CLARIN & Libraries

Martin Wynne

martin.wynne@ling-phil.ox.ac.uk

Faculty of Linguistics, Philology and Phonetics,
University of Oxford

National Coordinator, CLARIN-UK

CLARIN and Libraries Workshop

KB National Library of the Netherlands

Monday 9th May 2018

The motivation for the workshop

- A number of major initiatives were emerging which aimed to delivering digital text collections, for readers and also for research use
- but projects were often somewhat disconnected from each other
- and also disconnected from research infrastructures
- So, it was a good time to discuss sharing experiences, methods, tools, and plans, and looking for some potential areas of collaboration

CLARIN and Libraries,
also known as:
Interoperability of Text Platforms for Digital Libraries

CLARIN: data, tools, and methods

- *Data*: a corpus containing multiple texts representative of a particular language variety
- *Tools*: Concordance, collocation, clusters, keywords; working with annotation and metadata
- *Methods*: Data-driven, evidence based research into language; distant reading

Digital Libraries: data, tools, and methods

- *Data*: digitized copies of books and other text collections
- *Tools*: Library records search, text display for online reading and download
- *Methods*: Accessing the content of books; close reading

CLARIN: data, tools, and methods

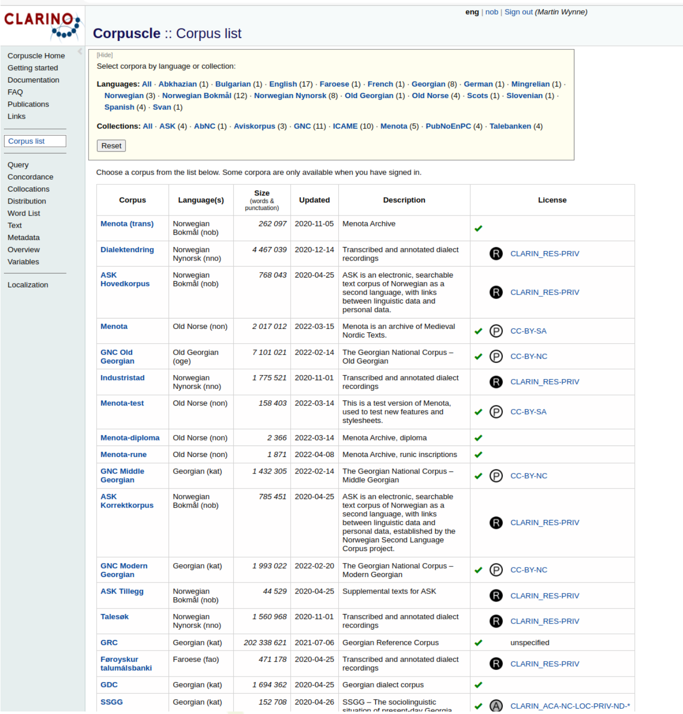
- *Data*: a corpus containing multiple texts representative of a particular language variety
- *Tools*: Concordance, collocation, clusters, keywords; working with annotation and metadata
- *Methods*: Data-driven, evidence based research into language; distant reading

Language Resource Repositories

- *Data*: corpora **and** digital text collections
- *Tools*: Sometimes with tools that can be connected to the data
- *Methods*: Facilitating corpus linguistics methods

Digital Libraries: data, tools, and methods

- *Data*: digitized copies of books and other text collections
- *Tools*: Library records search, text display for online reading and download
- *Methods*: Accessing the content of books; close reading



Corpuscle :: Corpus list

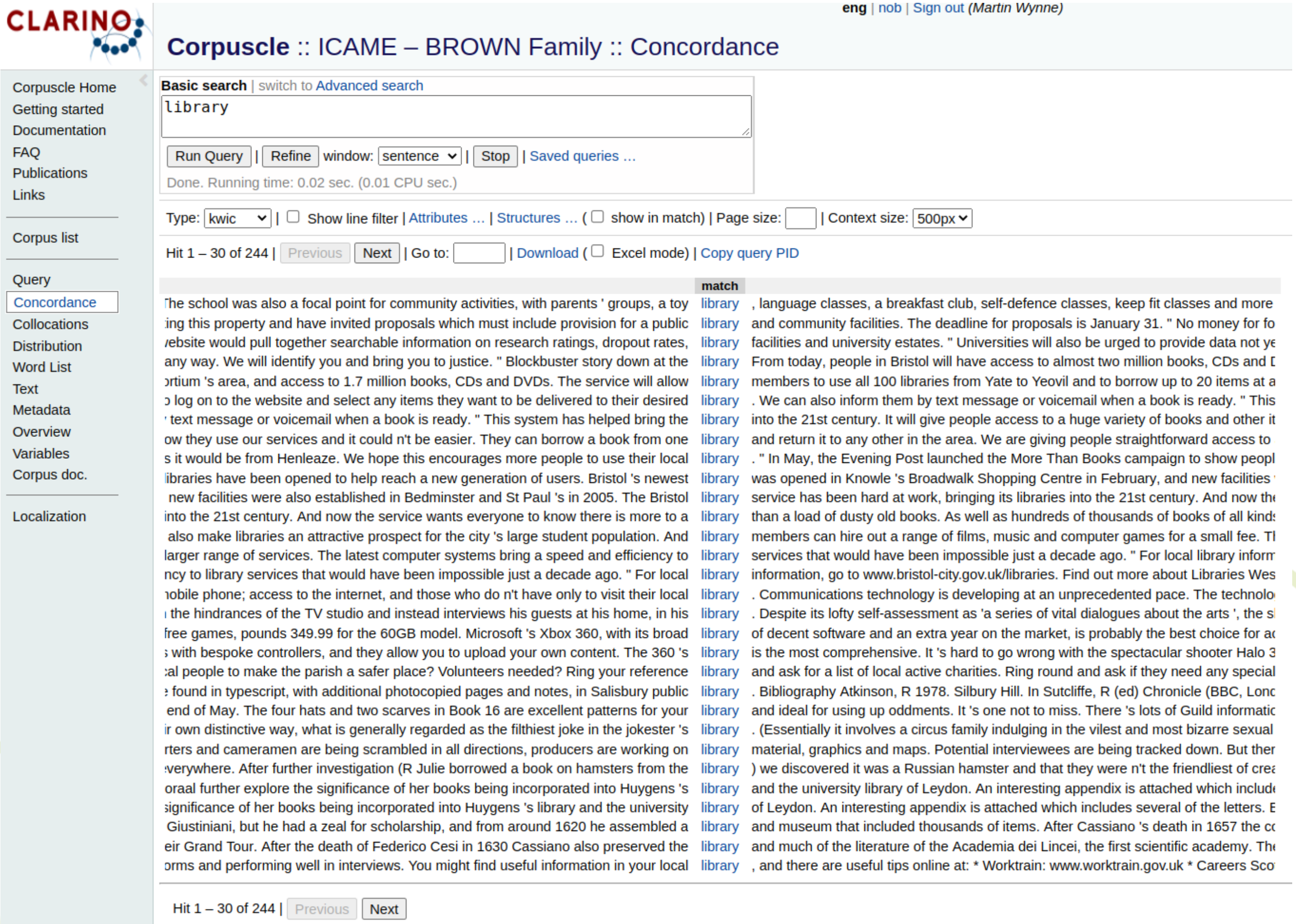
Select corpora by language or collection:

Languages: All · Abkhazian (1) · Bulgarian (1) · English (17) · Faroese (1) · French (1) · Georgian (8) · German (1) · Mingrelian (1) · Norwegian (3) · Norwegian Bokmål (12) · Norwegian Nynorsk (8) · Old Georgian (1) · Old Norse (4) · Scots (1) · Slovenian (1) · Spanish (4) · Svan (1)

Collections: All · ASK (4) · AbNC (1) · Aviskorpus (3) · GNC (11) · ICAME (10) · Menota (5) · PubNoEnPC (4) · Talebanken (4)

Choose a corpus from the list below. Some corpora are only available when you have signed in.

