

CLARIN

Common Language Resources and Technology Infrastructure



Common Language Resources and Technology Infrastructure

Martin Wynne

Oxford e-Research Centre

martin.wynne@oerc.ox.ac.uk

CLARIN: the mission



The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable for the whole European Humanities communities (and Social Sciences and more widely).

CLARIN is committed to establishing an integrated and interoperable research infrastructure of language resources and tools. It aims to overcome the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure, and therefore enabling e-Research in the Humanities.

CLARIN: the basics



2006: CLARIN formed from existing networks and initiatives and merger of 3 potential submissions to ESFRI

2007: CLARIN selected for ESFRI roadmap in first tranche of social sciences and humanities initiatives

2008: Launch of CLARIN Preparatory Phase Research Infrastructure project, funded by FP7 with 32 partners

2010: 150+ members of CLARIN network

Future:

- FP7 proposal *CLARICLE* submitted 25.11.10
- CLARIN ERIC negotiations start (early 2011)
- End of Preparatory Phase project (mid 2011)
- Construction phase commences (?mid 2011)

Language resources and tools?



The following is a quick example of the sort of research question that language resources and tools can enable:

- When did the word Holocaust start being used to refer to the genocide against the Jews in the Second World War? Has the frequency of its use fluctuated?
- Looking at the *Corpus of Historical American English*, a representative sample of the language over a period of two hundred years, we can see frequencies and examine meanings by reading the occurrences of the word in context.
- Electronic datasets and tools are essential to do this type of research.

CORPUS OF HISTORICAL AMERICAN ENGLISH

400 MILLION WORDS, 1810s-2000s

COMPARE RESULTS: COCA COHA TIME BNC

EMAIL PASSWORD (HELP) LOG IN

DISPLAY

LIST CHART KWIC COMPARE

SEARCH STRING

WORD(S)

COLLOCATES

POS LIST

SECTIONS SHOW

1 IGNORE
SPOKEN
FICTION
MAGAZINE
NEWSPAPER
NON-ACAD

2 IGNORE
SPOKEN
FICTION
MAGAZINE
NEWSPAPER
NON-ACAD

SORTING AND LIMITS

SORTING

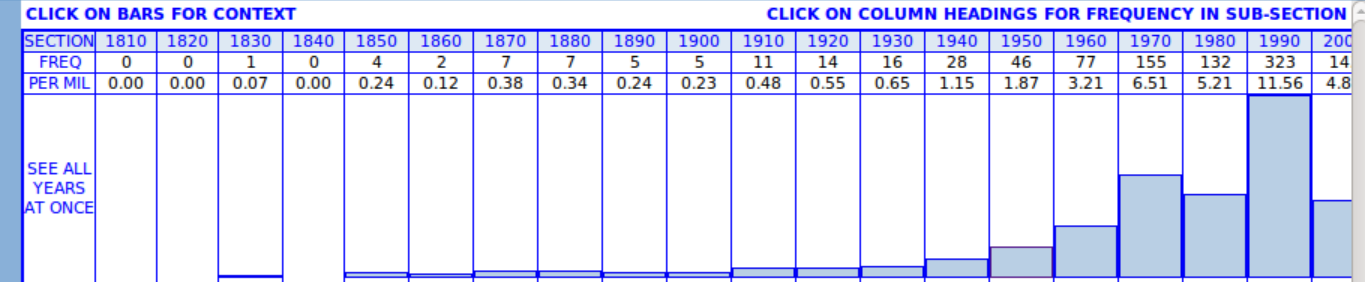
MINIMUM 5

CLICK TO SEE OPTIONS

GROUP BY

DISPLAY

SAVE LISTS



Help / information / contact

SECTION: 1950 (46)

CLICK FOR MORE CONTEXT [?]

[?]

