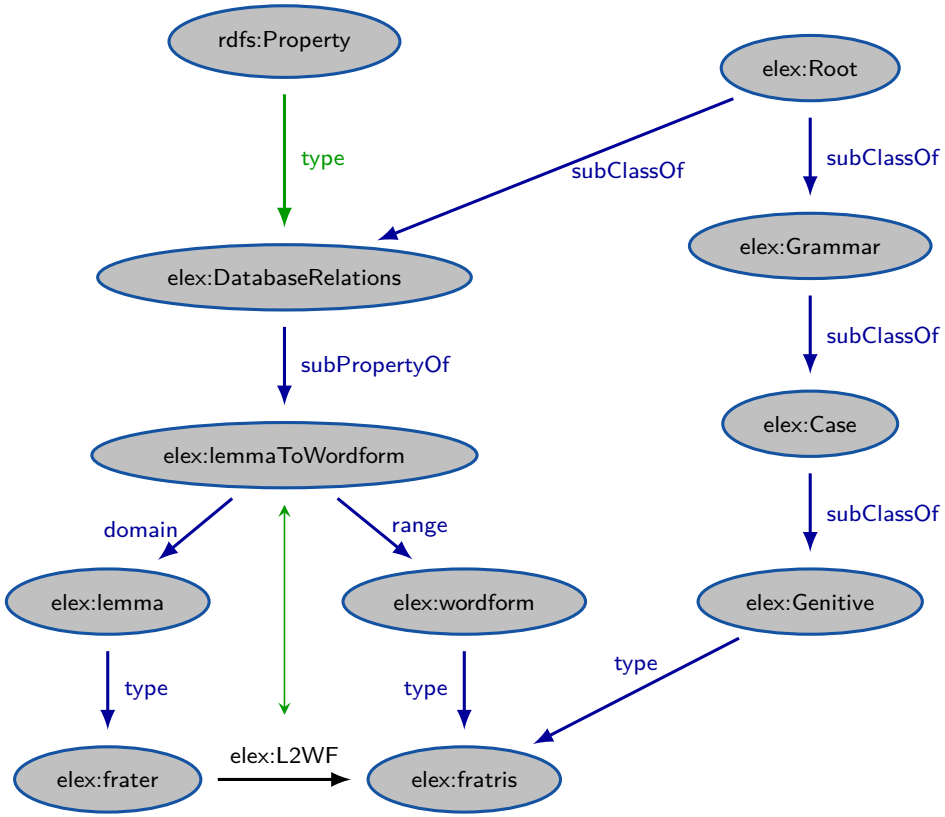
- Ontology (STTS, Olia, DCR Gold, OntoTag, TDS + special types)
- Lexicon (divisible)

The Model




```
<rdf:Description rdf:about="int_id_4">
      <rdfs:label>fratris</rdfs:label>
      <rdf:type rdf:resource="id_183"/>
      <rdf:type rdf:resource="id_12"/>
      <rdf:type rdf:resource="id_2"/>
</rdf:Description>
```

The Model




```
<rdf:Description rdf:about="int_id_4">
      <rdfs:label>fratris</rdfs:label>
      <rdf:type rdf:resource="id_183"/>
      <rdf:type rdf:resource="id_12"/>
      <rdf:type rdf:resource="id_2"/>
</rdf:Description>
```

Lemma

Lemma exemplified:

- lemmas contain one entry for each corresponding wordform
- wordforms and lemmas contain class references

```
<rdf:Description rdf:about="int_id_18154">
      <rdfs:subClassOf rdf:resource="elex;id_185"/>
      <rdf:type rdf:resource="elex;id_78"/>
      <rdfs:label>absento</rdfs:label>
      <rdfs:comment>absento, absentare</rdfs:comment>
      <rdfs:comment xml:lang="EN">be absent</rdfs:comment>
      <elex:id_216 rdf:resource="int_id_17860"/>
      <elex:id_216 rdf:resource="int_id_17861"/>
      <elex:id_216 rdf:resource="int_id_17862"/>
      <elex:id_216 rdf:resource="int_id_17863"/>
      <elex:id_216 rdf:resource="int_id_17864"/>
</rdf:Description>
```