Corpus	Language(s)	Size (words & punctuation)	Updated	Description	License
Menota (trans)	Norwegian Bokmål (nob)	*262 097*	2020-11-05	Menota Archive	✓
Dialektendring	Norwegian Nynorsk (nno)	*4 467 039*	2020-12-14	Transcribed and annotated dialect recordings	Ⓡ CLARIN_RES-PRIV
ASK Hovedkorpus	Norwegian Bokmål (nob)	*768 043*	2020-04-25	ASK is an electronic, searchable text corpus of Norwegian as a second language, with links between linguistic data and personal data.	Ⓡ CLARIN_RES-PRIV
Menota	Old Norse (non)	*2 017 012*	2022-03-15	Menota is an archive of Medieval Nordic Texts.	✓ Ⓟ CC-BY-SA
GNC Old Georgian	Old Georgian (oge)	*7 101 021*	2022-02-14	The Georgian National Corpus – Old Georgian	✓ Ⓟ CC-BY-NC
Industristad	Norwegian Nynorsk (nno)	*1 775 521*	2020-11-01	Transcribed and annotated dialect recordings	Ⓡ CLARIN_RES-PRIV
Menota-test	Old Norse (non)	*158 403*	2022-03-14	This is a test version of Menota, used to test new features and stylesheets.	✓ Ⓟ CC-BY-SA
Menota-diploma	Old Norse (non)	*2 366*	2022-03-14	Menota Archive, diploma	✓
Menota-rune	Old Norse (non)	*1 871*	2022-04-08	Menota Archive, runic inscriptions	✓
GNC Middle Georgian	Georgian (kat)	*1 432 305*	2022-02-14	The Georgian National Corpus – Middle Georgian	✓ Ⓟ CC-BY-NC
ASK Korrektkorpus	Norwegian Bokmål (nob)	*785 451*	2020-04-25	ASK is an electronic, searchable text corpus of Norwegian as a second language, with links between linguistic data and personal data, established by the Norwegian Second Language Corpus project.	Ⓡ CLARIN_RES-PRIV
GNC Modern Georgian	Georgian (kat)	*1 993 022*	2022-02-20	The Georgian National Corpus – Modern Georgian	✓ Ⓟ CC-BY-NC
ASK Tillegg	Norwegian Bokmål (nob)	*44 529*	2020-04-25	Supplemental texts for ASK	Ⓡ CLARIN_RES-PRIV
Talesøk	Norwegian Nynorsk (nno)	*1 560 968*	2020-11-01	Transcribed and annotated dialect recordings	Ⓡ CLARIN_RES-PRIV
GRC	Georgian (kat)	*202 338 621*	2021-07-06	Georgian Reference Corpus	✓ unspecified
Føroyskur talumálsbanki	Faroese (fao)	*471 178*	2020-04-25	Transcribed and annotated dialect recordings	Ⓡ CLARIN_RES-PRIV
GDC	Georgian (kat)	*1 694 362*	2020-04-25	Georgian dialect corpus	✓
SSGG	Georgian (kat)	*152 708*	2020-04-26	SSGG – The sociolinguistic situation of present-day Georgia	✓ Ⓐ CLARIN_ACA-NC-LOC-PRIV-ND-*



Corpuscle :: ICAME – BROWN Family :: Concordance

Basic search | switch to Advanced search

```
library
```

Run Query | Refine window: sentence | Stop | Saved queries …

Done. Running time: 0.02 sec. (0.01 CPU sec.)

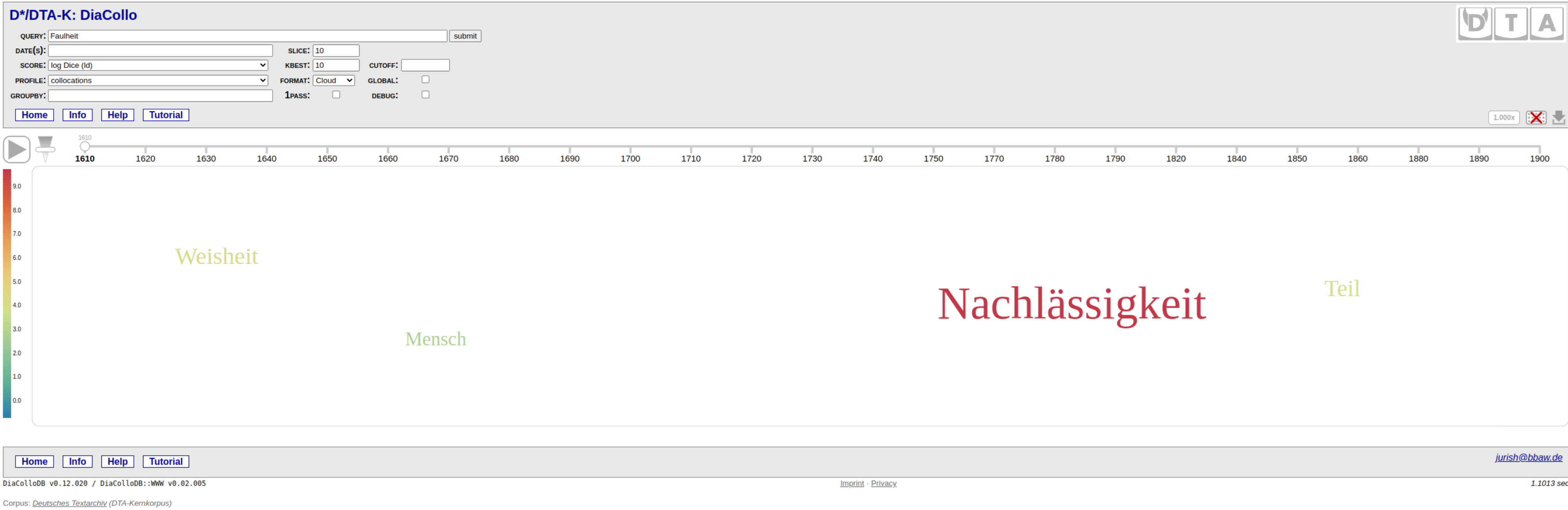
Type: kwic | □ Show line filter | Attributes … | Structures … (□ show in match) | Page size: | Context size: 500px

