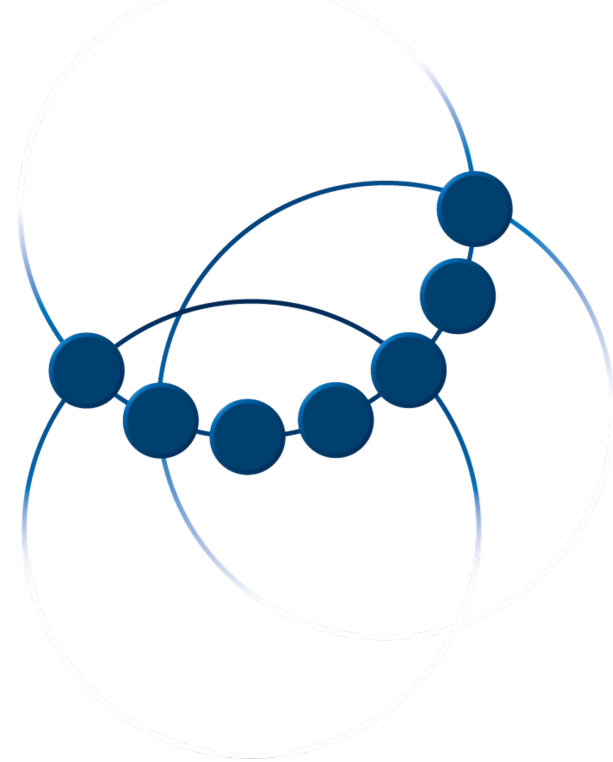


CLARIN

Common Language Resources and
Technology Infrastructure



Things to do with data: The CLARIN European Research Infrastructure

Martin Wynne

Research Support Team, Academic IT
martin.wynne@it.ox.ac.uk

IT Services &
Oxford e-Research Centre &
Faculty of Linguistics, Philology and Phonetics,
University of Oxford
Director of User Involvement, CLARIN ERIC



ITLP

Wednesday 28th January 2015

Outline of today's talk

- **What sort of data and what sort of research?** Language resources and tools in research in the humanities and social sciences
- **What are the problems?** Sharing, sustaining, access, ease of use, etc.
- **What's the solution?** CLARIN!
- **How's that going?** Barriers, problems, success stories and lessons
- **How do I find out more?**

What sort of data?

- A language corpus is a collection of **naturally occurring** language data compiled to represent a language or language variety
- In electronic form
- It can contain more than text (e.g. annotations, metadata, other media)

TEI Texts Corpora Legacy formats

Search: Show **10** entries [First](#) [Previous](#) [1](#) [2](#) [3](#) [4](#) [5](#) [Next](#) [Last](#)

ID	Title	Author	Date	Language	Availability	Genre
3193	The Lampeter Corpus of Early Modern English Tracts	Josef Schmied, Claudia Claridge, Rainer Siemund (ed.)	1800	English	CC BY-SA	
2555	French Oral Narrative Corpus	Janice Carruthers	2013-08	French	Free for non-commercial use	Linguistic corpora
2554	British National Corpus, XML edition	BNC Consortium	2007-03-06	English	BNC User Licence	Linguistic corpora
2553	British National Corpus, Baby edition	BNC Consortium	2007-03-06	English	BNC User Licence	Linguistic corpora
2551	British National Corpus Sampler	BNC Consortium	2007-03-06	English	BNC User Licence	Linguistic corpora
2544	GerManC	Martin Durrell; Paul Bennett; Silke Scheible; Richard J. Whitt;	2012-05	German	Free for non-commercial use	Linguistic corpora
2542	VOICE: Vienna-Oxford International Corpus of English	Barbara Seidlhofer ; Angelika Breiteneder; Theresa Klimpfinger; Stefan Majewski; Ruth Osimk-Teasdale (POS-tagged versions); Marie-Luise Pitzl; Michael Radeka (POS-tagged versions)	2011-08	English	Free for non-commercial use	Linguistic corpora
2541	VU Amsterdam Metaphor Corpus	Gerard J Steen; Aletta G Dorst; J Berenike Herrmann; Anna A Kaal; Tina Krennmayr	2010-09	English	Free for non-commercial use	Linguistic corpora
2540	Speech, Thought and Writing Presentation Corpus (STWP)	Culpeper, Jonathon; Semino, Elena; Short, Mick; Wynne, Martin	2008-10	English	Free for non-commercial use	Linguistic corpora
2539	British Academic Written English Corpus	Nesi, Hilary; Gardner, Sheena; Thompson, Paul; Wickens, Paul	2008-10	English	Restricted	Linguistic corpora

