

Measuring the Impact of e-Research: Accounting for Disciplinary Differential in Patterns of Diffusion

Jenny Fry and Mike Thelwall

Department of Information Science, Loughborough University, Leicestershire
LE11 3TU, UK.

School of Computing & Information Technology, University of Wolverhampton,
Wulfruna Street, Wolverhampton WV1 1SB, UK.

Email address of corresponding author: j.fry@lboro.ac.uk.

Abstract: This paper presents preliminary findings from a scoping and impact study of the CLARIN research infrastructure project, a long-term pan-European effort to build a virtual, distributed research infrastructure to support researchers of all fields dealing with language based material (text, speech, multimodal media), especially humanities and social sciences (Váradi et al, 2008). The aim of the scoping and impact study is to evaluate the diffusion of the infrastructure-related resources and tools being developed by CLARIN project members. The methodology is taken from the relatively new field of 'webometrics', whereby selected websites and the hyperlinks that connect them are taken to represent intellectual or social dynamics within and between intellectual fields or disciplines. The preliminary findings suggest that the resources and tools developed by the linguistics and language resources research communities have a broad non-specialist audience and that interlinking between the resources and tools themselves is low.

Introduction and background

A core characteristic of social science approaches to e-research has been a concern with research methods. This is reflected not only in the projects and nodes funded by the National Centre for e-Social Science (NCeSS)¹, but also in broadly defined initiatives such as the ESRC Research Methods Network², the ESRC's annual Methods Festival³, and the NCeSS Agenda Setting workshops⁴. This is in contrast with the direction in

¹ NCeSS Website available at: <http://www.ncess.ac.uk/>. Accessed 10 February 2008.

² Details of ESRC Research Methods Programme available at: <http://www.ccsr.ac.uk/methods/>. Accessed 10 February 2008.

³ Details of ESRC Research Methods Festival available: <http://www.ncrm.ac.uk/TandE/other/festival2008/dates.php>. Accessed 10 February 2008.

⁴ A list of forthcoming and past Agenda Setting Workshops available at: <http://www.ncess.ac.uk/>. Accessed 10 February 2008.

which e-research has evolved in the natural, biomedical and engineering sciences where there has been an emphasis on solving computational issues relating to the high-performance processing of data (Hey and Trefethen, 2003). At the same time there is a burgeoning interest in the development of innovative methods for mapping and evaluating the technologies and communities that constitute e-research (Cumplings and Kiesler, 2007; Ackland, Fry and Schroeder, 2007; Ackland and Antony, 2007).

In recent times the term ‘e-Infrastructure’ has started to dominate the language of e-social science governance⁵. Early in the UK e-social science programme it was recognized that sustainability would be a crucial factor in the success of an e-Infrastructure for the social sciences. Sustainability in this context includes being sufficiently embedded both in the research practices of social scientists and in the research policies of institutions such as funding bodies and universities as to ensure a renewable source of resources. Not merely financial resources, but also in terms of the availability of appropriately trained researchers to develop and use an e-Infrastructure together with its related resources, tools and services. In short, diffusion of e-research technologies across a broad range of social science disciplines will be crucial for the success of any e-Infrastructure. Efforts have been made by the NCeSS to generate support for e-social science and encourage users through an intensive focus on education, training and resolving usability issues. Ways in which any subsequent appropriation will be evaluated, what criteria will be used, and the extent to which disciplinary differential should be accounted for is an unresolved issue. These are central questions that any social science method being developed to study the impact of e-research needs to take into account.

As Rogers (1995, p. xvii) has demonstrated “the diffusion of innovations⁶ is essentially a social process in which subjectively perceived information about a new idea is communicated”. By definition an innovation introduces uncertainty into a social system. Rogers (1995) defines uncertainty as unpredictability, lack of structure and lack of information. We know that academic knowledge creation takes place across a series of interconnected social systems. Kuhn’s (1962) controversial work illustrated this and studies in the sociology of science that came after it have shown how particular phases in the development of science social systems impact on knowledge structures (Mullins, 1972). Rogers (1995) argues that diffusion is a kind of social change that alters the structure and functions of a social system. The adoption of an innovation has, therefore, a disruptive influence. To understand what the implications of this are for the appropriation of technological innovation, such as e-research, it is important to understand the dynamics of uncertainty within knowledge structures and how they vary across research communities.