Hit 1 – 30 of 244 | Previous Next | Go to: | Download (□ Excel mode) | Copy query PID

	match	
The school was also a focal point for community activities, with parents ' groups, a toy	library	, language classes, a breakfast club, self-defence classes, keep fit classes and more
ing this property and have invited proposals which must include provision for a public	library	and community facilities. The deadline for proposals is January 31. " No money for fo
ebsite would pull together searchable information on research ratings, dropout rates,	library	facilities and university estates. " Universities will also be urged to provide data not ye
any way. We will identify you and bring you to justice. " Blockbuster story down at the	library	From today, people in Bristol will have access to almost two million books, CDs and
rtium 's area, and access to 1.7 million books, CDs and DVDs. The service will allow	library	members to use all 100 libraries from Yate to Yeovil and to borrow up to 20 items at a
log on to the website and select any items they want to be delivered to their desired	library	. We can also inform them by text message or voicemail when a book is ready. " This
text message or voicemail when a book is ready. " This system has helped bring the	library	into the 21st century. It will give people access to a huge variety of books and other it
ow they use our services and it could n't be easier. They can borrow a book from one	library	and return it to any other in the area. We are giving people straightforward access to
s it would be from Henleaze. We hope this encourages more people to use their local	library	. " In May, the Evening Post launched the More Than Books campaign to show peopl
libraries have been opened to help reach a new generation of users. Bristol 's newest	library	was opened in Knowle 's Broadwalk Shopping Centre in February, and new facilities
new facilities were also established in Bedminster and St Paul 's in 2005. The Bristol	library	service has been hard at work, bringing its libraries into the 21st century. And now the
into the 21st century. And now the service wants everyone to know there is more to a	library	than a load of dusty old books. As well as hundreds of thousands of books of all kind
also make libraries an attractive prospect for the city 's large student population. And	library	members can hire out a range of films, music and computer games for a small fee. T
larger range of services. The latest computer systems bring a speed and efficiency to	library	services that would have been impossible just a decade ago. " For local library inform
ncy to library services that would have been impossible just a decade ago. " For local	library	information, go to www.bristol-city.gov.uk/libraries. Find out more about Libraries Wes
obile phone; access to the internet, and those who do n't have only to visit their local	library	. Communications technology is developing at an unprecedented pace. The technolo
the hindrances of the TV studio and instead interviews his guests at his home, in his	library	. Despite its lofty self-assessment as 'a series of vital dialogues about the arts ', the s
free games, pounds 349.99 for the 60GB model. Microsoft 's Xbox 360, with its broad	library	of decent software and an extra year on the market, is probably the best choice for a
with bespoke controllers, and they allow you to upload your own content. The 360 's	library	is the most comprehensive. It 's hard to go wrong with the spectacular shooter Halo 3
al people to make the parish a safer place? Volunteers needed? Ring your reference	library	and ask for a list of local active charities. Ring round and ask if they need any special
found in typescript, with additional photocopied pages and notes, in Salisbury public	library	. Bibliography Atkinson, R 1978. Silbury Hill. In Sutcliffe, R (ed) Chronicle (BBC, Lon
end of May. The four hats and two scarves in Book 16 are excellent patterns for your	library	and ideal for using up oddments. It 's one not to miss. There 's lots of Guild informati
r own distinctive way, what is generally regarded as the filthiest joke in the jokester 's	library	. (Essentially it involves a circus family indulging in the vilest and most bizarre sexual
rters and cameramen are being scrambled in all directions, producers are working on	library	material, graphics and maps. Potential interviewees are being tracked down. But ther
verywhere. After further investigation (R Julie borrowed a book on hamsters from the	library	) we discovered it was a Russian hamster and that they were n't the friendliest of cre
oraal further explore the significance of her books being incorporated into Huygens 's	library	and the university library of Leydon. An interesting appendix is attached which includ
significance of her books being incorporated into Huygens 's library and the university	library	of Leydon. An interesting appendix is attached which includes several of the letters.
Giustiniani, but he had a zeal for scholarship, and from around 1620 he assembled a	library	and museum that included thousands of items. After Cassiano 's death in 1657 the c
eir Grand Tour. After the death of Federico Cesi in 1630 Cassiano also preserved the	library	and much of the literature of the Academia dei Lincei, the first scientific academy. Th
orms and performing well in interviews. You might find useful information in your local	library	, and there are useful tips online at: * Worktrain: www.worktrain.gov.uk * Careers Sc

Hit 1 – 30 of 244 | Previous Next

Diacollo: visualizing collocations over time



D*/DTA-K: DiaCollo

QUERY: Faulheit [submit]

DATE(S):

SLICE: 10

SCORE: log Dice (ld)

KBEST: 10

CUTOFF:

PROFILE: collocations

FORMAT: Cloud

GLOBAL: □

GROUPBY:

1PASS: □

DEBUG: □

Home Info Help Tutorial

Home Info Help Tutorial

jurish@bbaw.de

DiaColloDB v0.12.020 / DiaColloDB::WWW v0.02.005

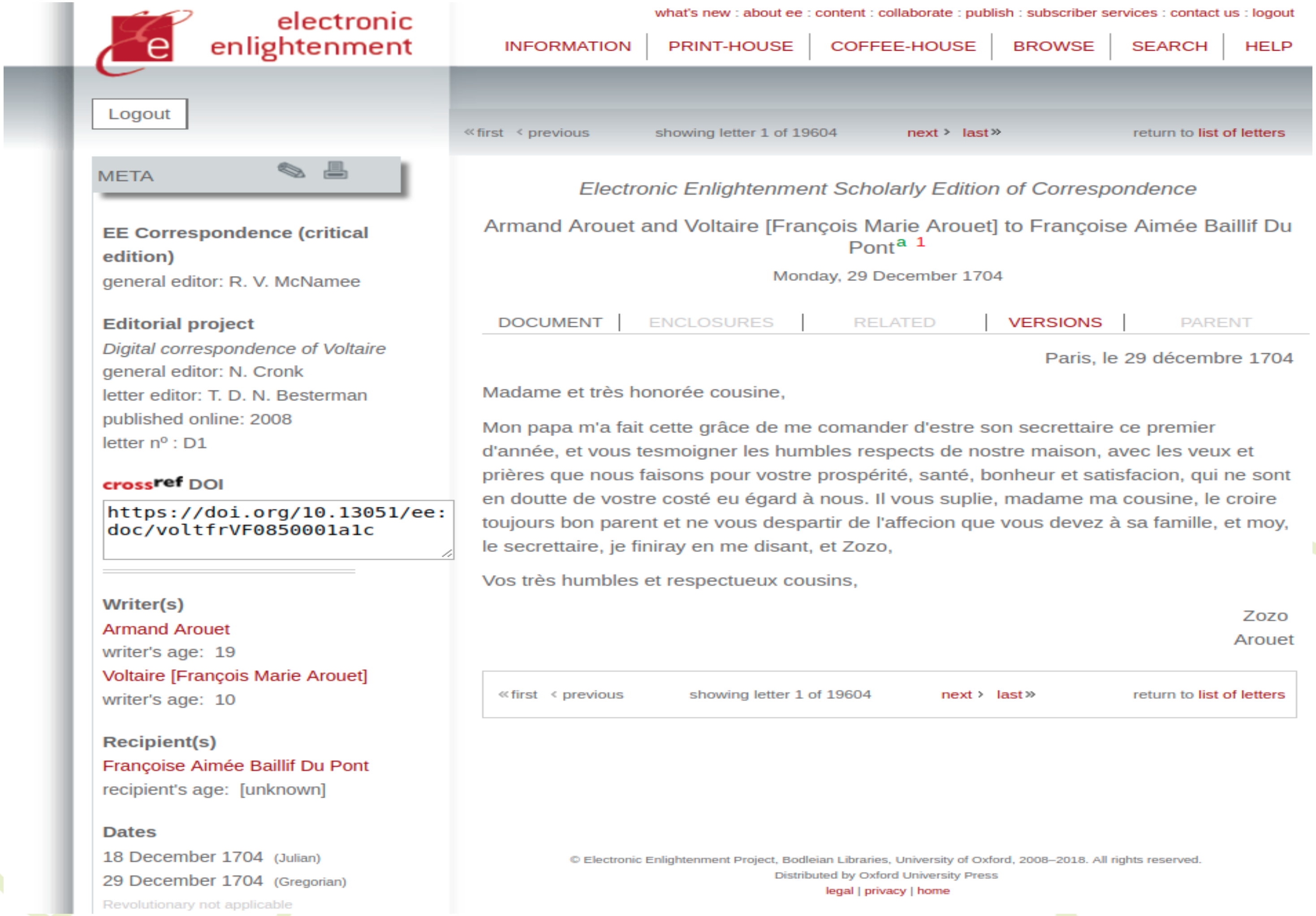
Imprint · Privacy

1.1013 sec

Corpus: *Deutsches Textarchiv* (DTA-Kernkorpus)

=cloud&groupby=&eps=0

Už dnes o tom ale na summitu NATO v Istanbulu	mluvil	s prezidentem Václavem Klausem . Fotbalistům Boleslavi se v lize
pomoc požádala její matka . Následujícím popisem . A to	nemluvě	o několika desítkách tisíc kostelů v majetku měst . Který
nepěstoval žádná zvláštní pouta s vlastí svého otce , ani	nemluví	maďarsky . Více informací nebo CD s obrázky si můžete
pokud to počasí dovolí . V nejbližší době spolu budeme	mluvit	, ale momentálně můj odchod nepřichází v úvahu . “
neplatit za odvedeou práci dělníkům ? Sice teď zrovna rusky	nemluvím	, ale myslím si , že tak během 14 dnů
do přečíslení dva ba jednoho . Ne že bych dokázal	mluvit	, ale už docela dobře rozumím . Slovenská metoda ,
Přestože je u nás už mnoho vietnamských dětí , které	mluví	lépe česky než vietnamsky , většina Vietnamců hovoří česky velmi
a kariéra sáňkaře Georga Hackla Více jak třetina obyvatel tu	mluvila	německy . Není to jistě málo , ale i kdyby
. Od chvíle , co se o tom dozvěděl ,	nemluvil	o ničem jiném než o svatbě . Návrh předkládal vládě
Moderátorka : Říká doktor Otakar Hulec a když už tady	mluvíme	o Kalinenu , paní doktorko , můžete říci , jak
o vyvlastňování ve " veřejném zájmu " , má smysl	mluvit	o ekologii jako o levicovém tématu ? Rozhodčí : K
od smrti princezny Diany v roce 1997 nikdy s policí	nemluvila	, protože měla strach , že bude zavražděna , ale
silnic a o valivém hluku od kol snad ani nelze	mluvit	; v kabině Pola je prostě opravdu ticho . S
kteří ji o to požádali . Znovu prý spolu budou	mluvit	ve čtvrtek v poledne a poté budou připraveni všechna tři
. Iversen (Nor .) , 10 . "	Mluví	ze mě osobní zkušenost , protože vaše knihy a nahrávky
zdrojů v příslušné oblasti apod . O ženách obecně pak	mluví	jako o čemsi , co nechápe , ale na druhé
Nečas by zřejmě chtěl , aby se o jeho vládě	mluvilo	jako o té , která vyrovnala rozpočty . Hned zkouším
fiktivních pravomocí) , jako někdo , kdo si dovolil	mluvit	se svým poslancem (a tím podezřelý z ovlivňování /
ke konzervativnější monetární politice . Konečně britský deník The Independent	mluví	o morální básnivosti , vytěžené z popela holocaustu : „
se dá . A jen pětina říká , že raději	nemluví	o nemocech . ◄ nová verze 3G mobilu READIUS Už
Hudební aranžmá je přítomno i v designu . Když jsem	mluvil	o přírodě , při chůzi po lese člověka napadnou zajímavější
rozhovoru sdělil později na tiskové konferenci : " Dnes jsem	mluvil	s velmi chytrým a milujícím mužem a musím přiznat ,
tématem a že se o nich bude v příhodný čas	mluvit	. A jisté city , jak je dobře známo ,
že jsem na technopárty byl , dobře se bavil a	mluvil	jak s účastníky (bylo mezi nimi dost studentek a



electronic enlightenment
what's new : about ee : content : collaborate : publish : subscriber services : contact us : logout
INFORMATION | PRINT-HOUSE | COFFEE-HOUSE | BROWSE | SEARCH | HELP
Logout
META

EE Correspondence (critical edition)
general editor: R. V. McNamee

Editorial project
Digital correspondence of Voltaire
general editor: N. Cronk
letter editor: T. D. N. Besterman
published online: 2008
letter nº : D1

crossref DOI

https://doi.org/10.13051/ee:doc/voltfrVF0850001a1c

Writer(s)
Armand Arouet
writer's age: 19
Voltaire [François Marie Arouet]
writer's age: 10

Recipient(s)
Françoise Aimée Baillif Du Pont
recipient's age: [unknown]

Dates
18 December 1704 (Julian)
29 December 1704 (Gregorian)
Revolutionary not applicable

« first ‹ previous showing letter 1 of 19604 next › last » return to list of letters

Electronic Enlightenment Scholarly Edition of Correspondence

Armand Arouet and Voltaire [François Marie Arouet] to Françoise Aimée Baillif Du Pont[a] [1]

Monday, 29 December 1704

Paris, le 29 décembre 1704

Madame et très honorée cousine,

Mon papa m'a fait cette grâce de me comander d'estre son secrettaire ce premier d'année, et vous tesmoigner les humbles respects de nostre maison, avec les veux et prières que nous faisons pour vostre prospérité, santé, bonheur et satisfacion, qui ne sont en doutte de vostre costé eu égard à nous. Il vous suplie, madame ma cousine, le croire toujours bon parent et ne vous despartir de l'affecion que vous devez à sa famille, et moy, le secrettaire, je finiray en me disant, et Zozo,