Showing 1 to 10 of 71 entries [First](#) [Previous](#) [1](#) [2](#) [3](#) [4](#) [5](#) [Next](#) [Last](#)

TEI Texts

Created following the guidelines of the Text Encoding Initiative, available in a variety of formats to read, download or link to.

Corpora

Collections of language data comprising a number of texts from different sources, usually compiled for the purposes of linguistic research

Legacy formats

A variety of resources which have been collected since the OTA came into existence in 1976. Some have restricted access, which means that you have to ask for permission to use them, and many are in formats which some users may find difficult to use, although most are plain text. The OTA cannot offer support for dealing with legacy formats.

Some types of corpus

- Reference corpus:
 - General, big and ‘balanced’, representative of a language or variety, static and closed (no more material added)
- Monitor corpus:
 - Big, with new material added over time, used to monitor change
- Historical corpus:
 - Synchronic (from one time) or diachronic (from several time periods)
- Parallel corpus:
 - Same texts in two or more languages
- Multi-media/multi-modal corpus:
 - Not only text (sound, images, gesture)

More than text

The corpus may also contain other information vital to searching and interpreting results, e.g.:

- information about the texts, speakers, context etc. (*metadata*)
- linguistic information: parts of speech, lemmata, grammatical structure, etc (*annotation*)
- other information: discourse structure, extra-linguistic features, etc. (*structural markup*)

How do you use a corpus?

A typical workflow:

- lexical starting point (i.e. start by searching for a word or phrase)
- retrieve examples
 - examine quantitatively (count things)
 - look for patterns
 - try to account for meanings and functions
 - adjust and search parameters and search again
- Compare and contrast, summarise
- Construct a hypothesis, investigate further

Your query "spooft" returned 62 hits in 44 different texts (98,313,429 words [4,048 texts]; frequency: 0.63 instances per million words)

No	Filename	Hits 1 to 50	Page 1 / 2
1	A2D 38	with his conservative values,' he said. A series of	spooft or obsessively realistic 1960s westerns (Cat Balou, Butch Cassidy and
2	A7L 1262	's class structure, but seemed all at sea with his Bond	spooft , Modesty Blaise (1966). There simply isn't any
3	A9T 40	' choice What's going on. By Stephanie Nettell FAST-MOVING comic-strip	spooft on US life and politics, Phoenix Beam's cult American Eagle
4	AAH 199	Harry Enfield. Having worked his way up via his own live	spooft chat show Vic's Big Night Out, he has latterly been
5	ABW 322	. Later, there was frenzied activity in the creation of a	spooft campaign for Tunnel Vision, a 'new' TV channel.
6	ABW 389	with loose shoes. Why travel thousands of miles to visit a	spooft Paris? It was inevitable, she supposed, when America was
7	ACP 126	the street, this club flyer advertises Newcastle's Rockshots. A	spooft on washing powder packets, the colourful T-shirts are available from the
8	ACP 2313	small screen for the last two years (the South Bank Show	spooft , Norbert Smith , was a revamp of an old idea)
9	ADR 450	factor'. Some declared Minogue-free zones and preferred to play a	spooft record instead, 'I Should Be So Yucky'. The
10	AE6 1463	Thackeray began, but does not appear to have finished, a	spooft , now in the New York Public Library, which begins:
11	AN0 351	the 1955 Berlin Tunnel (see Chapter 2), proceeded to	spooft Western intelligence for several years with false information. But because so
12	AP0 902	. Polanski arrived in Hollywood from Europe on the back of a	spooft horror film called The Fearless Vampire Killers , a parody of the
13	APU 1490	Lorna made a comedy western, a beach musical and a spy	spooft in slow succession. Hitchcock wanted her for The Birds but she
14	B3J 2977	leave here, so does Fred. He thinks he can play	spooft , but I usually come out of the pub legless with as
15	B7J 2129	I love about New Scientist is the way you print the odd	spooft article or two in the issue nearest 1 April. It does
16	B7J 2130	. It does rather spoil the fun, though, when the	spooft is too obvious — the one in the 31 March issue was
17	BNT 109	had rejected it. Funnily enough, Punch ran exactly the same	spooft , with the same book, on 10 April 1991. If
18	BP4 523	the time word got out that the whole thing had been a	spooft , there was a sizeable demand for the book and, rather
19	CBC 10844	Wind — these are just some of the hilarious works of the	spooft heavy metal band whose new book Inside Spinal Tap will fill many
20	CBC 11760	22) Mrs Robinson Lemonheads Diana's got it taped as royal	spooft hits street By Tim Satchell THE Princess of Wales has obtained a
21	CBC 11762	obtained a copy of the video version of Pallas 2, the	spooft soap opera which pokes fun at the Royal Family and which will
22	CBE 1464	Stallone, 46, says he is 'flattered' that a	spooft of his Rambo films is being made. Charlie Sheen, who
23	CBE 1465	Shots 2, said: 'I'd be flattered if a	spooft was made of one of my movies.' Travolta stays quiet
24	CBE 2549	the Conference on European Security and said his speech had been a	spooft . He wanted to show how bad things could become if the
25	CBE 8446	down bets on a Charles and Diana split. Drama at TV	spooft SATIRICAL film producer Geoff Atkinson has been forced to hurriedly re-edit his