	SECTION	POS	CONTEXT	A	B	C
1	1950	FIC	Play:LittleBlueLight			of martyrs has slumped. Since the bombings and the concentration camps and the Stalingrad holocaust
2	1950	FIC	LoveHonor			who gave all they had and didn't come through. My part in the holocaust is ancient history. Now I wish
3	1951	FIC	ProudNewFlags			. Will that do the trick? " Commander Robb thought it would. XXIV HOLOCAUST SLEEP SIMPLY WOULD r
4	1951	FIC	ProudNewFlags			. " Similar to but infinitely larger than that rocket which had initiated this terrifying holocaust , balls of c
5	1951	MAG	Harpers			the 10,000,000,000 or so stars in the Galaxy. It was out of such a holocaust that the Earth and planets
6	1951	NEWS	NYT-Let			order. Let us hope that this comes about before the outbreak of another senseless holocaust . The Uni
7	1952	FIC	SteamboatGothic			horrible war, she had lost her only daughter and her only granddaughter in a holocaust which, though
8	1952	FIC	SteamboatGothic			the wedded pair had perished so close together and as the result of the same holocaust , such criticism

CORPUS OF HISTORICAL AMERICAN ENGLISH

400 MILLION WORDS, 1810s-2000s

COMPARE RESULTS: COCA COHA TIME BNC

EMAIL

PASSWORD

(HELP) [LOG IN](#)

[Help / information / contact](#)

KEYWORD IN CONTEXT DISPLAY

SECTION: 1960 (77)

CLICK FOR MORE CONTEXT		<input type="checkbox"/>	[?]	SAVE LIST	CHOOSE LIST	CREATE NEW LIST	[?]
1	1960 FIC	Adaptation	A B C	in power we of the Rigel planets will be destroyed, probably in an atomic holocaust . In self-protection we have found it necessary to unite, we Genoese ar			
2	1960 FIC	HuntersOutSpace	A B C	sparkling void. Odin stood transfixed by the sight. Then, momentarily, the holocaust of flame was over. New suns and new worlds drifted calmly, with only			
3	1960 FIC	HuntersOutSpace	A B C	, and Ato gave orders to return fire. For nearly an hour, the holocaust continued. Both ships rocked upon their steady foundations. They were bathed in fl			
4	1960 FIC	Carrington	A B C	whose membershad not been able to save jacket and tails or watered silk from the holocaust , no matter what courage or defiance marked their pasts. Ma			
5	1960 MAG	NatGeog	A B C	comradeships with other men who bore arms. one true good that comes of the holocaust . A man goes back, but he goes alone; his friends are not			
6	1960 NEWS	NYT-Ed	A B C	which Unexpected children view the possibility of fire, flood, earthquake or some similar holocaust that might cause school to close, Any act of God or ma			
7	1961 FIC	GoneFishing	A B C	constructed it as a secure personal retreat in the event of something like a nuclear holocaust . But, in that case, why vacate it now for Barney Chard?			
8	1961 MAG	NewRepublic	A B C	have only two choices - appeasement or war, suicide or surrender, humiliation or holocaust , to be either Red or dead. Each side sees only hard and soft			
9	1961 NEWS	NYT-Ed	A B C	situation escape control. Nevertheless, peril is still real and the shadow of potential holocaust perceptible. Because Mr. Khrushchev now sees our determin			
10	1961 NEWS	NYT-Ed	A B C	will give nothing essential away and are calmly if sadly resigned to the alternative of holocaust if they insist too much. He must also convince our allies th			
11	1961 NEWS	NYT-Ed	A B C	education today embraces the ancient past, outer space and how to survive a nuclear holocaust . The famous son of the Autocrat of the Breakfast Table, t			
12	1961 NF	AmericanRailroad	A B C	collision at Revere, Massachusetts, killed 29 and injured 57 in 1871; a holocaust at Ashtabula, Ohio, cremated dozens as the crack Pacific Express on Vanc			
13	1962 FIC	LettingGo	A B C	day were 34 killed and 102 wounded. Speculations as to why, with this holocaust , today's casualtieswere only slightly higher than yesterday's, were left w			
14	1962 FIC	SevenDaysInMay	A B C	city in the world. Men looked at it, felt the weight of nuclear holocaust lift from their shoulders, and wept. But later the reaction set in.			
15	1962 FIC	ShadeDifference	A B C	kind had come from the Russians, and anything from kidnaping to an immediate nuclear holocaust could have immediate independence for Terrible Terry			
16	1962 FIC	ShadeDifference	A B C	except war; and he sighs again as he contemplates the possibilities of such a holocaust and wonder, what, if anything, a man even in his position can			
17	1962 FIC	ThinRedLine	A B C	day were 34 killed and 102 wounded. Speculations as to why, with this holocaust , today's casualties were only slightly higher than yesterday's, were left v			
18	1962 NEWS	NYT-Let	A B C	take his position, refuses to unequivocally answer the question, " Is a nuclear! holocaust now possible, " This refusal allows him to identify measures which			
19	1962 NEWS	NYT-Let	A B C	to name a specific move by the Soviet Union that would be worth a nuclear holocaust . We must never be trapped into regarding any of the Russian action			
20	1962 NF	HeavenInvention	A B C	be no one to look at them, nor will the pictures themselves escape the holocaust . The middle classes, moreover, are busily acquiring real estate in housin			

