

Interoperability of Second Language Resources and Tools

Elena Volodina¹, Maarten Janssen², Therese Lindström Tiedemann³,
Nives Mikelić Preradović⁴, Silje Ragnhildstveit⁵, Kari Tenfjord⁶,
Koenraad de Smedt⁶

¹ University of Gothenburg, **Sweden**; ² University of Coimbra, **Portugal**; ³ University of Helsinki, **Finland**;

⁴ Western Norway University of Applied Sciences, **Norway**; ⁶ University of Bergen, **Norway**



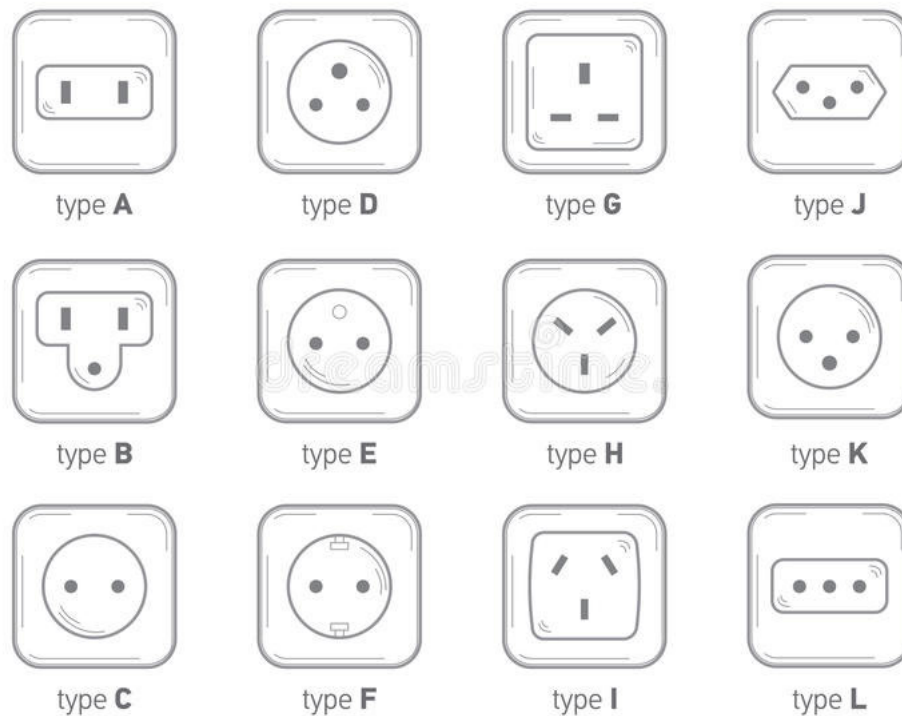
What is **Interoperability** (of Language Resources) ?

Ideal picture



What is **Interoperability** (of Language Resources) ?

Real-life picture



What is **Interoperability of Language Resources** ?

by Chiarcos, 2012

- Structural
 - Annotations of different origin are represented using the same formalism (e.g. stand-off XML or RDF databases)
- Conceptual
 - Annotations of different origin are linked to a common vocabulary (terminological reference repository)

Chiarcos, C. (2012). Interoperability of corpora and annotations. In *Linked Data in Linguistics*. Springer.

What is **Interoperability of Language Resources** ?

by Foulonneau & Riley, 2014

- **Metadata**

- Descriptions of the data; resource discovery in search engines, portals and registries. (+filtering?)

- **Technical**

- Data aggregation

- **Content**

- Comparable content of the resources – based on metadata

Foulonneau, M., & Riley, J. (2014). *Metadata for digital resources: implementation, systems design and interoperability.* Elsevier.

What is **Interoperability of Language Resources** ?

by Ide & Pustejovsky, 2010

...a measure of the degree to which diverse systems, organizations and/or individuals are able to work together to achieve a common goal.

- For computers
 - **syntactic** interoperability (data formats, communication protocols, data exchange)
 - **semantic** interoperability (ability to automatically interpret exchanged information via a common information exchange reference model)
- For language resources
 - focus is rather on **semantic** interoperability, since syntactic ones are technically mappable via a trivial conversion

Ide, N., & Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*.

What is **Interoperability of Language Resources** ?

by Ide & Pustejovsky, 2010

- Metadata
 - characteristics of data expressed through a set of labels (syntactic dimension) and categories (semantic dimension)
- Data categories and their semantics
 - e.g. morpho-syntax, syntax, text typologies, etc.
- Requirements for publication of data and notations
 - common practices for creating, documenting and evaluating language resources, e.g. agreement on formats and access; encoding; copyright; etc.
- Requirements for software sharing
 - software formats, data formats, software integration platforms; possibility to combine different tools; evaluation of software; copyright

Ide, N., & Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*.

Interoperability of Second Language Resources and Tools

Ide, N., & Pustejovsky, J. (2010).