Vos très humbles et respectueux cousins,

Zozo
Arouet

« first ‹ previous showing letter 1 of 19604 next › last » return to list of letters

Hard problems with old texts

Mon papa m'a fait cette grâce de me comander d'estre son secrettaire ce premier d'année, et vous tesmoigner les humbles respects de nostre maison, avec les veux et prières que nous faisons pour vostre prospérité, santé, bonheur et satisfacion, qui ne sont en doutte de vostre costé eu égard à nous. Il vous suplie, madame ma cousine, le croire toujours bon parent et ne vous despartir de l'affecion que vous devez à sa famille, et moy, le secrettaire, je finiray en me disant, et Zozo,

Vos très humbles et respectueux cousins,

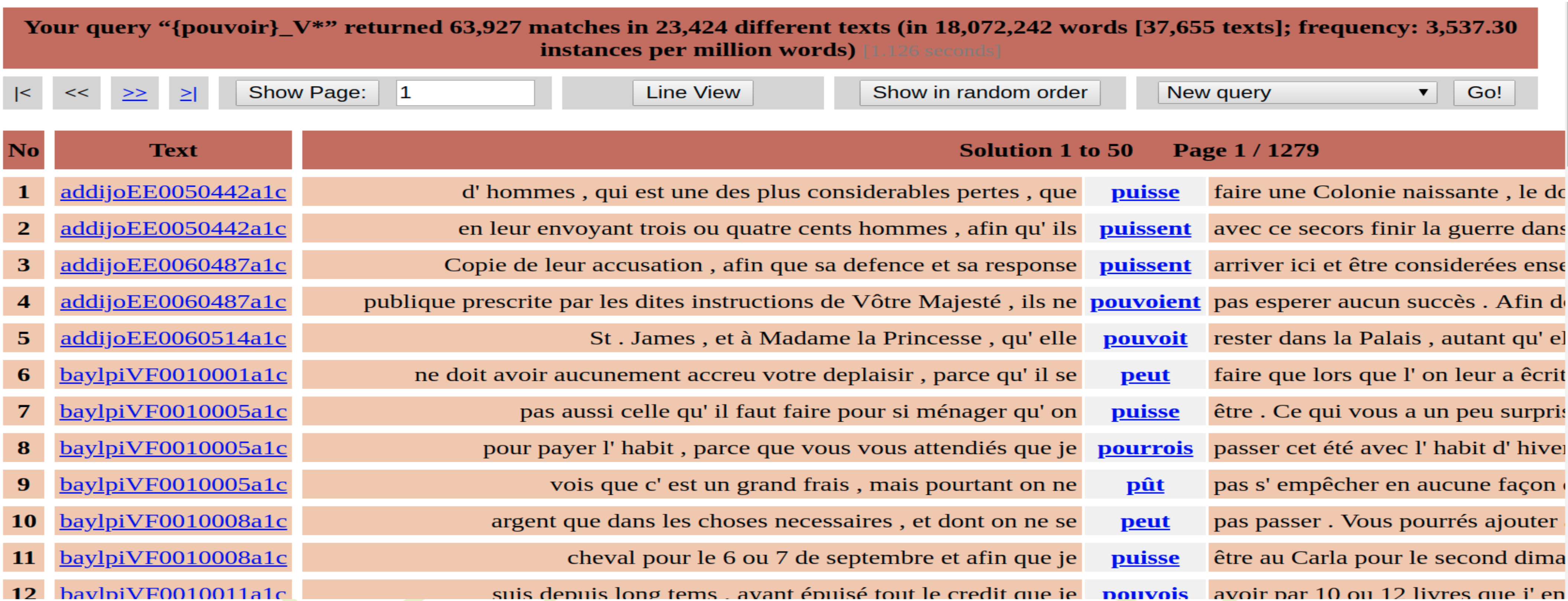
Zozo

Arouet

https://doi.org/10.13051/ee:doc/voltfrVF0850001a1c

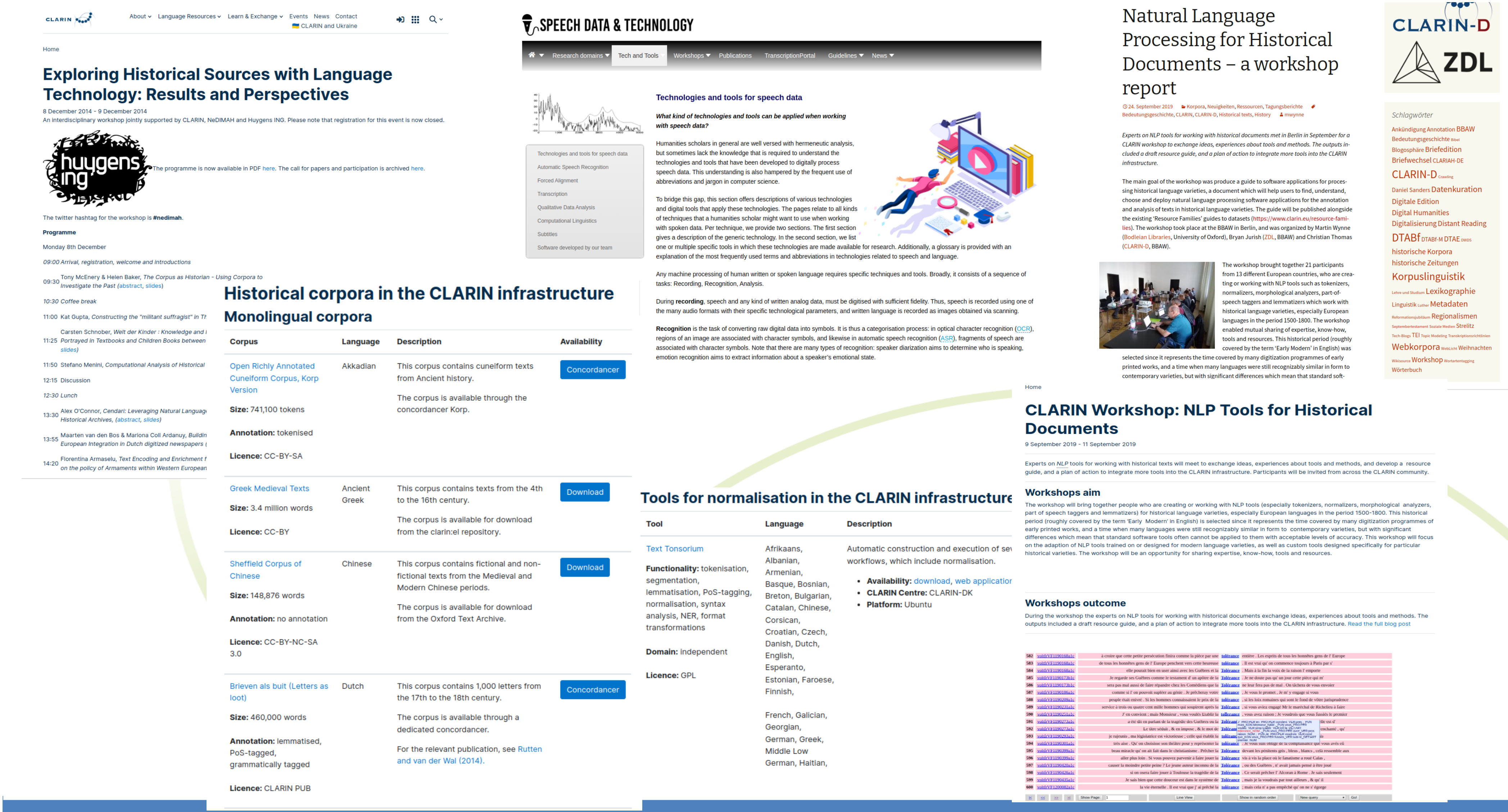
```
<text id="voltfrVF0850001a1c">
<p>
Paris       Paris       NAM       Paris
,           ,           PUN       ,
le          le          DET:ART   le
29          29          NUM       @card@
décembre    décembre    NOM       décembre
1704        1704        NUM       @card@
</p>
<p>
Madame      Madame      NOM       Madame
et          et          KON       et
très        très        ADV       très
honorée     honorée     ADJ       honoré
cousine     cousine     NOM       cousin
,           ,           PUN       ,
</p>
<p>
Mon         Mon         DET:POS   mon
papa        papa        NOM       papa
m'          m'          PRO:PER   me
a           a           VER:pres  avoir
fait        fait        VER:pper  faire
cette       cette       PRO:DEM   ce
grâce       grâce       NOM       grâce
de          de          PRP       de
me          me          PRO:PER   me
comander    commander             VER:infi   commander
estre       être        VER:infi  être
son         son         DET:POS   son
secrettaire             secrettaire NOM   secrettaire
ce          ce          PRO:DEM   ce
premier     premier     NOM       premier
d'          d'          PRP       de
année       année       NOM       année
,           ,           PUN       ,
et          et          KON       et
vous        vous        PRO:PER   vous
tesmoigner              témoigner VER:infi    témoigner
les         les         DET:ART   le
humbles     humbles     ADJ       humble
respects    respects    NOM       respect
de          de          PRP       de
nostre      nostre      DET:POS   notre
maison       maison       NOM maison
```