http://corpus.byu.edu/coha/x3.asp?r=16&w11=holocaust

FoxyProxy: Disabled

CORPUS OF HISTORICAL AMERICAN ENGLISH

400 MILLION WORDS, 1810s-2000s

COMPARE RESULTS: COCA COHA TIME BNC

EMAIL

PASSWORD

(HELP) [LOG IN](#)

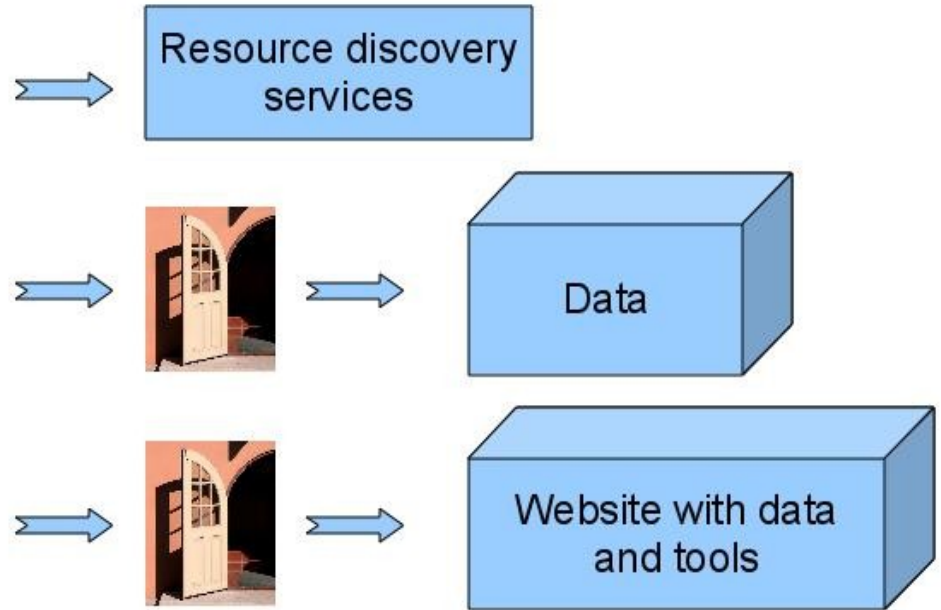
[Help / information / contact](#)