- Metadata
- Error taxonomies
- Tools
- User interfaces

Workshop on Interoperability of Second Language Resources and Tools

6-8 Dec. 2017 | University of Gothenburg, Sweden

<https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools>

CONTRIBUTION TO CLARIN GOALS

- Promotion of the interoperability of resources and tools in CLARIN by working towards common guidelines for L2 annotation and metadata
- Promotion of the integration of data, tools and services making L2 corpora compatible with corpus tools in CLARIN
- Creating and expanding the CLARIN network of experts in the area of L2 corpora and tools

INVITED TALKS and main program points

Towards standardization of metadata for L2 corpora:
Sylviane Granger, University of Louvain, Belgium

Program in a nutshell:

- Existing corpora
- Corpora under construction
- L2 infrastructures
- Metadata and ethics
- Error annotation
- Tools and software
- Happy user
- Developments on top of L2 corpora
- Discussion *à la World café*
- Social program

Participants and organizers

Organizers:

- Elena Volodina, Sweden
- Kari Tenfjord & Silje Ragnhildstveit, Norway
- Therese Lindström Tiedemann, Finland
- Nives Mikelić Preradović, Croatia
- Maarten Janssen, Portugal

Participants:

15 countries; 27 participants



L2 metadata

by Granger and Paquot, 2018

- Administrative
 - title, license, availability, ...
- Corpus design
 - L1s, L2s, size, mode, levels, guidelines, ...
- Annotation
 - type: POS, syntax, errors; tagsets, guidelines, tools, ...
- Text
 - mode, author, title, statistics, task types & instructions, ...
- Learner
 - age, gender, L1s, L2s, level, school, education, ...

L2 metadata

in present-day LCR projects

- Varied between L2 corpora
 - no track of various aspects, e.g. Tasks, guidelines, etc
- Restricted by laws and agreements
 - e.g. aggregated birth year spans in one corpus *versus* exact birth year in another
- Incompatible
 - e.g. Bosnian, Serbian and Croatian L2s separate in one corpus *versus* BSC in another

Granger, S., & Paquot, M. (2018). *Towards standardization of metadata for L2 corpora*. Presentation at the workshop on Interoperability of Second Language Resources and Tools. Gotheburg, Sweden, Dec 2017.

Examples: *Korp, Swedish edition*

KORP SW1203-uppsatser selected — 51.97K of 13.26G tokens

Simple Extended Advanced Compare

Filter: Type: C: Slutprovsuppsats and Proficiency Level: C1 B2 and Add native language

gender is Kvinna

or

and lemgram is barn (noun)

or

Search within sentence

KWIC: hits per page: 25 sort within corpora: not sorted Statistics

KWIC Statistics Word picture

Results: 6

« < 1 > » Go to page of 1 Show context

De måste aktivt använda barn att...
Många föräldrar tror att tvåspråkiga barn har fördelar.
Orsaken till detta är att många föräldrar tror att barn som talar flera språk har bättre kognitiva förmågor.
Särskilt i Afrika finns det många språk som är hotade eftersom folkgrupper lever i områden där språket inte används längre.
Sverige till exempel har sluttande språk.

Arabic 354
Bengali 424
Bosnian 501
Catalan 478
German 1105
Greek 358
English 1503
Finnish 799
French 841
Serbo-Croatian 444
Hungarian 470
Japanese 428
Latvian 410
Lithuanian 390
Macedonian 422
Dutch 425
Persian (Dari) 708
Russian 822
Spanish 1731
Chinese 1906
Estonian 9

KORP TISUS-texter selected — 59.64K of 13.26G tokens

Simple Extended Advanced Compare

written proficiency is 5

written proficiency is 4

or

and gender is Kvinna

or

✓ Afrikaans
Azerbaijani
Bulgarian
Croatian
Czech
Dutch
English
Estonian
Finnish
French
Galician
German
Greek
Hungarian
Italian
Japanese
Korean
Latvian
Lithuanian
Mandarin Chinese
Polish
Portuguese
Russian
Serbian
Serbo-Croatian
Spanish
Swedish

KWIC: hits per page: 25 sort within corpora: not sorted Statistics

KWIC Statistics Word picture

Results: 6

« < 1 > » Go to page of 1 Show context

De måste aktivt använda barn att...
Många föräldrar tror att tvåspråkiga barn har fördelar.
Orsaken till detta är att många föräldrar tror att barn som talar flera språk har bättre kognitiva förmågor.
Särskilt i Afrika finns det många språk som är hotade eftersom folkgrupper lever i områden där språket inte används längre.
Sverige till exempel har sluttande språk.