2 intuitive files

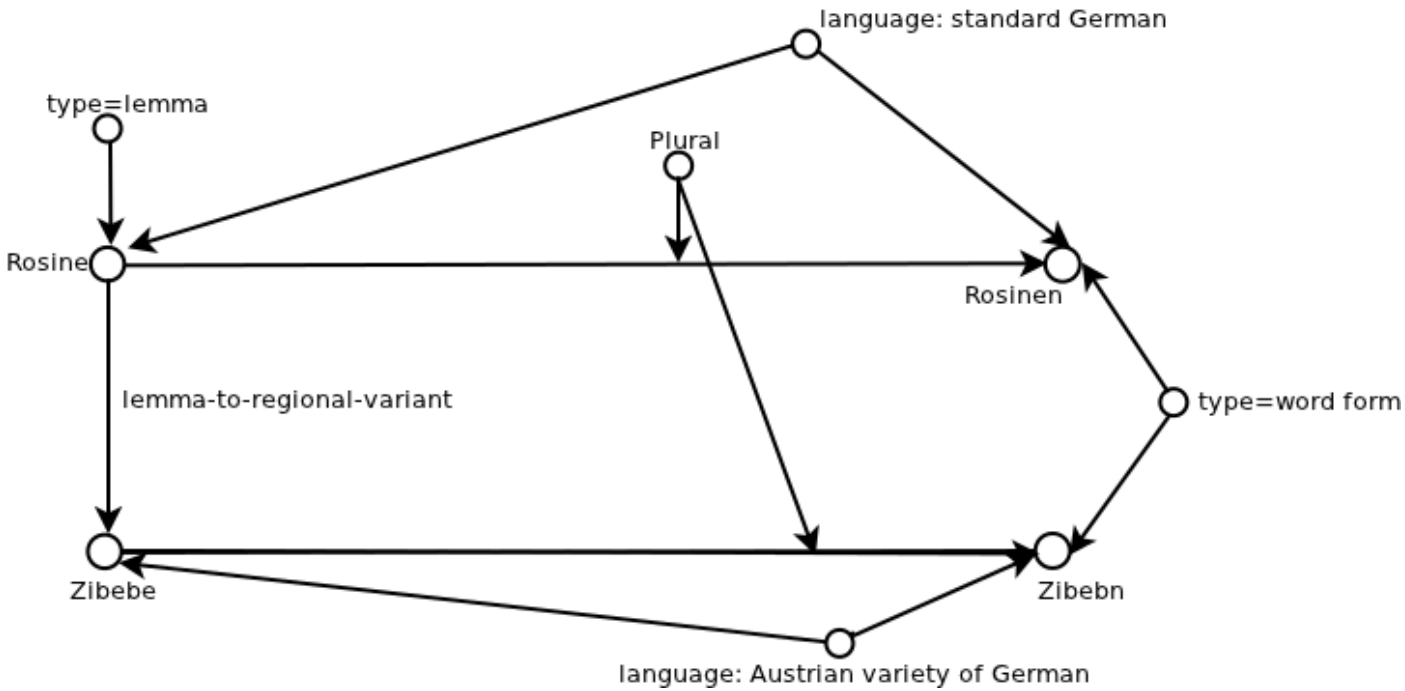
- Ontology (STTS, Olia, DCR Gold, OntoTag, TDS + special types)
- Lexicon (divisible)

dealing with special characteristics

historical corpora

- word form orthographic variant (type)
- earlier- inflection vs. later- inflection (types)
- era (type)
- new variants like genre, dialect etc.

Example



the importance of an abstract lemma

orthographic variants to the standard

sruiie.

the importance of an abstract lemma

orthographic variants to the standard

sruiie.

sruiiae.

the importance of an abstract lemma

orthographic variants to the standard

sruiie.

sruiiae.

sruue.

the importance of an abstract lemma

orthographic variants to the standard

sruiie.

sruiiae.

sruue.

sraoe.

the importance of an abstract lemma

orthographic variants to the standard

sruiie.

sruiiae.

sruue.

sraoe.

sraoiie.

the importance of an abstract lemma

orthographic variants to the standard

sruiie.

sruiiae.
sruue.
sraoe.
sraoiie.
sraoi.

the importance of an abstract lemma

orthographic variants to the standard

sruiie.

sruiiae.
sruue.
sraoe.
sraoiie.
sraoi.
sraōi.

the importance of an abstract lemma

orthographic variants to the standard

sruiie.

sruiiae.
sruue.
sraoe.
sraoiie.
sraoi.
sraōi.
srōi.

the importance of an abstract lemma

orthographic variants to the standard

sruiie.