KEYWORD IN CONTEXT DISPLAY

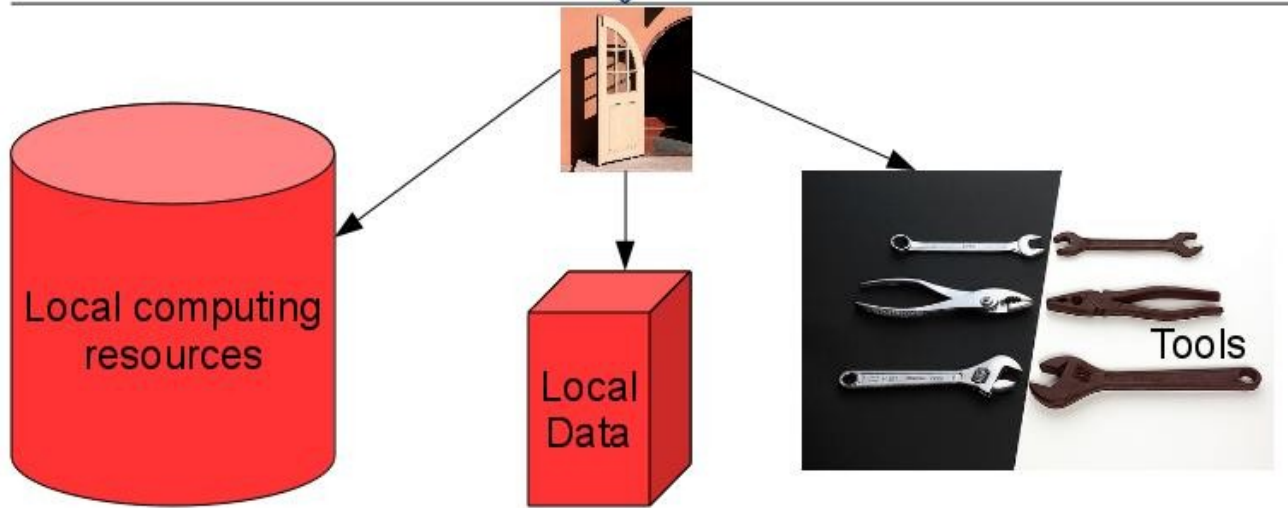
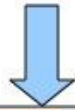
SECTION: 1990 (323)

PAGE: << 1 / 4 >>
SAMPLE: 100 200

CLICK FOR MORE CONTEXT	<input type="checkbox"/>	[?]	SAVE LIST	CHOOSE LIST	CREATE NEW LIST	[?]
1	1990	MAG	Time	A B C	340765	ONE, BY ONE, BY ONE: FACING THE HOLOCAUST # by Judith Miller; Simon & Schuster # 319 pages; \$21.95 # Judith
2	1990	MAG	Time	A B C		pages; \$21.95 # Judith Miller's challenging thesis is that many countries remember the Holocaust in different ways, and from these different perceptions
3	1990	MAG	Time	A B C		in different ways, and from these different perceptions come different distortions of what the Holocaust actually was. A veteran reporter and editor for the
4	1990	MAG	Time	A B C		Nazis asked them to do so; the Soviets steadfastly denigrate the Jewishness of most Holocaust victims; and all too many Americans are turning memories
5	1990	MAG	Time	A B C		of most Holocaust victims; and all too many Americans are turning memories of the Holocaust into a vulgar fund-raising carnival. # All of that is more or
6	1990	MAG	Time	A B C		Auschwitz. The other is Israel, which probably would not exist but for the Holocaust and which still tends to cite the 6 million dead as justification for what
7	1990	MAG	Time	A B C		mainly to Gentile distortions and evasions, she writes about American reactions as though the Holocaust were purely a Jewish question. " While it is now
8	1990	MAG	Time	A B C		States did not do enough to prevent the genocide in Europe... the Holocaust is not an American experience, " she claims. " Americans did not do
9	1990	MAG	Time	A B C		is widely respected in Europe for his blunt acknowledgments of Germany's blame for the Holocaust . Both leaders repeatedly emphasized that their visits
10	1990	MAG	NewRepublic	A B C		uproar that followed. He recalls his role in successfully nominating the Israeli writer and Holocaust survivor Elie Wiesel for the Nobel Peace Prize, and not
11	1990	MAG	NewRepublic	A B C		remained steadfastly silent about the Vatican's conduct during what later became known as the Holocaust . # Hook responded to Adler with an article in T
12	1991	MAG	Time	A B C		laughs where only Roth can find them: a nutty Auschwitz survivor hustling his pornographic Holocaust novel. But elsewhere, readers may find themselv
13	1991	MAG	NewRepublic	A B C		and his anti-Semitic statements, including his stated beliefs that the gas chambers of the Holocaust actually were used for delousing. One group, the Coa
14	1992	FIC	Mov:MedicineMan	A B C		haven't studied history. Everywhere the white man has been there's been a holocaust . It's a way of life with us. Stalin. Hitler... Look
15	1992	FIC	Commentary	A B C		Trois Rivieres was on the trip because of a book she had read on the Holocaust ; the purpose of her coming to Israel was to understand the Jews of whom
16	1992	FIC	TwoGuysFromFuture	A B C		We are here to save art works that will otherwise be lost in the coming holocaust by sending them through a chronoslot to our century in what is, to you
17	1992	FIC	TwoGuysFromFuture	A B C		allowing us to save great art works that otherwise would be destroyed in the coming holocaust . " # " Oh dear. What coming holocaust? " " We're
18	1992	FIC	TwoGuysFromFuture	A B C		be destroyed in the coming holocaust. " # " Oh dear. What coming holocaust ? " " We're not allowed to say, " said Shorty. It
19	1992	FIC	TwoGuysFromFuture	A B C		translate. " # I ignored her. " You know, in this coming holocaust thing. What happens to the people? " # " We're not allowed
20	1992	FIC	TwoGuysFromFuture	A B C		means it's in their catalog. " # " Which means it survives the holocaust , " I said. # " Which means they pick it up at midnight