Error taxonomy

an ideal

- Same error classification approach across L2 corpora
 - e.g. based on linguistic description (phonology, orthography, morphology, ...) (Dobrić 2015)
- Same granularity
 - 22 tags versus 65 tags
- Theory-independent approach
 - (Tenfjord et al 2006)
- Piloting
 - test on project members first to avoid unreliable / confusing tags
- Annotation
 - e.g. normalization first, error code afterwards (Volodina et al. 2018)
- Annotation quality
 - documented inter-annotator agreement, etc. (Fort 2016)

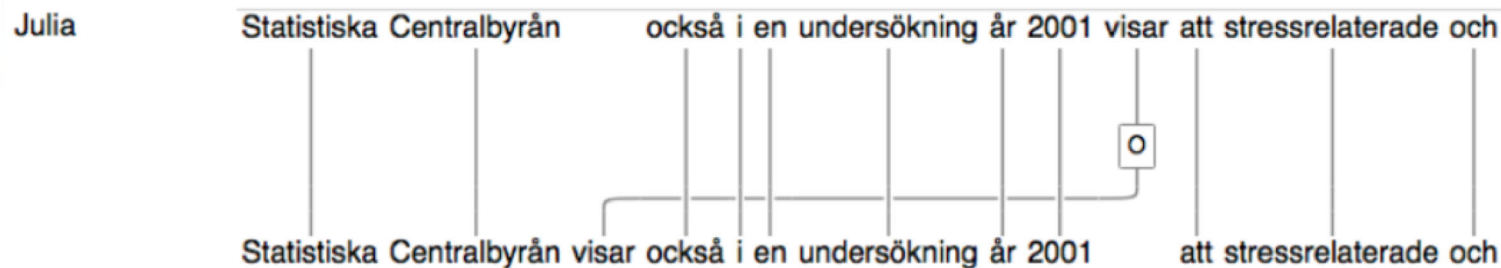
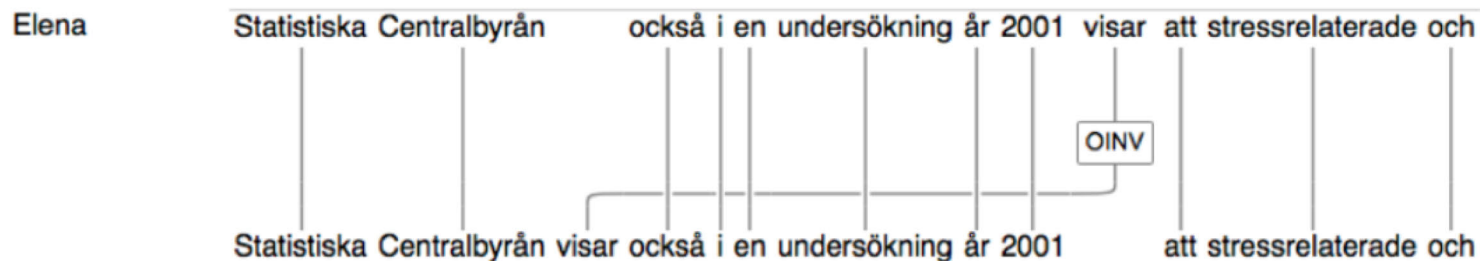
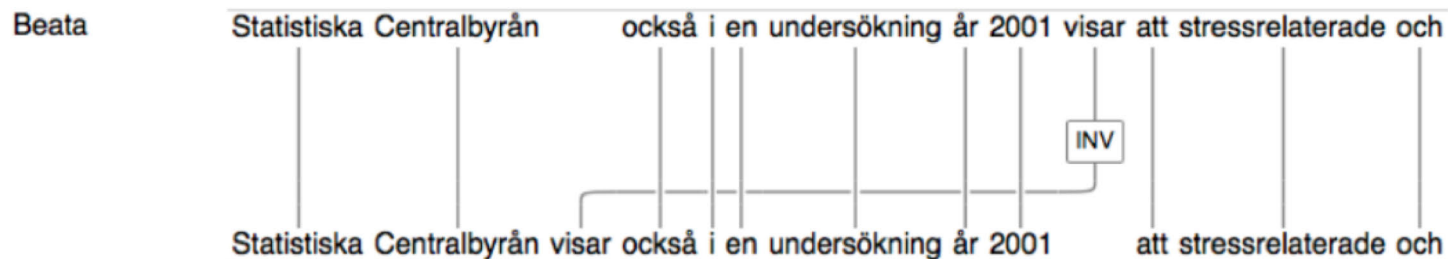
Error taxonomy

in present-day LCR projects

Taxonomies are like underwear; everyone needs them, but no one wants someone else's.

(From a presentation by Egon Stemle at CLARIN workshop on interoperability of L2 resources and tools)

Example: ASK taxonomy in SweLL pilot



Gloss: Central Statistical Agency [...] also in a report from 2001 [*shows (finite verb)*] that stress-related and...

Error code explanations: *INV*: Non-application of subject/verb inversion, *OINV*: Application of subject/verb inversion in inappropriate contexts, *O*: other word (or phrase) order error.