What sorts of research can you do with a corpus?

- Insights into real language usage
- Insights into history and culture via the writing and debates of the past
- Tracking social attitudes and social change
- Identifying interesting areas to explore in the archives in the textual record

Big data leads to digital transformations

- Data-driven research into language
“What words and structures do I find in this sample of language?”
- Asking new research questions
“How did the grammar of everyday English speech and writing change across the twentieth century?”
- Answering old questions in different ways
“How much free indirect speech does Jane Austen use in her novels?”
- Re-examine old questions and statements which were formulated in the absence of systematic use of data
“Is it true that writers only started to refer to 'the state' in its modern political sense in the early sixteenth century?”

But...

- How do you read a million books?
- How do you reconcile the deep knowledge of texts, and close reading of them, with broad brush overviews, statistics, digests and trends?
- Do we still need language corpora in the age of the online data deluge?
- Have we got the infrastructure to support work like this in the humanities?

Interoperability and sustainability for digital textual scholarship



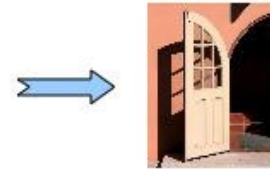
Well-known problems with digital resources relating to:

- fragmentation of communities, resources, tools;
- lack of connections, integration, and interoperability;
- sustainability of online services;
- lack of deployment of tools as reliable and available services

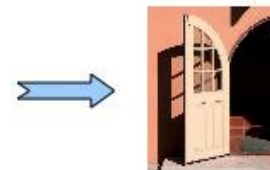
There is a potential solution in distributed, federated infrastructure services.