Single sign on?



Advisory services?



?Cloud computing resources?



The Vision



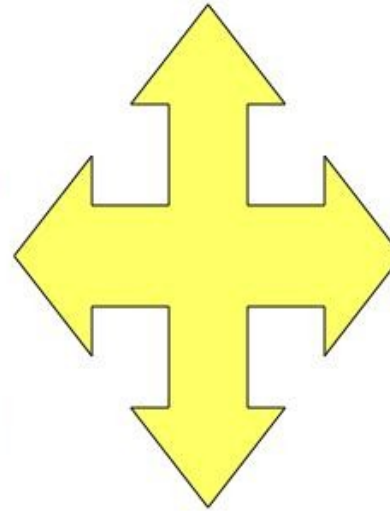
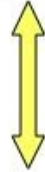
- A researcher in Lancaster from his desktop computer can:
- do a single sign-on with local authentication, and then:
 - search for, find and obtain authorization to use corpora in Oxford, Prague and Berlin
 - select the precise dataset to work on, and save that selection
 - run semantic analysis tools from Budapest and statistical tools from Tübingen over the dataset
 - use computational power from the local or national computing centre where necessary
 - 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



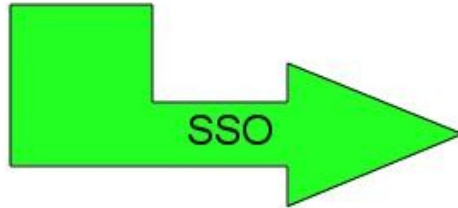
Advisory services



Resource discovery services



Cloud computing resources



Local computing resources

Local Data

Datasets



CLARIN technical work



Promoting collaboration and interoperability between European language resource repositories, particularly in relation to:

- Persistent identifiers
- Component metadata
- Metadata aggregator (Virtual Language Observatory)
- Secure distributed authentication layer
- Trust domain and service provider federation
- Certified centres hosting services with high availability
- Converting tools to operate as web services

See the CLARIN Short Guides at <http://www.clarin.eu/>

Requirements



Service providers

Persistent identifiers

Access & identities

Domain of trust

Web services

Resource discovery

Resource preservation

Monitoring of service availability

Access to compute resources

Requirements



Service providers	<i>existing and emerging</i>
Persistent identifiers	<i>EPIC handle service</i>
Access & identities	<i>national identity federations</i>
Domain of trust	<i>CLARIN service provider federation</i>
Web services	<i>RESTful web services</i>
Resource discovery	<i>our own metadata sets, OAI-PMH</i>
Resource preservation	<i>separate centre policies</i>
Monitoring of service availability	?
Access to compute resources	?

Some outstanding questions



- How can we monitor the availability of services?
- How we deal with access to researchers outside of the SP and IP federations?
- How do we make services interoperable with external service providers, and how do we manage authentication and accounting in these cases?
- What is the business model for advisory services, and who has access to them?
- How can we control access to web services? Do we need to?
- Do we need to establish a CLARIN Virtual Organization to provide identity attributes to users?
- Who does the computing? Who pays for it?

Some outstanding questions



In summary:

- How do we build an infrastructure sufficiently open to allow interoperation with external applications and data services?
- How do reconcile the architecture of the CLARIN infrastructure with the other research computing infrastructures?