Tools

an ideal

- Accessible
- User-friendly
- Well-documented
- Accompanied by user manuals
- Collected in one repository for re-use
- Annotation quality

Tools

in present-day LCR projects

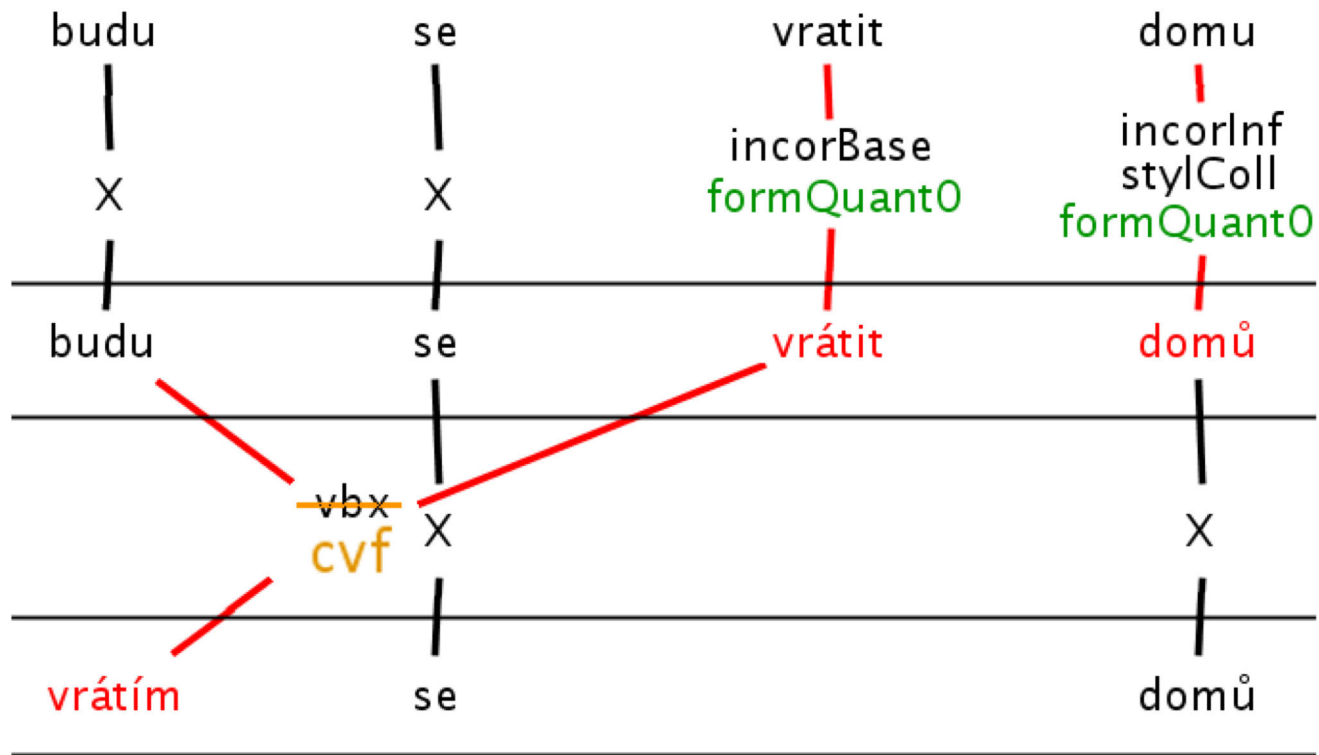
Pluralism of formats and outputs, often inaccessible, or proprietary.

Some examples:

- Feat (Hana et al 2010)
- TEITOK (Janssen 2016)
- SVALA (Rosén et al 2018)
- Falko-tools (Müller & Strube 2006)

Some tools

feat



SVALA

revert

auto

disconnect

merge

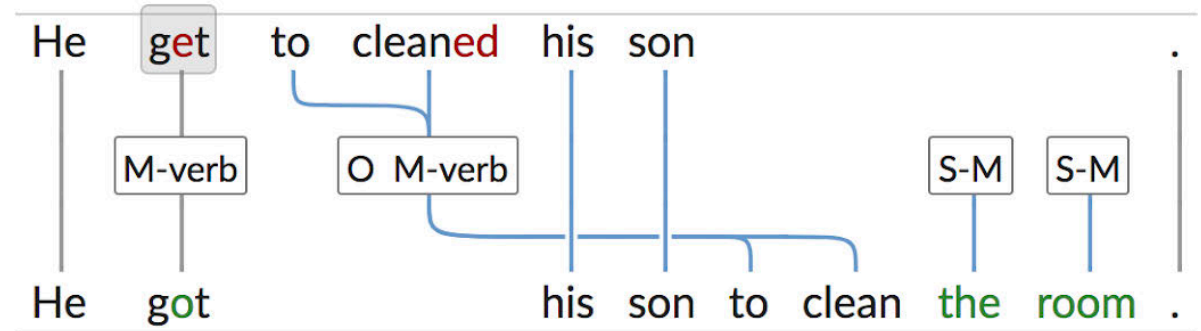
group

deselect

Enter label...

• M-verb

×



User interfaces

an ideal

- Basic and advanced modes
- Selection of error types to be correlated with speaker features
- Metadata re-use & filtering

User interfaces

in present-day LCR projects

- Various **formats** exist, but most of the search tools rely on xml format
→ need to have a TEI-conformant version of all corpora
- **Search builder** – varies between interfaces
 - Not all **metadata** is visualized or is made searchable

Some user interfaces

Swedish Korp

Search query & filters

KORP SW1203-uppsatser selected — 51,97K of 13,26G tokens

Single Extended Advanced Compare

Filter: Add type and Add proficiency level and Add native language

A: Inträdesuppsats (1761)
B: Mitterminsuppsats (1801)
C: Slutprovsuppsats (2019)
D: Ömprow (1762)

is

or

and

+

Add token

Generellt för alla uppsatser i SW1203-korpusen

Studenterna skrev i tentasalen.
Inga hjälpmedel var tillåtna.
Tid 3,5 timmar.