Which allows us to search by lemma, pos, original form, or modernized form



Your query "{pouvoir}_V*" returned 63,927 matches in 23,424 different texts (in 18,072,242 words [37,655 texts]; frequency: 3,537.30 instances per million words) [1.126 seconds]

No	Text	Solution 1 to 50 Page 1 / 1279		
1	addijoEE0050442a1c	d' hommes , qui est une des plus considerables pertes , que	puisse	faire une Colonie naissante , le do
2	addijoEE0050442a1c	en leur envoyant trois ou quatre cents hommes , afin qu' ils	puissent	avec ce secors finir la guerre dans
3	addijoEE0060487a1c	Copie de leur accusation , afin que sa defence et sa response	puissent	arriver ici et être considerées ense
4	addijoEE0060487a1c	publique prescrite par les dites instructions de Vôtre Majesté , ils ne	pouvoient	pas esperer aucun succès . Afin de
5	addijoEE0060514a1c	St . James , et à Madame la Princesse , qu' elle	pouvoit	rester dans la Palais , autant qu' el
6	baylpiVF0010001a1c	ne doit avoir aucunement accreu votre deplaisir , parce qu' il se	peut	faire que lors que l' on leur a êcrit
7	baylpiVF0010005a1c	pas aussi celle qu' il faut faire pour si ménager qu' on	puisse	être . Ce qui vous a un peu surpris
8	baylpiVF0010005a1c	pour payer l' habit , parce que vous vous attendiés que je	pourrois	passer cet été avec l' habit d' hiver
9	baylpiVF0010005a1c	vois que c' est un grand frais , mais pourtant on ne	pût	pas s' empêcher en aucune façon c
10	baylpiVF0010008a1c	argent que dans les choses necessaires , et dont on ne se	peut	pas passer . Vous pourrés ajouter
11	baylpiVF0010008a1c	cheval pour le 6 ou 7 de septembre et afin que je	puisse	être au Carla pour le second dima
12	baylpiVF0010011a1c	suis depuis long tems , ayant épuisé tout le credit que je	pouvois	avoir par 10 ou 12 livres que j' en



Home

Exploring Historical Sources with Language Technology: Results and Perspectives

8 December 2014 - 9 December 2014
An interdisciplinary workshop jointly supported by CLARIN, NeDIMAH and Huygens ING. Please note that registration for this event is now closed.

The programme is now available in PDF here. The call for papers and participation is archived here.

The twitter hashtag for the workshop is **#nedimah**.

Programme

Monday 8th December

09:00 Arrival, registration, welcome and introductions

09:30 Tony McEnery & Helen Baker, *The Corpus as Historian - Using Corpora to Investigate the Past* (abstract, slides)

10:30 Coffee break

11:00 Kat Gupta, *Constructing the "militant suffragist" in Th*

11:25 Carsten Schnober, *Welt der Kinder : Knowledge and I Portrayed in Textbooks and Children Books between slides)*

11:50 Stefano Menini, *Computational Analysis of Historical*

12:15 Discussion

12:30 Lunch

13:30 Alex O'Connor, *Cendari: Leveraging Natural Languag Historical Archives, (abstract, slides)*

13:55 Maarten van den Bos & Mariona Coll Ardanuy, *Buildin European Integration in Dutch digitized newspapers (*

14:20 Florentina Armaselu, *Text Encoding and Enrichment f on the policy of Armaments within Western European*

Technologies and tools for speech data
Automatic Speech Recognition
Forced Alignment
Transcription
Qualitative Data Analysis
Computational Linguistics
Subtitles
Software developed by our team

Technologies and tools for speech data

What kind of technologies and tools can be applied when working with speech data?

Humanities scholars in general are well versed with hermeneutic analysis, but sometimes lack the knowledge that is required to understand the technologies and tools that have been developed to digitally process speech data. This understanding is also hampered by the frequent use of abbreviations and jargon in computer science.

To bridge this gap, this section offers descriptions of various technologies and digital tools that apply these technologies. The pages relate to all kinds of techniques that a humanities scholar might want to use when working with spoken data. Per technique, we provide two sections. The first section gives a description of the generic technology. In the second section, we list one or multiple specific tools in which these technologies are made available for research. Additionally, a glossary is provided with an explanation of the most frequently used terms and abbreviations in technologies related to speech and language.

Any machine processing of human written or spoken language requires specific techniques and tools. Broadly, it consists of a sequence of tasks: Recording, Recognition, Analysis.

During **recording**, speech and any kind of written analog data, must be digitised with sufficient fidelity. Thus, speech is recorded using one of the many audio formats with their specific technological parameters, and written language is recorded as images obtained via scanning.