⁵ See the recent call for a special issue of *Social Science Computer Review on e-Social Science*. Available at: <http://www.ncess.ac.uk/news/item/?item=82> Accessed on 21/01/08.

⁶ Rogers (1995, p. xvii) defines innovation as an idea, practice, or object that is perceived as new by an individual or another unit of adoption.

Methodology

The CLARIN consortium consists of 32 partners from 22 countries and is currently working on the preparatory phase. CLARIN intends to build a federation of trusted archive centers that will provide resources and tools through web services with a single sign-on access. Computer-aided language processing is used by a wide variety of intellectual fields across the humanities and social sciences. Current methods and research objectives across these disparate fields have a lot in common with each other. The cost of collecting, digitising and annotating large text or speech corpora, dictionaries or language descriptions is huge, and the creation of tools to manipulate these linguistic data is very demanding in terms of skills and expertise. The challenge for the CLARIN consortium is to turn the existing, fragmented resources and tools into accessible and stable services that any user can share or customize for their own research objectives (Váradi et al, 2008). Within the linguistic and language resources communities there has been a tradition of making resources and tools accessible since the early days of computer networks and these communities were early adopters of the web as a forum for dissemination (Fry, 2006), with many being free and others available at cost. As with other e-research initiatives the aim of the CLARIN consortium is to make the infrastructure accessible via a web-based service. It is appropriate, therefore, that any scoping and impact study would include an evaluation of hyperlinking patterns.

In order to investigate the hyperlinking patterns it was necessary to generate an initial seed set of URLs. For the first phase of the scoping and impact study we asked CLARIN members for URLs that represented the resources and tools that they had developed prior to the start of the project in January 2008. We defined resources as structured data objects, e.g. large text or speech corpora, dictionaries or language descriptions, and tools, e.g. lemmatisers, parsers, summarizers, and information extractors, as the technologies by which those data objects are processed, manipulated and analyzed. The initial response to the request for URLs, approximately seventy-five percent of project members, yielded webpages and websites with varying granularity, e.g. from institutions, academic departments and research groups, to corpora, dictionaries and analytical tools. The dataset was manually cleaned, so that all URLs in the initial seed set were of equal type and represented the finest level of granularity e.g. the resources and tools themselves.

The resulting snapshots of hyperlinking patterns could then be used as a benchmark against which to map the diffusion of resources and tools as the infrastructure project developed and matured. For example, by conducting subsequent periodic crawls and content analyses the rate of diffusion, disciplinary constitution, centrality of the infrastructure, and the emergence of new communities around the infrastructure could be measured.

Inlinks

For each of the seed URLs, a list of pages containing links that point to the URL was generated as follows. The web site domain name of each URL was extracted and a Yahoo! search generated for pages outside the domain but linking to the URL was constructed. This used the advanced search link: command to match pages linking to the

URL and –site: to exclude pages from the same web site (based upon domain name). For example, the URL <http://ucrel.lancs.ac.uk/llwizard.html> originated from the Lancaster University site lanc.ac.uk and so the following search was used to match all pages linking to the URL except pages from Lancaster University.

link:<http://ucrel.lancs.ac.uk/llwizard.html> -site:lancs.ac.uk

The searches constructed as above were submitted to Yahoo! via its Applications Programming Interface, which allows online automatic searches, in May 2008. All pages of results were downloaded and the matching URLs extracted. In cases where there were more than 1,000 results, Yahoo! only returned the first 1,000 and the “query-splitting” technique (Thelwall, 2008) was used to gain additional URLs.