⇒ Resource discovery services

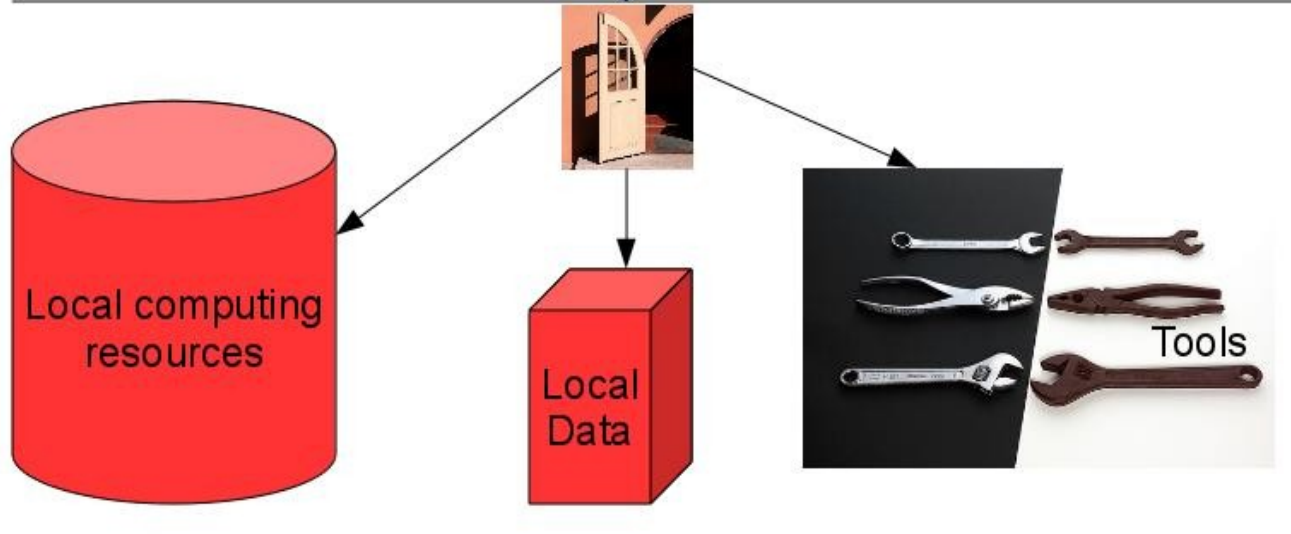
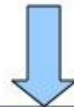


⇒ Data



⇒ Website with data and tools

Single sign on?



Advisory services?



?Cloud computing resources?

The CLARIN Vision



A researcher in the Oxford, from his desktop computer, can:

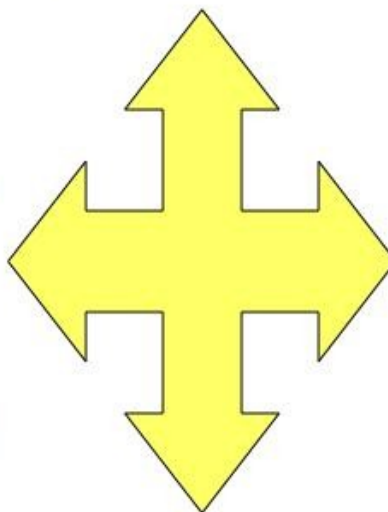
- Sign on with local authentication, and then:
- search for, find and obtain authorization to use corpora in Lancaster, Prague and Berlin
- select the precise dataset to work on, and save that selection
- run semantic analysis tools from Budapest and statistical tools from Tübingen over the dataset
- use computational power from the local, national or other computing centre where necessary
- obtain advice and support for carrying out all technical and methodological procedures
- save the workflow and results of the analysis, and share those results with collaborators in Paris, Vienna and Zagreb
- discuss and iteratively adopt and re-run the analyses with collaborators



Advisory services



Resource discovery services



Cloud computing resources

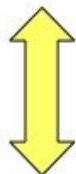


Local computing resources

Local Data



Datasets



Technical requirements for a distributed data system

- Data format standards
- Annotation standards
- Metadata standards
- Persistent identifiers and locations
- High availability for services
- A common authentication realm
- Documentation, support, and training for service providers
- Support for users
- Citable data and workflows
- Long-term business models



WHY?

- Promote effective use language resources and technologies across the social sciences and humanities
- Overcome the fragmentation of existing language resources and tools
- Help to combine datasets and tools
- Overcome some of the high barriers of entry to digital research
- Help to make valuable resources emerging from research projects available for sharing and reuse

WHAT IS CLARIN?

- European Research Infrastructure Consortium (**ERIC**) on the European Strategy Forum on Research Infrastructures (ESFRI) roadmap
- Set of national **infrastructures**
- **Network** of technical centres, data providers, researchers, developers
- **Community** of researchers using language resources and tools to do new forms of digital research in the humanities and social sciences

WHO?