sruiiae.
sruue.
sraoe.
sraoiie.
sraoi.
sraōi.
srōi.
sarauuae.
affiliation of manuscripts

the importance of an abstract lemma

stem variants as a sign of language development

the importance of an abstract lemma

stem variants as a sign of language development

1. napāt- „grandchild“ → inherited

the importance of an abstract lemma

stem variants as a sign of language development

1. napāt- „grandchild“ → inherited
2. naptar- → semantic class assimilation
   (cf. pitar-, brātar-, mātar-, etc.)

the importance of an abstract lemma

stem variants as a sign of language development

1. napāt- „grandchild“ → inherited
2. naptar- → semantic class assimilation (cf. pitar-, brātar-, mātar-, etc.)
3. napa- → generalization of A-stems

the importance of an abstract lemma

ablaut and fusion of morphemes

$\sqrt{}$ vac “to say”

the importance of an abstract lemma

ablaut and fusion of morphemes

√ vac “to say”

vašaŋhe	2sg. ind. pres. m.

the importance of an abstract lemma

ablaut and fusion of morphemes

√ vac “to say”

vašaŋhe	2sg. ind. pres. m.
vaxšiiā	1sg. ind. fut. a.

the importance of an abstract lemma

ablaut and fusion of morphemes

√ vac “to say”

vašaŋhe	2sg. ind. pres. m.
vaxšiiā	1sg. ind. fut. a.
frauuaočəm	1sg. inj. aor. a.

the importance of an abstract lemma

ablaut and fusion of morphemes

√ vac “to say”

vašaŋhe	2sg. ind. pres. m.
vaxšiiā	1sg. ind. fut. a.
frauuaočəm	1sg. inj. aor. a.
vaočat̲	3sg. inj. aor. a.

the importance of an abstract lemma

ablaut and fusion of morphemes

√ vac “to say”

vašaŋhe	2sg. ind. pres. m.
vaxšiiā	1sg. ind. fut. a.
frauuaočəm	1sg. inj. aor. a.
vaočat̲	3sg. inj. aor. a.
auuāčī	3sg. inj. aor. p.

the importance of an abstract lemma

ablaut and fusion of morphemes

√ vac “to say”

vašaŋhe	2sg. ind. pres. m.
vaxšiiā	1sg. ind. fut. a.
frauuaočəm	1sg. inj. aor. a.
vaočat̲	3sg. inj. aor. a.
auuāčī	3sg. inj. aor. p.
vauuaxδa	2sg. ind. perf. a.

the importance of an abstract lemma

ablaut and fusion of morphemes

√ vac “to say”

vašaŋhe	2sg. ind. pres. m.
vaxšiiā	1sg. ind. fut. a.
frauuaočəm	1sg. inj. aor. a.
vaočat̲	3sg. inj. aor. a.
auuāčī	3sg. inj. aor. p.
vauuaxδa	2sg. ind. perf. a.
vaoxəmā	1pl. ind. perf. a.

the importance of an abstract lemma

ablaut and fusion of morphemes

√ vac “to say”

vašaŋhe	2sg. ind. pres. m.
vaxšiiā	1sg. ind. fut. a.
frauuaočəm	1sg. inj. aor. a.
vaočat̲	3sg. inj. aor. a.
auuāčī	3sg. inj. aor. p.
vauuaxδa	2sg. ind. perf. a.
vaoxəmā	1pl. ind. perf. a.
fraoxta-	past participle

Approach I

merger of existing lexica

1. *Perseus*: 10,186,278 Latin Words [2]
2. *LEMLAT*: 399,226 wordforms 34,625 lemmata
3. *Words*: 39,000 entries [3]

Problems

→ Problems

- different sets of POS (Tagsets)
- different orthographic varieties
- different treatment of clitics
- erroneous analyses

Approach II

morphological Expansion:

1. automatic expansion of regular forms according to the word class
2. manual or semi-automatic expansion of irregular forms
3. removal of duplicates
4. **human control**

characteristics: overgeneration

Approach II

morphological Expansion:

1. automatic expansion of regular forms according to the word class
2. manual or semi-automatic expansion of irregular forms
3. removal of duplicates
4. **human control**

characteristics: overgeneration

Approach II

morphological Expansion:

1. automatic expansion of regular forms according to the word class
2. manual or semi-automatic expansion of irregular forms
3. removal of duplicates
4. **human control**

characteristics: overgeneration

Approach II

morphological Expansion:

1. automatic expansion of regular forms according to the word class
2. manual or semi-automatic expansion of irregular forms
3. removal of duplicates
4. **human control**

characteristics: overgeneration

Approach II

morphological Expansion:

1. automatic expansion of regular forms according to the word class
2. manual or semi-automatic expansion of irregular forms
3. removal of duplicates
4. **human control**

characteristics: overgeneration

Approach II

morphological Expansion:

1. automatic expansion of regular forms according to the word class
2. manual or semi-automatic expansion of irregular forms
3. removal of duplicates
4. **human control**

characteristics: overgeneration

analytic approaches

caesar*e* **noun m.***ablative*

- lex (open class): caesar
- lex (closed class) + morpheme lexicon
- rules: derivation rules; phonological processes; classes; exceptions
- **advantage**: neatly small lexica
- **disadvantage**: hidden errors or ambiguities might persist unnoticed

generative approaches

caesare

- one merged lexicon, no bound morphemes, only word forms
- **advantage**: better control of forms in lexion; allow only controlled forms to enter
- **disadvantage**: large lexicon

method: morphological Expansion

Shoebox


```
\lx gebhaa
\ps n
\sn 1
\ge husband
\lt big person.
\lf SynD = namorit
\le Rana dialect
\sn 2
\ge clan_head
\lf SynD = tean elen
\le Rana dialect
\mr geba-haa
\cf haa
\ce big, important, loud
```

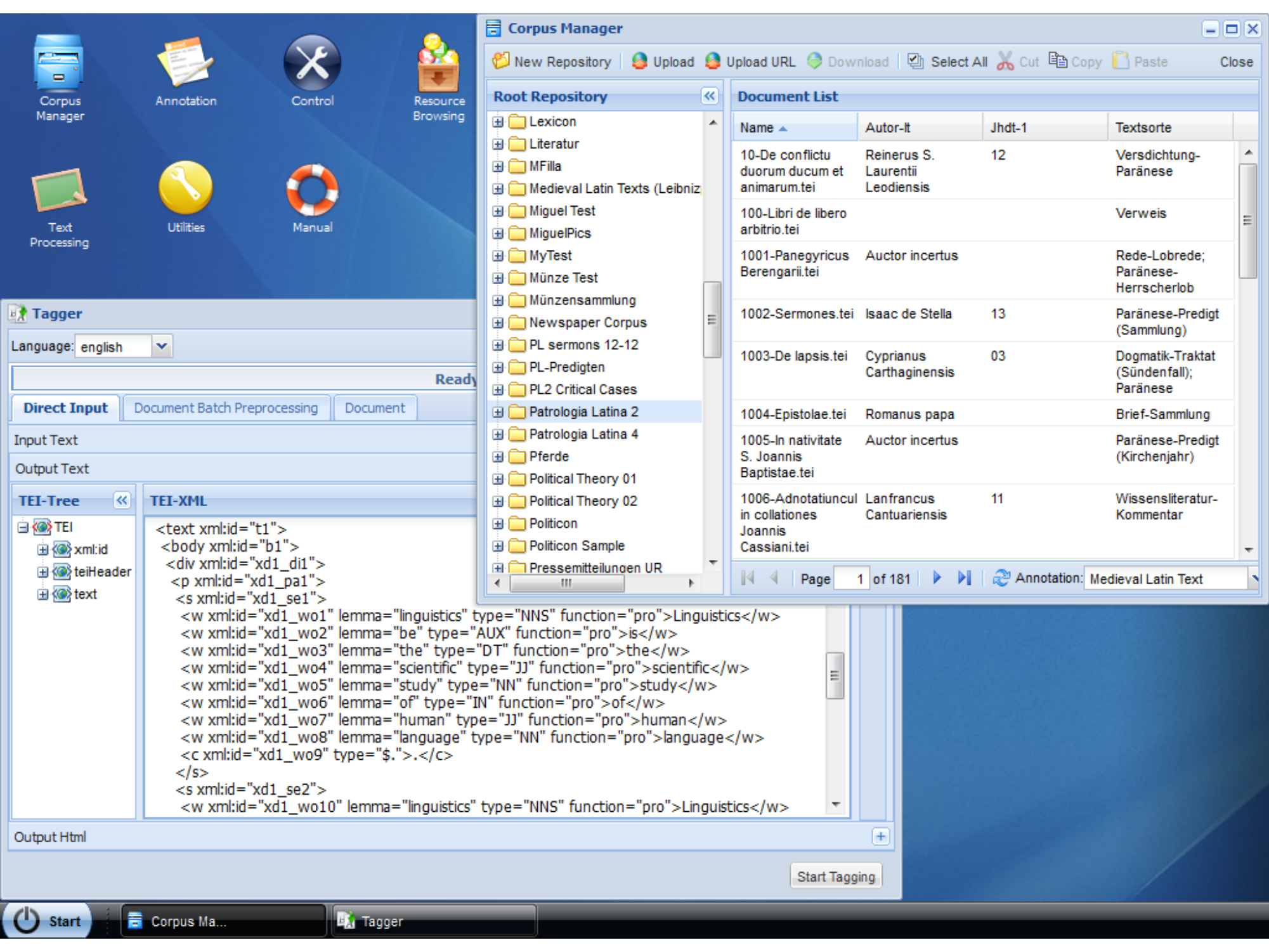
conversion??