SKRIVUPPGIFT: Ett brev till politikerna i kommunen

Du har fått två kladdpapper och ett dubbelvikt ark. Skriv din text först på kladdpapper och sedan sedan resten på det dubbelvikta arket. Kom ihåg att chatta namn och

Task description

Hits in KWIC format

Go to page 1 of 143 Show context

SW1203-uppsatser

Jag gillar mycket min omvärld där jag bor. min omvärld där jag bor nu.

Jag har väldigt bra tillgång till kollektivtrafik, till exempel bus. bra tillgång till kollektivtrafik, till exempel bus.

Det finns många busar och spårvagnar på vilka man kan åka. till alla platser på stan.

En viktig del av det är biblioteket som jag tycker om. om.

Det finns också mycket tid där. mycket tid där.

ett tid där. där.

över också sina läxor. också sina läxor.

Jag är inte nöjd med avtal system. nöjd med avtal system.

hittas lätt i soprum i varje byggnad och de måste vara i stor. i soprum i varje byggnad och de måste vara i stor.

antal. Dessutom måste de delas ut ordentligt. Dessutom måste de delas ut ordentligt.

les ut ordentligt. bra idé att vi får ge en förslag för att förbättra vår k.

är en mycket hit för 4 månader sedan. hit för 4 månader sedan.

ttade sedan. sedan.

ett bo här därför att det finns flera mataffärer i närområdet c. därför att det finns flera mataffärer i närområdet c.

so här därför att det finns flera mataffärer i närområdet och det. därför att det finns flera mataffärer i närområdet och det.

h det inte tar lång tid att gå till hållplatsen. tar lång tid att gå till hållplatsen.

Dessutom finns det en stor skog omkring i vilken man kan pr. finns det en stor skog omkring i vilken man kan pr.

skog omkring i vilken man kan promenera. omkring i vilken man kan promenera.

in jag inte är så nöjd med. är så nöjd med.

nte är så nöjd med. så nöjd med.

n kan inte sortera sopor i vår sopstation så att vi bara kan så. sortera sopor i vår sopstation så att vi bara kan så.

ation så att vi bara kan sätta alla sopor i en påse och sedan. att vi bara kan sätta alla sopor i en påse och sedan.

att vi bara kan sätta alla sopor i en påse och sedan kastar de. bara kan sätta alla sopor i en påse och sedan kastar de.

of 143

Download hit page as...

Metadata

Corpus

SW1203-uppsatser

Text attributes

student: I

type: A: Inträdesuppsats

task URL:

http://sprakbanken.gu.se/_TillPolitiker.pdf

C/D: +

A: +

B: +

age: 27

gender: F

native language:

Persian (Dari)

proficiency level: B1

task: Ett brev till politikerna i kommunen

semester: HT12

birth year: 1985

Metadata explanation

Metadata beskrivning för SW1203

Födelseår 1963-1993

Kön Kvinna, Man

A, B, C (Type) Uppsatstyper:

A: Inträdesuppsats (52 st)

B: Mitterminsuppsats (41 st)

C: Slutprovsuppsats (45 st)

D: Ömprow (3 st)

Student

Semester

HT12 VT13

Student-id i korpusen (1-84)

HT12 VT13

Statistics

KWIC Statistics Word picture Graph

Show Trend Diagram Show map

Number of rows: 750

	proficiency level	Total
-program		
aldrig (adverb)	B0	37 (1)
allra (adverb)	C1	18,2 (1)
allra (adverb)	B0	51,7 (1)
alla (adverb)	C1	77 (1)
alla (adverb)	B0	18,2 (1)
allt (adverb)	C2	51,7 (1)
allt (adverb)	C1	77 (1)
allt (adverb)	B0	171,2 (1)
allt (adverb)	B0	96,2 (1)
alltid (adverb)	C2	51,7 (1)
alltid (adverb)	C1	134,7 (1)
alltid (adverb)	B0	363,4 (1)
alltid (adverb)	B0	230,9 (1)
alltid (adverb)	C2	51,7 (1)
alltid (adverb)	C1	115,4 (1)
alltid (adverb)	B0	151,9 (1)
alltid (adverb)	B0	77 (1)
alltid (adverb)	C1	18,2 (1)

Some user interfaces

Corpuscle :: ASK Hovedkorpus :: Concordance

Advanced search | [switch to Basic search](#)

[Query history ...](#)

[type = "R|M"] [pos = "det"] \ <> \\ :: language != "norsk"

[Run Query](#)

[Stop](#)

[Saved queries ...](#)

[Save Query](#)

Done. Running time: 2.23 sec. (2.34 CPU sec.)