- Austria, Bulgaria, Czech Republic, Denmark, Estonia, Germany, Greece, Lithuania, Netherlands, Norway (observer), Poland, Portugal, Sweden, Dutch Language Union.
- Croatia, Finland, France, Italy, Slovenia and the UK in the process of joining
- Contacts and activities in all other European countries
- Millions of users can log in to access resources, thanks to the CLARIN Federation

HOW MUCH?

- Preparatory phase c. €4m from FP7 project
- National infrastructure investments of c. €10m in 2014
- €1m p.a. paid by members towards central coordination
- Project funding income to CLARIN ERIC and centres

WHAT DOES CLARIN DO?

- Provides access to language data
- Makes software available as as services
- Links together repositories
- Links tools and data
- Offers training, advice, support
- Promotes the use of digital text and speech among scholars across the humanities and social sciences

WHEN?

- Preparation: 2008-2011
- Construction: 2011-2015
- Operation: 2013-
- Long-term perspective – no fixed end date

WHO IS USING CLARIN?

Centres, tools and datasets bring users with them – not yet all measured as 'CLARIN users', but we do know:

- Virtual Language Observatory: 700,000 records, 32,000 visits in 2014
- More than 1,500 monthly unique visits to clarin.eu website
- More than 150 research papers published in 2013 mentioning CLARIN, according to recent bibliometric research

The first part of the jigsaw – resource discovery across multiple repositories (including Oxford)...

Virtual Language Observatory
Explore the world of language resources and technology from different perspectives

VLO > [Faceted search](#) > [Selections](#): Old English ✕ Oxford Text Archive ✕ [Link](#) | [Report](#) | [Help](#)

SEARCH

SEARCH RESULTS

20 results << < 1 2 >>> Showing 1 to 10

[Alfred's Orosius](#) [Electronic resource] Expand
Resources: | 1 other |

[Ancrene Wisse Preface](#) [Electronic resource] Expand
Ancrene Wisse, a Middle English 'rule' or guide for female recluses, was composed in the West Midlands in the early thirteenth century. It is a carefully-constructed work, divided into the Preface edited here and eight parts (called by the author distinctions). This edition offers a corrected form of the text of the P...
Resources: | 1 other |

[Anglo Saxon Chronicle : selections / compiled by Anita Dowsing](#) Expand
Resources: | 1 other |

[Anglo-Saxon Charter Boundaries material](#) [Electronic resource] Expand
The resource consists of files made at different times and on different machines during the time of compilation. It is useful from an historical perspective only since the material is now being set up as the AHRB-funded 'Language of Landscape: Reading the Anglo-Saxon Countryside' project (started in October 2004), whic...

NARROW DOWN
Use the categories below to limit the search results to those matching the selected value(s).

- LANGUAGE
Old English ✕
- + COLLECTION
Oxford Text Archive ✕
- + RESOURCE TYPE
- + SUBJECT
- FORMAT

Type to search for more

unknown type (20)

- + ORGANISATION
- + DATA PROVIDER

[Expand all](#) +
[Collapse all](#) -
[Deselect all](#) ✕

+ SEARCH OPTIONS



...and from separate online interfaces to resources...

Your query "spooft" returned 62 hits in 44 different texts (98,313,429 words [4,048 texts]; frequency: 0.63 instances per million words)