Recognition is the task of converting raw digital data into symbols. It is thus a categorisation process: in optical character recognition (OCR), regions of an image are associated with character symbols, and likewise in automatic speech recognition (ASR), fragments of speech are associated with character symbols. Note that there are many types of recognition: speaker diarization aims to determine who is speaking, emotion recognition aims to extract information about a speaker's emotional state.

Natural Language Processing for Historical Documents – a workshop report

24. September 2019 Korpora, Neuigkeiten, Ressourcen, Tagungsberichte Bedeutungsgeschichte, CLARIN, CLARIN-D, Historical texts, History mwynne

Experts on NLP tools for working with historical documents met in Berlin in September for a CLARIN workshop to exchange ideas, experiences about tools and methods. The outputs included a draft resource guide, and a plan of action to integrate more tools into the CLARIN infrastructure.

The main goal of the workshop was produce a guide to software applications for processing historical language varieties, a document which will help users to find, understand, choose and deploy natural language processing software applications for the annotation and analysis of texts in historical language varieties. The guide will be published alongside the existing 'Resource Families' guides to datasets (https://www.clarin.eu/resource-families). The workshop took place at the BBAW in Berlin, and was organized by Martin Wynne (Bodleian Libraries, University of Oxford), Bryan Jurish (ZDL, BBAW) and Christian Thomas (CLARIN-D, BBAW).

The workshop brought together 21 participants from 13 different European countries, who are creating or working with NLP tools such as tokenizers, normalizers, morphological analyzers, part-of-speech taggers and lemmatizers which work with historical language varieties, especially European languages in the period 1500-1800. The workshop enabled mutual sharing of expertise, know-how, tools and resources. This historical period (roughly covered by the term 'Early Modern' in English) was selected since it represents the time covered by many digitization programmes of early printed works, and a time when many languages were still recognizably similar in form to contemporary varieties, but with significant differences which mean that standard soft-

Schlagwörter

Ankündigung Annotation BBAW Bedeutungsgeschichte Bibel Blogosphäre Briefedition Briefwechsel CLARIAH-DE CLARIN-D Crawling Daniel Sanders Datenkuration Digitale Edition Digital Humanities Digitalisierung Distant Reading DTABf DTABf-M DTAE DWDS historische Korpora historische Zeitungen Korpuslinguistik Lehre und Studium Lexikographie Linguistik Luther Metadaten Reformationsjubiläum Regionalismen Septembertestament Soziale Medien Strelitz Tech-Blogs TEI Topic Modeling Transkriptionsrichtlinien Webkorpora WebLicht Weihnachten Wikisource Workshop Wortartentagging Wörterbuch

Historical corpora in the CLARIN infrastructure

Monolingual corpora

Corpus	Language	Description	Availability
Open Richly Annotated Cuneiform Corpus, Korp Version **Size:** 741,100 tokens **Annotation:** tokenised **Licence:** CC-BY-SA	Akkadian	This corpus contains cuneiform texts from Ancient history. The corpus is available through the concordancer Korp.	Concordancer
Greek Medieval Texts **Size:** 3.4 million words **Licence:** CC-BY	Ancient Greek	This corpus contains texts from the 4th to the 16th century. The corpus is available for download from the clarin:el repository.	Download
Sheffield Corpus of Chinese **Size:** 148,876 words **Annotation:** no annotation **Licence:** CC-BY-NC-SA 3.0	Chinese	This corpus contains fictional and non-fictional texts from the Medieval and Modern Chinese periods. The corpus is available for download from the Oxford Text Archive.	Download
Brieven als buit (Letters as loot) **Size:** 460,000 words **Annotation:** lemmatised, PoS-tagged, grammatically tagged **Licence:** CLARIN PUB	Dutch	This corpus contains 1,000 letters from the 17th to the 18th century. The corpus is available through a dedicated concordancer. For the relevant publication, see Rutten and van der Wal (2014).	Concordancer

Tools for normalisation in the CLARIN infrastructure

Tool	Language	Description
Text Tonsorium **Functionality:** tokenisation, segmentation, lemmatisation, PoS-tagging, normalisation, syntax analysis, NER, format transformations **Domain:** independent **Licence:** GPL	Afrikaans, Albanian, Armenian, Basque, Bosnian, Breton, Bulgarian, Catalan, Chinese, Corsican, Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Faroese, Finnish, French, Galician, Georgian, German, Greek, Middle Low German, Haitian,	Automatic construction and execution of sev workflows, which include normalisation. - **Availability:** download, web applicatior - **CLARIN Centre:** CLARIN-DK - **Platform:** Ubuntu

Home

CLARIN Workshop: NLP Tools for Historical Documents

9 September 2019 - 11 September 2019

Experts on NLP tools for working with historical texts will meet to exchange ideas, experiences about tools and methods, and develop a resource guide, and a plan of action to integrate more tools into the CLARIN infrastructure. Participants will be invited from across the CLARIN community.