Hit 1 - 30 of 3611 | [Previous](#) | [Next](#) | Go to: | [Download](#) (☐ Excel mode) | Type: [context](#) | ☐ Show ann

count	cpos	context
1	853	Nettopp denne uken organiserte skolen til barna mine et foreldremøte hvor vi diskuterte undervisningssystem kommet fram med.
3	1267	På { den } _M andre siden {hver {enkelte enkelt } _F av oss forberedet forberedet hver enkelt av oss } _O seg sammen samlet seg for å vente med glede { på } _M det nye året.
5	2450	Men når { den } _R mannen blir så trygg på familien sin at han tror at han {trenger ikke ikke trenger } _O å vis
6	3393	Men likevel er det vanskelig for meg å {adaptere tilpasse } _W meg { til } _R et fremmed samfunn.
8	4871	Det var en av { de } _R grunnene { til } _M at hun ble mobbet hver eneste dag.
9	6470	Ei bok for barn i småskulen burde ikke {være i første omgang i første omgang være } _O skrevet med { de }
10	9704	Gjennom historien da samfunnreglene { kunne } _M {var være } _F ustabile og uklare { , } _{PUNCM} var familien organisere seg { på } _M i { { en et } _F slags } _R samfunnslivet.
12	9776	I gamle dager var { det } _M de eldste som bestemte alt.
14	11287	Bedriften kan {selfølgelig selvfølgelig } _{ORT} påvirke { i } _R en person som liker å {trakasere trakassere } _{ORT} } _{PUNCM} eller noe liknende.
16	11565	Nå har jeg bare en familie og en kjæreste - fantastisk, forresten - men snart skal jeg ha to familier (eller tre i
18	13970	Bedrifter utnytter netthandelsmetoden { { en et } _F } _R {økonomisk økonomisk } _{ORT} { { viss vis } _{ORT} } _R .
19	14321	Dette er ikke bare { et } _R læreres {probleme problem } _{ORT} , det er { et } _R {samfunns samfunnets } _F {pro
21	14563	Læreren blir trøtt og sint fortere enn { i } _M en klasse med færre elever.
23	15597	Gi {lærere lærerne } _F høyere lønn { , } _{PUNCM} slik at barna {våres våre } _{INFL} er {sikre d te sikret } _F til { av
25	16487	Det som kan være negativt {til med } _W å ha et mer åpent forhold mellom {kongefamiliet kongefamilien } _{IN} skulle miste { den } _R betydningen og {rolle rollen } _F i det norske samfunnet.
26	16617	Jeg ser på det norske monarkiet som { på } _R et {gammel gammelt } _F og fredelig {kongeriket kongerike }
28	17115	Jeg vet ikke, fordi det bestemmer { de } _R nordmenn {nordmenn nordmennene } _F selv.

Choose Corpus:
Hovedkorpus

Query

Concordance
Collocations
Distribution
Word List
Text
Overview
Variables

Basic search | **Advanced search**

Use this input form to write a textual query.

Query: [pos = "prep"] \ <> \\ \del :: language = "polsk|tsk"

Run Query | **Reset query** | Build graphical query

Here you can compose a query graphically.

Choose a **subcorpus**:

Morsmål = "polsk|tsk"
add:
attribute: -

Choose **positional constraints**. | ☒ Ignore structural positions

☐ target
Ordklasse = ""
repetition: 1

add:
attribute: -
struct: -

konj
prep
sbu
subst
sybm
ukjent
verb

Run Query

OK Cancel

[illegible]

Choose Corpus:
Hovedkorpus

Basic search | **Advanced search**

Query: [pos = "prep"] <> \\\ :: language = "polsk|tysk"

Query

Concordance

Collocations

Distribution

Word List

Text

Overview

Variables

Match size: 18907, unique words or phrases: 202. Attribute: Ord

Page 1 of 1.

3204 (16,95%) i	32 (0,17%) hjem	6 (0,03%) langs	2 (0,01%) vek
2164 (11,45%) p	30 (0,16%) fram	5 (0,03%) pga	1 (0,01%) Angående
2113 (11,18%) til	30 (0,16%) per	5 (0,03%) unna	1 (0,01%) Bland
2049 (10,84%) for	29 (0,15%) Med	4 (0,02%) angående	1 (0,01%) Etter
1614 (8,54%) med	28 (0,15%) unsett	4 (0,02%) ifra	1 (0,01%) Fram
1599 (8,46%) av	25 (0,13%) imot	4 (0,02%) nedover	1 (0,01%) Får
776 (4,10%) om	24 (0,13%) tross	3 (0,02%) Bland	1 (0,01%) Heretter
536 (2,83%) som	23 (0,12%) Av	3 (0,02%) Hjemsne	1 (0,01%) Hos
467 (2,47%) I	23 (0,12%) Om	3 (0,02%) Pa	1 (0,01%) Iblant
458 (2,42%) fra	23 (0,12%) Som	3 (0,02%) Tross	1 (0,01%) Intil

Choose Corpus:
Hovedkorpus

Display: original

Query
Concordance
Collocations
Distribution
Word List
Text
Overview
Variables

Oppgave C |

Norge er en av landene som har ledige arbeidsplasser. De finnes ikke nok nordmenn for alle stillingene særlig når det gjelder helsepersonell og ingeniører. Sykepleierne og legene gjør en veldig viktig, ansvarlig og tung jobb, men betaling for denne jobben er ikke så høy. Det kan være grunnen til at mennesker vil ikke jobbe i sånne yrker.