No	Filename	Hits 1 to 50	Page 1 / 2
1	A2D 38	with his conservative values,' he said. A series of	spooft or obsessively realistic 1960s westerns (Cat Balou, Butch Cassidy and
2	A7L 1262	's class structure, but seemed all at sea with his Bond	spooft , Modesty Blaise (1966). There simply isn't any
3	A9T 40	' choice What's going on. By Stephanie Nettell FAST-MOVING comic-strip	spooft on US life and politics, Phoenix Beam's cult American Eagle
4	AAH 199	Harry Enfield. Having worked his way up via his own live	spooft chat show Vic's Big Night Out, he has latterly been
5	ABW 322	. Later, there was frenzied activity in the creation of a	spooft campaign for Tunnel Vision, a 'new' TV channel.
6	ABW 389	with loose shoes. Why travel thousands of miles to visit a	spooft Paris? It was inevitable, she supposed, when America was
7	ACP 126	the street, this club flyer advertises Newcastle's Rockshots. A	spooft on washing powder packets, the colourful T-shirts are available from the
8	ACP 2313	small screen for the last two years (the South Bank Show	spooft , Norbert Smith , was a revamp of an old idea)
9	ADR 450	factor'. Some declared Minogue-free zones and preferred to play a	spooft record instead, 'I Should Be So Yucky'. The
10	AE6 1463	Thackeray began, but does not appear to have finished, a	spooft , now in the New York Public Library, which begins:
11	AN0 351	the 1955 Berlin Tunnel (see Chapter 2), proceeded to	spooft Western intelligence for several years with false information. But because so
12	AP0 902	. Polanski arrived in Hollywood from Europe on the back of a	spooft horror film called The Fearless Vampire Killers , a parody of the
13	APU 1490	Lorna made a comedy western, a beach musical and a spy	spooft in slow succession. Hitchcock wanted her for The Birds but she
14	B3J 2977	leave here, so does Fred. He thinks he can play	spooft , but I usually come out of the pub legless with as
15	B7J 2129	I love about New Scientist is the way you print the odd	spooft article or two in the issue nearest 1 April. It does
16	B7J 2130	. It does rather spoil the fun, though, when the	spooft is too obvious — the one in the 31 March issue was
17	BNT 109	had rejected it. Funnily enough, Punch ran exactly the same	spooft , with the same book, on 10 April 1991. If
18	BP4 523	the time word got out that the whole thing had been a	spooft , there was a sizeable demand for the book and, rather
19	CBC 10844	Wind — these are just some of the hilarious works of the	spooft heavy metal band whose new book Inside Spinal Tap will fill many
20	CBC 11760	22) Mrs Robinson Lemonheads Diana's got it taped as royal	spooft hits street By Tim Satchell THE Princess of Wales has obtained a
21	CBC 11762	obtained a copy of the video version of Pallas 2, the	spooft soap opera which pokes fun at the Royal Family and which will
22	CBE 1464	Stallone, 46, says he is 'flattered' that a	spooft of his Rambo films is being made. Charlie Sheen, who
23	CBE 1465	Shots 2, said: 'I'd be flattered if a	spooft was made of one of my movies.' Travolta stays quiet
24	CBE 2549	the Conference on European Security and said his speech had been a	spooft . He wanted to show how bad things could become if the
25	CBE 8446	down hats on a Charles and Diana split. Drama at TV	spooft SATIRICAL film producer Geoff Atkinson has been forced to hurriedly re-edit his

...to cross-searching distributed resources...



Zukunft

DWDS Standardsicht +Ressourcen

DWDS-Wörterbuch

in Zukunft soll die Produktion von synthetischen Fasern gesteigert werden

etw. hat Zukunft etw. lässt eine günstige Entwicklung erwarten, ist erfolgversprechend

diese Technologie hat (keine) Zukunft

künftiger Lebensweg eines Menschen

wie stellst du dir deine Zukunft vor?

Sprachwissenschaft **Tempus des zukünftigen Geschehens, Futur**

in dem Satz »ich werde ihm schreiben« steht »schreiben« in der Zukunft

Dazu:

- Zukunftsabsicht, Zukunftsaussicht, Zukunftsbild, Zukunftsforschung, Zukunftsgemälde, Zukunftsglaube, Zukunftsmusik, Zukunftsmöglichkeit, Zukunftsperspektive, Zukunftsplan, Zukunftsproblem, Zukunftsprognose, Zukunftsprojekt, Zukunftsroman, Zukunftsicherheit, Zukunftsraum, Zukunftsvision, zukunftsfröhlich, zukunftsreich, zukunftsreich, zukunftsreich, zukunftsreich, zukunftsreich
- zukünftig