Workshops aim

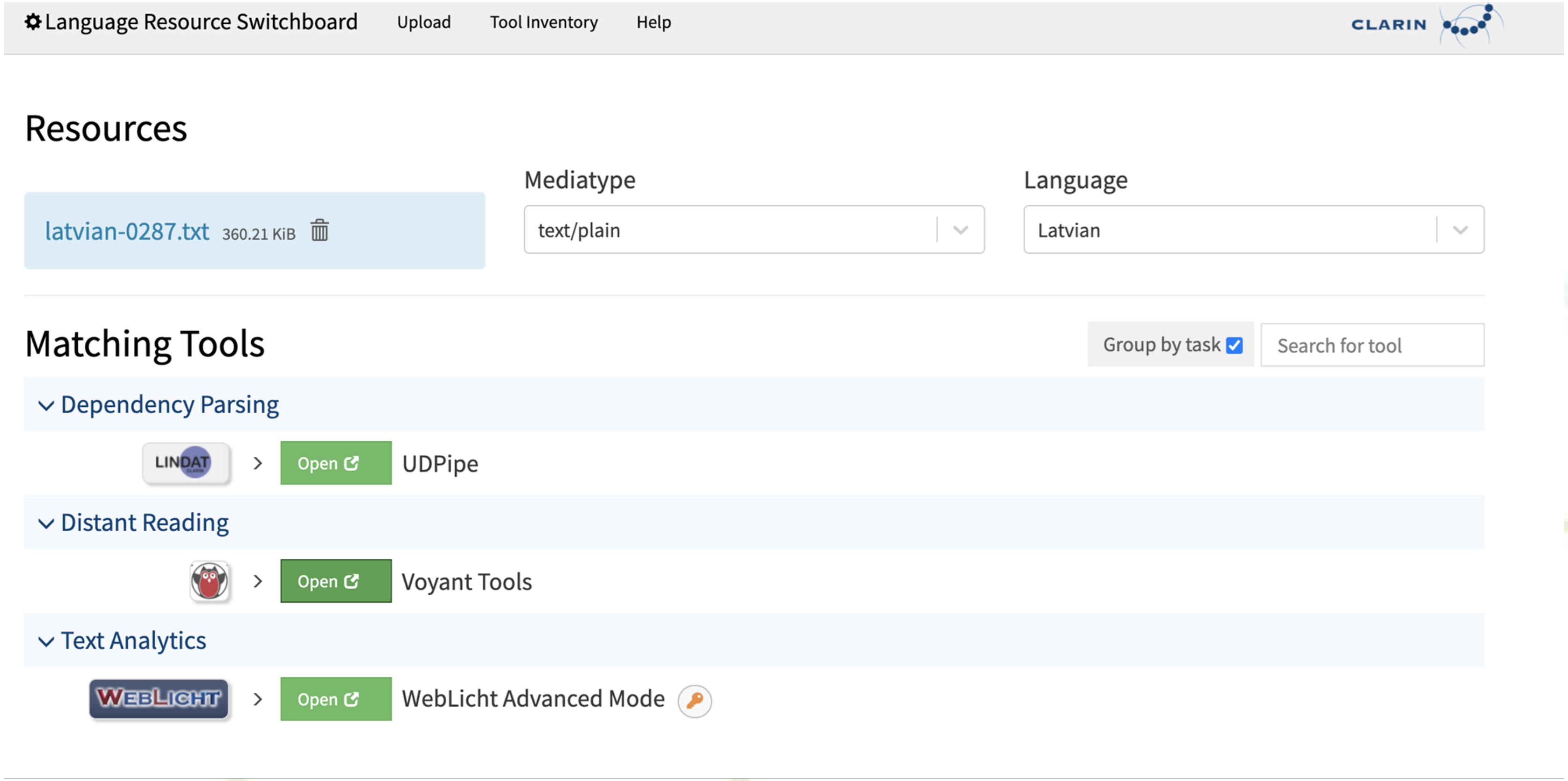
The workshop will bring together people who are creating or working with NLP tools (especially tokenizers, normalizers, morphological analyzers, part of speech taggers and lemmatizers) for historical language varieties, especially European languages in the period 1500-1800. This historical period (roughly covered by the term 'Early Modern' in English) is selected since it represents the time covered by many digitization programmes of early printed works, and a time when many languages were still recognizably similar in form to contemporary varieties, but with significant differences which mean that standard software tools often cannot be applied to them with acceptable levels of accuracy. This workshop will focus on the adaption of NLP tools trained on or designed for modern language varieties, as well as custom tools designed specifically for particular historical varieties. The workshop will be an opportunity for sharing expertise, know-how, tools and resources.

Workshops outcome

During the workshop the experts on NLP tools for working with historical documents exchange ideas, experiences about tools and methods. The outputs included a draft resource guide, and a plan of action to integrate more tools into the CLARIN infrastructure. Read the full blog post

582	voltfrVF1190168a1c	à croire que cette petite persécution finira comme la pièce par une	tolérance	entière . Les esprits de tous les honnêtes gens de l' Europe
583	voltfrVF1190168a1c	de tous les honnêtes gens de l' Europe penchent vers cette heureuse	tolérance	. Il est vrai qu' on commence toujours à Paris par s'
584	voltfrVF1190168a1c	elle pourait bien en user ainsi avec les Guèbres et la	Tolérance	. Mais à la fin la voix de la raison l' emporte
585	voltfrVF1190173b1c	Je regarde ses Guèbres comme le testament d' un apôtre de la	Tolérance	. Je ne doute pas qu' un jour cette pièce qui m'
586	voltfrVF1190173b1c	sera pas mal aussi de faire répandre chez les Comédiens que la	Tolérance	ne leur fera pas de mal . On tâchera de vous envoier
587	voltfrVF1190186a1c	comme si l' on pouvoit supléer au génie . Je prêcheray votre	tolérance	, Je vous le promet , Je m' y engage si vous
588	voltfrVF1190209a1c	peuple était enivré . Si les hommes connaissaient le prix de la	tolérance	, si les loix romaines qui sont le fond de vôtre jurisprudence
589	voltfrVF1190231a1c	service à trois ou quatre cent mille hommes qui soupirent après la	Tolérance	, si vous aviez engagé Mr le maréchal de Richelieu à faire
590	voltfrVF1190251a1c	J' en convient ; mais Monsieur , vous voulés Etablir la	tollerance	, vous avez raison ; Je voudrois que vous fussiés le premier
591	voltfrVF1190273a1c	a été dit en parlant de la tragédie des Guèbres ou la	Tolérance	elle est d'
592	voltfrVF1190273a1c	Le titre séduit , & en impose , & le mot de	Tolérance	enchanté , qu'
593	voltfrVF1190293a1c	je rajeunis , ma législatrice est victorieuse ; celle qui établit la	tolérance	is
594	voltfrVF1190301a1c	très aise . Qu' on choisisse son théâtre pour y représenter la	tolérance	. Je vous suis obligé de la complaisance que vous avés eû
595	voltfrVF1190398a1c	beau miracle qu' on ait fait dans le christianisme . Prêcher la	Tolérance	devant les pénitents gris , bleus , blancs , celà ressemble aux
596	voltfrVF1190398a1c	aller plus loin . Si vous pouvez parvenir à faire jouer la	Tolérance	vis à vis la place où le fanatisme a roué Calas ,
597	voltfrVF1190420a1c	causer la moindre petite peine ? Le jeune auteur inconnu de la	Tolérance	, ou des Guèbres , n' avait jamais pensé à être joué
598	voltfrVF1190426a1c	si on osera faire jouer à Toulouse la tragédie de la	Tolérance	. Ce serait prêcher l' Alcoran à Rome . Je sais seulement
599	voltfrVF1190435a1c	Je sais bien que cette douceur est dans le systéme de	Tolérance	, mais je la voudrais par tout ailleurs , & qu' il
600	voltfrVF1200082a1c	la vie éternelle . Il est vrai que j' ai prêché la	tolérance	; mais cela n' a pas empêché qu' on ne s' égorge

J' PRO:PER en PRO:PER convient VER:pres ; PUN mais KON Monsieur NAM , PUN vous PRO:PER voulés VER:simp Etablir VER:infi la DET:ART tollerance NOM , PUN vous PRO:PER avez VER:pres raison NOM ; PUN Je PRO:PER voudrois VER:cond que KON vous PRO:PER fussiés VER:subi le DET:ART premier NUM



Resources

latvian-0287.txt 360.21 KiB

Mediatype: text/plain

Language: Latvian

Matching Tools

Group by task ☑ Search for tool

Dependency Parsing

LINDAT › Open UDPipe

Distant Reading

› Open Voyant Tools

Text Analytics

WebLicht › Open WebLicht Advanced Mode

switchboard.clarin.eu

Some possible next steps

So some key areas for discussion about actual and potential collaboration are:

- Sharing the platforms and tools for delivering textual data that come from the computational and corpus linguistics communities, many now curated and developed by CLARIN centres,
- Tools for analysis and annotation of texts to enable more effective search
- How to connect tools and texts, flexibly and sometimes in processing pipelines

Other goals for the workshop:

- Find out more about each other: our data, tools, methods, plans, etc.
- Use CLARIN-LIB email list
- Pursue joint projects?
- A joint paper for TPDL?
- Another workshop – topic?