Norge er nødt til å hente arbeidskraft fra utlandet. Landet vil ikke funksjonere bra hvis det blir stort arbeidsledighet. Utvikling av landet blir hindret. Det er ikke positivt for Norges økonomi. I dag er det viktig å konkurrere med andre land, men det blir ikke mulig hvis Norge blir rammet av arbeidsledighet.

Jeg tror at ~~det tiltak~~ er også mulighet for mennesker som kan komme til Norge for å jobbe . Kanskje de har problemer med å finne en jobb i sitt eget land. Så det er et fantastisk sjanse for ~~å skaffe~~ seg et bedre liv. Noen ganger går det på bekostning av familie og venner som de må forlate.

Det at **N** Norge kommer folk fra mange forskjellige land er veldig bra for norske folk. De kan være kjent med andre kulturer, andre religioner er . Kanskje de kan finne seg venner.

Integrering er meget viktig for utlendinger. Hvis en skal trives her i Norge må han tilpasse seg det norske samfunnet. Myndighetene og arbeidsgiverne kan gjøre mye på dette området. Noen ganger trenges det å skifte loven , eller å

Some user interfaces

TEITOK

Editable metadata

KWIC/XML

Search & Filters

Query Builder

Text Search

Written form matches

Teacher corrected form matches

Orthographically corrected form matches

Syntactically corrected form matches

Lexically corrected form matches

POS tag matches

Lemma matches

Document Search

Nationality [select]

Mother tongue [select]

Proficiency [select]

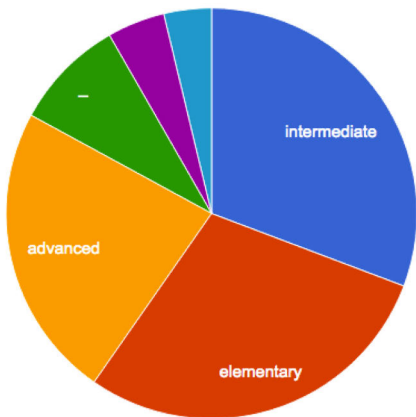
Collection phase [select]

Search query:

Group query: Group by: **Proficiency**

Statistics

Graph: Count: Download:



Entry exam

Template: telHeader-edit.tpl

Text title	Entry exam
Responsibility Statement	Andrej Žečević-Padić, Luka Dominiković
Project title	
Project language	
Person ID	afr00213
Year of Birth	1996
Sex	
Level of language proficiency for the language being learned	2A
First language	Afrikaans
Foreign languages	
Mother first language	
Father first language	
Stay in Croatia	
Duration of language learning	
Continuing from lower language learning level	
Task ID	2A_00
Task Title	Entry exam
Genre of the task	essay
Extent of the task	10-15 sentences
Type of task	test

vezes não
vezes não
quando p
coisa
demais e
online
à Polónia,
à Polónia,
Torre de B
base da s
duas da ta
Torre de B
solidão, o
maior part
que a qual
M - Ade
Aim

Cascais de 05 de Julho de 2010

Caro Nuno,

Tenho sorte de vir para cá, mesmo a vida é muito diferente da nossa, mas gostei de muitas coisas, sobretudo da simp^{at}ia do povo português. No s^{ab}ado conheci a FF e o seu marido, o MM depois de 6 anos de relação virtual.

Levaram-me às praias de Cascais, é muito perto de Lisboa, a praia não é muito grande, a sua areia é dourada, também a água é muito sana ?? para banhar-se.

A zona é muitas cheia de turistas, ao andar podes ouvir muitas línguas estrangeiras, grupos de jovens, idosos, casais, surfistas.

Ao n^{ível} de vigilância e higiene, há muitas tropas equipas de nadadores-salvadores, assegurando a vida daos pessoas banhistas. A vida anda muito bem sem conflitos, há uma organização muito cuidado. Com certeza enviarei-te enviar-te-ei mais cartas e mais notícias. Aproveita o teu tempo também. Na espera deas tuas novidades espero que tenhas um mês tranquilo e cheio de felicidade.

Beijinhos!

Document

GRUPO IV

A Joana está a passar férias numa praia e quer dar notícias ao seu melhor amigo, que se chama Nuno. Para isso, ela escreve-lhe uma carta, na qual:

- descreve a praia;
- conta como soube os seus dias;
- dá outras informações interessantes.

Escreve a carta de Joana, num texto com um mínimo de 60 e um máximo de 100 palavras.

Handwritten text in Portuguese, likely a student response to the writing task.