Version: 0.4.23 | Quelle: WDG | Artikeltyp: Vollartikel Kompakt | Details

Etymologisches Wörterbuch (nach Pfeifer)

kommen, ferner: Kunft, Abkunft, Ankunft, Auskunft, Auskunftei, Einkünfte, Herkunft, Niederkunft, Übereinkunft, Unterkunft, Zukunft, zukünftig, Zusammenkunft, abkommen, Abkommen, Abkomme, abkömmlich, ankommen, Ankömmling, aufkommen, auskommen, Auskommen, auskömmlich, einkommen, Einkommen, entkommen, herkommen, Herkommen, herkömmlich, nachkommen, Nachkomme, Nachkommenschaft, Nachkömmling, niederkommen, übereinkommen, überkommen, überkommen, umkommen, unterkommen, verkommen, Verkommenheit, vorkommen, Vorkommnis, zukommen

kommen Vb. 'sich einem Ort nähern, sich von einem Ort hierher begeben, einen Ort erreichen, eintreffen, erscheinen', ahd. *queman* (8. Jh.), mhd. (mit Verschmelzung von *u* und *e* zu *o*) *komen*, asächs. *kuman*, mnd. *kōmen*, *kāmen*, *kommen*, mnl. *cōmen*, nl. *komen*, aengl. *cuman*, engl. *to come*, got. *qiman*, anord. *koma*, schwed. *komma* führt mit aind. *gāmati* 'geht', toh. A *kām-*, *kam-*, *kum-*, toh. B *kā(n)m-*, *kam-*, *sem-* 'kommen', griech. *βαίνειν* (*baínein*) 'gehen', lat. *venire* 'kommen', lit. *gīmti* 'zur Welt kommen, geboren werden, entstehen' auf eine Wurzel ie. **gyem-* 'gehen, kommen, zur Welt kommen, geboren werden'. Im Dt. ist der ursprüngliche Anlaut *qu-* (germ. *kw-*) im Verbaladjektiv *bequem* (s. d.) erhalten. – **Kunft** f. 'das Kommen, Ankommen' (selbständiges Substantiv bis ins 19. Jh.), ahd. *kumft* (8. Jh.),

Version: 1.0.69

OpenThesaurus

Synonymgruppen für Zukunft

Synonymgruppe: Futur, Zukunft

Oberbegriffe: Tempus, Zeit, Zeitform

Version: 2014-04-14 | Quelle: OpenThesaurus

DWDS-Wortprofil 3.0

Abfragewort: Zukunft Vergleichswort: x

Substantiv logDice 36

Überblick zu 'Zukunft'

abhängt absehbarer als Gegenwart als Nostalgie als Vergangenheit berufliche entgegenblickt entgegengieht entgegensehen entschieden Erinnerung ferner Fernsehen Gegenwart Gegenwart als Gegenwart wie gesichert gestalten gestaltet Herausforderungen Jugend Markt offen sichern ungeklärt ungewisse ungewiß unklar unsicher Verantwortung Vergangenheit

Deutsches Textarchiv (wöchentlich aktualisiert)

Treffer: 4823 Filter

KWic	Datum ↓	Datum ↑	Zufällig	Links	Rechts
1	1905				Was die Zukunft dieser logischen Disziplin
2	1903				Aber die Zukunft ist kein Traum .
3	1903				Und die Menschheit der Zukunft mag wohl den Tag segnen
4	1903				takt mit der Menschheit , mit der Zukunft halten .
5	1903				tendes Farbenband in die reale Zukunft der Menschheit , in die gro
6	1903				n ein ganz fernes Lichtideal der Zukunft sein könnte , wird an den A
7	1903				Ideal , die wieder einmal zu der Zukunft sich auch die älteste Verg
8	1903				Die Zukunft lag so blau .
9	1903				n Wiederaufleben der höchsten Zukunft in der tiefsten Vergangenh
10	1903				gangenheit hingegeben um der Zukunft willen .
11	1903				nehr gleichgültig sein , ob diese Zukunft doch nur eine verkappte Vr
12	1903				Die sozialen Siege der Zukunft werden nichts Halbes sein
13	1902				die Ehe sei in den Sternen der Zukunft schon ausgelöscht , so ra