Shoebox	eLex
lx	lemma (type)
lc	citation form (type) + relation (?)
ph	as language variety; spoken form relation (?)
ps	single part of speeches
...	...

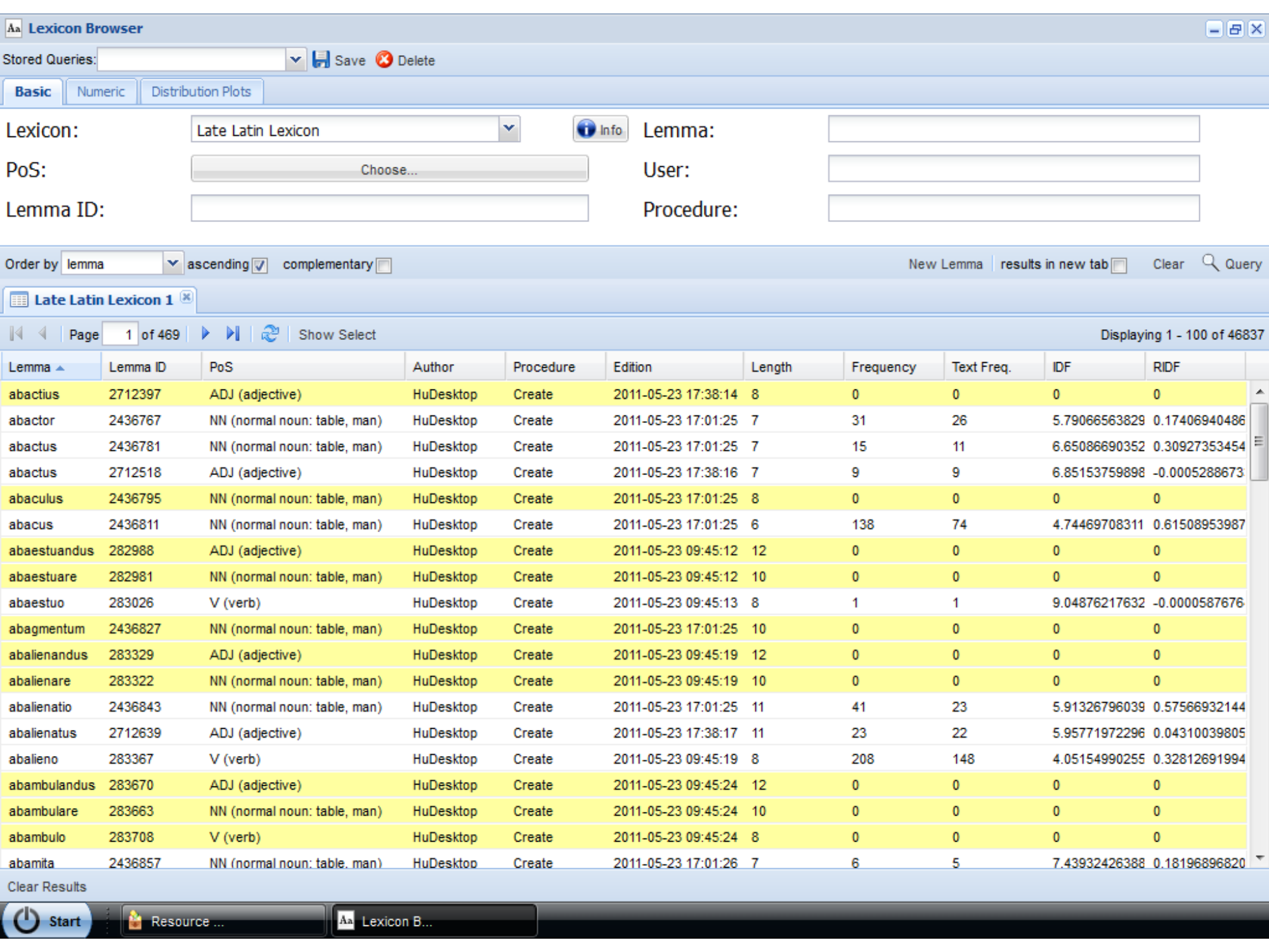
general conversion to eLex viz. RDF possible, if problematic cases can be discussed on a blog