Some user interfaces

ANNIS

learner	von	20	in	Stadt	X	exestierte	Baugenossenschaften
TH1	von	20	in	Stadt	X	existierten	Baugenossenschaften
TH1Diff						CHA	
TH2	von	20	in	Stadt	X	existenten	Baugenossenschaften
TH2Diff						CHA	
EA_category						G_Morphol_Wrong	
EA_category						O_Graph	
EA_category						V_semdenot_word_fs	

Prospects

- Non-trivial work, time-consuming, community depending
 - a lá Universal Tagset / Universal Dependencies
- Pluralism (in tools and formats) is healthy, BUT we need a common conversion mechanism, a lá transformers, between the pluralistic approaches
- Need for insights from several perspectives
 - Second Language Acquisition researchers, (learner corpora) linguists, teachers, language testing specialists
 - NLP researchers, software engineers, Systems developers

Follow-up

- **Suggestions for future:** <https://goo.gl/bW24Sq>
- **Joint publication** in LCR 2018 post-conference volume (accepted, publication date 2019):
 - Egon W. Stemle (Italy), Adriane Boyd (Germany), Maarten Janssen (Portugal), Therese Lindström Tiedemann (Finland), Nives Mikelić Preradović (Croatia), Alexandr Rosen (Czech Republic), Dan Rosén (Sweden), Elena Volodina (Sweden). *Working together towards an ideal infrastructure for language learner corpora.*
- **CLARIN survey of L2 corpora:**
 - *L2 learner corpus survey – Towards improved verifiability, reproducibility and inspiration in learner corpus research.* By Therese Lindström Tiedemann, Jakob Lenardič and Darja Fišer. CLARIN 2018
- **COST action:** application is planned
- **Follow-up workshop:** is planned



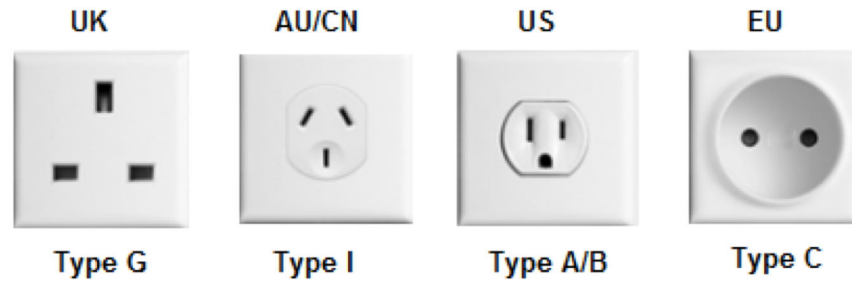
References 1

- **Ballier, N., & Martin, P.** (2013). Developing corpus interoperability for phonetic investigation of learner corpora. *Automatic treatment and analysis of learner corpus data*, 33-64.
- **Chiarcos, C.** (2012). Interoperability of corpora and annotations. In *Linked Data in Linguistics* (pp. 161-179). Springer, Berlin, Heidelberg.
- **Dobrić, N.** (2015). Quality Measurements of Error Annotation-Ensuring Validity Through Reliability. *The European Messenger*, Vol. 24.1, pp 36-42
- **Fang, A. C.** (2012). Creating an interoperable language resource for interoperable linguistic studies. *Language resources and evaluation*, 46(2), 327-340.
- **Fort, K. (2016).** *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.
- **Foulonneau, M., & Riley, J.** (2014). *Metadata for digital resources: implementation, systems design and interoperability*. Elsevier.
- **Hana, J., A. Rosen, S. Škodová and B. Štindlová** (2010). Error-tagged Learner Corpus of Czech. In Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV). Uppsala, Sweden.
- **Ide, N., & Pustejovsky, J.** (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*.
- **Janssen, M.** 2016. [TEITOK: Text-Faithful Annotated Corpora](#). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.
- **Müller, C. and Strube, M.** (2006). Multi-level annotation of linguistic data with MMAX2. *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, 3, pp. 197–214. < <https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiterinnen/marc/> >

References 2

- **Rosén, D., M. Wirén, and E. Volodina** (2018). Error Annotation of Second Language Learner Texts Based on Mostly Automatic Alignment of Parallel Corpora. CLARIN annual conference 2018.
- **Tenfjord, K., Johansen, H., & Hagen, J. E.** (2006). The "Hows" and the "Whys" of Coding Categories in a Learner Corpus (or "How and Why an Error-Tagged Learner Corpus is not "One Big Comparative Fallacy"). *Rivista di psicolinguistica applicata*, 6(3).
- **Volodina, E., L. Granstedt, S. Johansson, B. Megyesi, J. Prentice, D. Rosén, C-J. Schenström, G. Sundberg and M. Wirén.** (2018). Annotation of learner corpora: first SweLL insights. *Proceedings of SLTC 2018*.

Thank you!

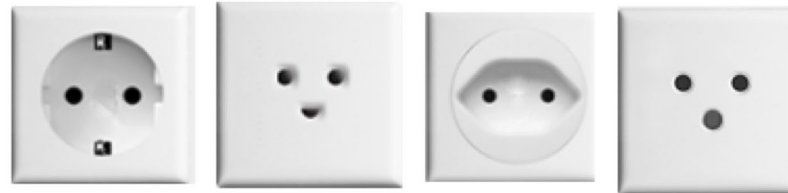


Germany France
Austria, Korea
Netherland Spain
Russia Belgium

Denmark
(happy face)

Switzerland
Italy Brazil
South Africa

Israel



Type C Europlug fits all, more than 130 countries

Our aim