Kernkorpus 20

Treffer: 13444, davon anzeigbar: 11790 Filter

KWic	Datum ↓	Datum ↑	Zufällig	Links	Rechts
1	1999				Allerdings beginnt ja die Zukunft mit jedem Augenblick neu
2	1999				ellschaft oder die Vision einer Zukunft namens Kommunismus au
3	1999				Lindens hätte es in Berlin keine Zukunft mehr gegeben .
4	1999				mbetriebes gegenübersähe , in Zukunft womöglich noch mehr als f
5	1999				o was schadet , jetzt und für die Zukunft .
6	1999				ε , um dann hoffnungsvoll in die Zukunft zu blicken : Wir haben Lich
7	1999				verser Anfang , der Streit für die Zukunft verhielß .
8	1999				ues Lebensgefühl mitzureißen : Zukunft war doch machbar !
9	1999				imens hat der Staatshaushalt in Zukunft nichts mehr zu tun .
10	1999				Vergangenheit bzw. der fernen Zukunft an , und so blieb es nicht a
11	1999				r Ausklang lag in unbestimmter Zukunft , und so mochte es manchr
12	1999				det , damit jeder Ehegatte seine Zukunft unabhängig von materielle
13	1900				mster zu nehmen als eine vane Zukunft

...and distributed web service orchestration

Browser address bar: <https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/>

Navigation: Main Page | Chain 1 x +

View Tool List | HELPDESK

Available Annotations for: German Plain Text

- POS Tags/Lemmas
- Morphology
- Constituent Parses
- Dependency Parses
- Named Entities**

German Named Entity Recognizer

Annotation Layers: language = de

Download as Excel sheet | Download as CSV

[1 - 30 / 91]

token ID	tokens	POStags	lemmas	namedEntities
t1	Damals	ADV	damals	
t2	in	APPR	in	
t3	Odessa	NE	Odessa	B-GPE
t4	war	VAFIN	sein	
t5	es	PPER	es	
t6	sehr	ADV	sehr	
t7	kalt	ADJD	kalt	

Download TCF

Input and Chain Selection

Run Tools | Clear Results | Download chain

My Input Plain Text Damals in Odessa war es sehr kalt. Wir fuhren jeden Morgen mit groÙen rappenden Lastwagen über das Kopfsteinpflaster zum	SFS To TCF Converter Language: German Document Type: TCF TCF Version: 0.4 Text	IMS Tokenizer Sentences Tokens	IMS TreeTagger Part of Speech: STTS Tagset Lemmas	SFS German Named Entity Named Entities: tuebadz8
---	---	---	--	--

Done running tools.

How does this work in the University of Oxford?

A collaboration between:

- IT Services
 - Oxford Text Archive, British National Corpus
- Oxford e-Research Centre
 - home for interdisciplinary research, technical R&D, expertise in e-infrastructures in other disciplines
 - BNCweb online service, and OxLangCloud service planned
- Humanities Division
 - users in English, Modern Languages, Linguistics and beyond
- CLARIN European Research Infrastructure Consortium
 - secondment of staff to work on European coordination

CLARIN: some lessons learned

Credit

- Specialist repositories embedded in user communities can offer facilities for doing things with data which generic repositories cannot
- Infrastructures can support bespoke facilities which are not sustainable or connected when run by individual research projects
- Ease of access can be enhanced by offering cross-repository services
- New types of research can be made possible by offering cross-repository services
- Cross-border authentication, authorization and licensing are not insurmountable barriers

Debit

- Research infrastructures are a new type of organization – their activities and their existence are potentially confusing and disruptive for existing institutions and practices
- Reaching and working with an enormous number of potential users in many communities is difficult

Read more...

CLARIN website: <http://www.clarin.eu/>

'CLARIN for beginners'

'Using large text collections for research'

'Text Encoding, text collections and the potential to transform the Humanities'

'Silos or Fishtanks?'

<http://blogs.it.ox.ac.uk/martinw/>

'The Role of CLARIN in Digital Transformations in the Humanities'

Martin Wynne

International Journal of Humanities and Arts Computing 7.1-2 (2013): 89–104

DOI: 10.3366/ijhac.2013.0083

Edinburgh University Press 2013