particularities

1. lots of self-defined labels
2. shoebox as annotation tool


Corpus Manager
Annotation
Control
Resource Browsing
Text Processing
Utilities
Manual
Corpus Manager
New Repository
Upload
Upload URL
Download
Select All
Cut
Copy
Paste
Close
Root Repository
Lexicon
Literatur
MFilla
Medieval Latin Texts (Leibniz
Miguel Test
MiguelPics
MyTest
Münze Test
Münzensammlung
Newspaper Corpus
PL sermons 12-12
PL-Predigten
PL2 Critical Cases
Patrologia Latina 2
Patrologia Latina 4
Pferde
Political Theory 01
Political Theory 02
Politicon
Politicon Sample
Pressemitteilungen UR
Document List
Name
Autor-lt
Jhdt-1
Textsorte
10-De conflictu duorum ducum et animarum.tei
Reinerus S. Laurentii Leodiensis
12
Versdichtung-Paränese
100-Libri de libero arbitrio.tei
Verweis
1001-Panegyricus Berengarii.tei
Auctor incertus
Rede-Lobrede; Paränese-Herrscherlob
1002-Sermones.tei
Isaac de Stella
13
Paränese-Predigt (Sammlung)
1003-De lapsis.tei
Cyprianus Carthaginensis
03
Dogmatik-Traktat (Sündenfall); Paränese
1004-Epistolae.tei
Romanus papa
Brief-Sammlung
1005-In nativitate S. Joannis Baptistae.tei
Auctor incertus
Paränese-Predigt (Kirchenjahr)
1006-Adnotatiuncul in collationes Joannis Cassiani.tei
Lanfrancus Cantuariensis
11
Wissensliteratur-Kommentar
Page 1 of 181
Annotation: Medieval Latin Text
Tagger
Language: english
Ready
Direct Input
Document Batch Preprocessing
Document
Input Text
Output Text
TEI-Tree
TEI
xml:id
teiHeader
text
TEI-XML
<text xml:id="t1">
<body xml:id="b1">

<p xml:id="xd1_pa1">
<s xml:id="xd1_se1">
<w xml:id="xd1_wo1" lemma="linguistics" type="NNS" function="pro">Linguistics</w>
<w xml:id="xd1_wo2" lemma="be" type="AUX" function="pro">is</w>
<w xml:id="xd1_wo3" lemma="the" type="DT" function="pro">the</w>
<w xml:id="xd1_wo4" lemma="scientific" type="JJ" function="pro">scientific</w>
<w xml:id="xd1_wo5" lemma="study" type="NN" function="pro">study</w>
<w xml:id="xd1_wo6" lemma="of" type="IN" function="pro">of</w>
<w xml:id="xd1_wo7" lemma="human" type="JJ" function="pro">human</w>
<w xml:id="xd1_wo8" lemma="language" type="NN" function="pro">language</w>
<c xml:id="xd1_wo9" type="$.">.</c>
</s>
<s xml:id="xd1_se2">
<w xml:id="xd1_wo10" lemma="linguistics" type="NNS" function="pro">Linguistics</w>
Output Html
Start Tagging
Start
Corpus Ma...
Tagger


Lexicon Browser
Stored Queries:
Save
Delete
Basic
Numeric
Distribution Plots
Lexicon:
Late Latin Lexicon
Info
Lemma:
PoS:
Choose...
User:
Lemma ID:
Procedure:
Order by lemma
ascending
complementary
New Lemma
results in new tab
Clear
Query
Late Latin Lexicon 1
Page 1 of 469
Show Select
Displaying 1 - 100 of 46837
Lemma
Lemma ID
PoS
Author
Procedure
Edition
Length
Frequency
Text Freq.
IDF
RIDF
abactius 2712397 ADJ (adjective) HuDesktop Create 2011-05-23 17:38:14 8 0 0 0 0
abactor 2436767 NN (normal noun: table, man) HuDesktop Create 2011-05-23 17:01:25 7 31 26 5.79066563829 0.17406940486
abactus 2436781 NN (normal noun: table, man) HuDesktop Create 2011-05-23 17:01:25 7 15 11 6.65086690352 0.30927353454
abactus 2712518 ADJ (adjective) HuDesktop Create 2011-05-23 17:38:16 7 9 9 6.85153759898 -0.0005288673
abaculus 2436795 NN (normal noun: table, man) HuDesktop Create 2011-05-23 17:01:25 8 0 0 0 0
abacus 2436811 NN (normal noun: table, man) HuDesktop Create 2011-05-23 17:01:25 6 138 74 4.74469708311 0.61508953987
abaestuandus 282988 ADJ (adjective) HuDesktop Create 2011-05-23 09:45:12 12 0 0 0 0
abaestuare 282981 NN (normal noun: table, man) HuDesktop Create 2011-05-23 09:45:12 10 0 0 0 0
abaestuo 283026 V (verb) HuDesktop Create 2011-05-23 09:45:13 8 1 1 9.04876217632 -0.000587676
abagmentum 2436827 NN (normal noun: table, man) HuDesktop Create 2011-05-23 17:01:25 10 0 0 0 0
abalienandus 283329 ADJ (adjective) HuDesktop Create 2011-05-23 09:45:19 12 0 0 0 0
abalienare 283322 NN (normal noun: table, man) HuDesktop Create 2011-05-23 09:45:19 10 0 0 0 0
abalienatio 2436843 NN (normal noun: table, man) HuDesktop Create 2011-05-23 17:01:25 11 41 23 5.91326796039 0.57566932144
abalienatus 2712639 ADJ (adjective) HuDesktop Create 2011-05-23 17:38:17 11 23 22 5.95771972296 0.04310039805
abalieno 283367 V (verb) HuDesktop Create 2011-05-23 09:45:19 8 208 148 4.05154990255 0.32812691994
abambulandus 283670 ADJ (adjective) HuDesktop Create 2011-05-23 09:45:24 12 0 0 0 0
abambulare 283663 NN (normal noun: table, man) HuDesktop Create 2011-05-23 09:45:24 10 0 0 0 0
abambulo 283708 V (verb) HuDesktop Create 2011-05-23 09:45:24 8 0 0 0 0
abamita 2436857 NN (normal noun: table, man) HuDesktop Create 2011-05-23 17:01:26 7 6 5 7.43932426388 0.18196896820
Clear Results
Start
Resource ...
Lexicon B...

Lexicon Browser

Stored Queries: Save Delete

Basic | Numeric | Distribution Plots

Lexicon: Late Latin Lexicon Info

Lemma:

PoS: Choose...

User:

Lemma ID:

Procedure:

Order by frequency ascending complementary

New Lemma | results in new tab Clear Query

Late Latin Lexicon 1 | Late Latin Lexicon - frequency | **Lemma ecclesia**

New Lemma Save Changes Delete Lemma Delete Superlemma Reload

Language	Latin	Part of Speech	normal noun: table, man
Base Part of Speech	noun	Length	8
Frequency	289693	Text Frequency	6008
IDF	0.347914982876157	RIDF	0.347914982876155

Lemma
ecclesia

Declination

Singular	Masculine	Feminine	Neuter	Plural	Masculine	Feminine	Neuter
Nominative		ecclesia		Nominative		ecclesiae	
Vocative		ecclesia		Vocative		ecclesiae	
Genitive		ecclesiae		Genitive		ecclesiarum	
Dative		ecclesiae		Dative		ecclesiis	
Accusative		ecclesiam		Accusative		ecclesias	
Ablative		ecclesia		Ablative		ecclesiis	

Clear Results

Start | Resource ... | Lexicon B...

Why RDF?

- prospective longevity
- de facto standard
- freedom of modelling but fixed form (definition file) -> eLexicon
- easy conversion

Gleim, R., Hoenen, A., Mehler, A., Ernst, A., Diewald, N. Modeling, Building and Maintaining Lexica for Corpus Linguistic Studies. CL 2011. Birmingham.
http://linguistics-ontology.org/gold/2010
http://www.isocat.org/

THANK YOU!