The process described above yielded lists of pages containing a link to each seed URL. The lists would not be exhaustive because search engines do not index all web pages (Lawrence & Giles, 1999) but are likely to be a reasonable substitute (Thelwall, Vaughan, & Björneborn, 2005). One defect with lists of pages obtained in this way is that there may be many pages from a single site, either because the site is duplicated elsewhere or because parts of the site, such as the links, are replicated on many pages (Thelwall, 2002). As a result it is better to count the number of sites containing pages that link to the seed URLs rather than count the individual URLs. Hence the former method was used, as calculated via the LexiURL Searcher software⁷. This produced a list of the number of different web *sites* linking to each seed URL.

Interlinking

In order to investigate interlinking between the CLARIN resources and tools, all the seed URLs were crawled using the research web crawler SocSciBot⁸. In many cases the resource or tool was represented by a single URL and only this URL was downloaded. In other cases, however, the resource or tool occupied its own domain name or a folder within another web site. In these cases, other pages were also downloaded if they were (a) within the resource/tool folder or domain name and (b) linked to by the home page.

All links from the crawled pages were extracted and links within the same web site (site self-links) were discarded, following standard webometrics practice. Next, all links that pointed to resources or tools outside the set of CLARIN resources and tools were discarded as irrelevant to the study. This left too many links to be usefully reproduced on a single network diagram and so an additional step was taken to reduce the information to a manageable quantity. This step was to condense the URLs down to just the domain name part and to remove duplicate links (i.e., pairs of links sharing the same source domain name and sharing the same target domain name). The resulting network interlinking information at the domain name level was useful to convey the degree and patterns of national and international interlinking (see the diagram in appendix I).

⁷ The LexiURL link analysis software is available at: <http://lexiurl.wlv.ac.uk/index.html>

⁸ The SocSciBot link crawler is available at: <http://socscibot.wlv.ac.uk/>

The link analysis method used has two main limitations. First, whilst it produces an accurate representation of web linking, the crawling is not able to be comprehensive in web presences for resources that are not distinct from other parts of the host web site. The method is ideal for resources with their own domain name, less ideal for resources with their own folder or page but problematic for resources that share a page with other information or that are only embedded in a database with other information. The reason is that the calculations rely on the web page of a resource as the unit of content representing the resource, and the domain name as the unit of representation in the diagram. The second problem is that the diagram may not reflect the real connections between the resource owners because resource owners may collaborate without creating hyperlinks between each other. Hence, although the diagram may be taken as an indicator of collaborative links, it is not reasonable to interpret it as a map of collaboration.

Initial findings

In order to carry out a content analysis of the pages that linked to the CLARIN seed URLs a random sample of 100 URLs was automatically generated from the dataset described above. The sample was then coded according to the context of the link, the item being linked to, page type (e.g. blog, wiki, or portal), subject domain, and organizational type.

The two most commonly linked to URLs in the seed set were the British National Corpus (BNC), *929 websites*, which is a 100 million word written and spoken corpus managed by an industrial/academic consortium, and Wavesurfer, *516 websites*, which is an Open Source tool for sound visualization and manipulation developed by the speech technology group at the Royal Institute of Technology Sweden. The remaining top five linked to URLs were the Modern Lithuanian Dictionary, *223 websites*, GATE (General Architecture for Text Engineering), *138 websites*, and NeXTeNS (a Dutch text-to-speech conversion tool), *105 websites*. See appendix II for the number of inlinking websites to all seed URLs.

Fourteen of the URLs in the seed set received no in-links. This does not necessarily mean that these resources or tools have not diffused beyond the original development group, but could reflect issues of limited web representation based on the resources or tools not having unique URLs, but instead being nested in a database of other resources or tools.

There was very little tool/resource to tool/resource linking between those URLs that were categorized as being within the linguistics and language resources community. The majority of the links represented a one-way dependency from outside the CLARIN consortium e.g. the context of the in-link was of the ‘this was/might be useful’ variety embedded within a general audience. For example, the BNC was most frequently linked to from a general language context alongside other examples of English language such as the BBC (British Broadcasting Corporation) website. Where the BNC was linked to from within the linguistics and language resources community it was used as an exemplar of corpus design. In contrast, Wavesurfer was most frequently linked to in the specific context of speech analysis and this was across a variety of domains from open source software and speech technology, to audio research and tracking wildlife. Both the BNC

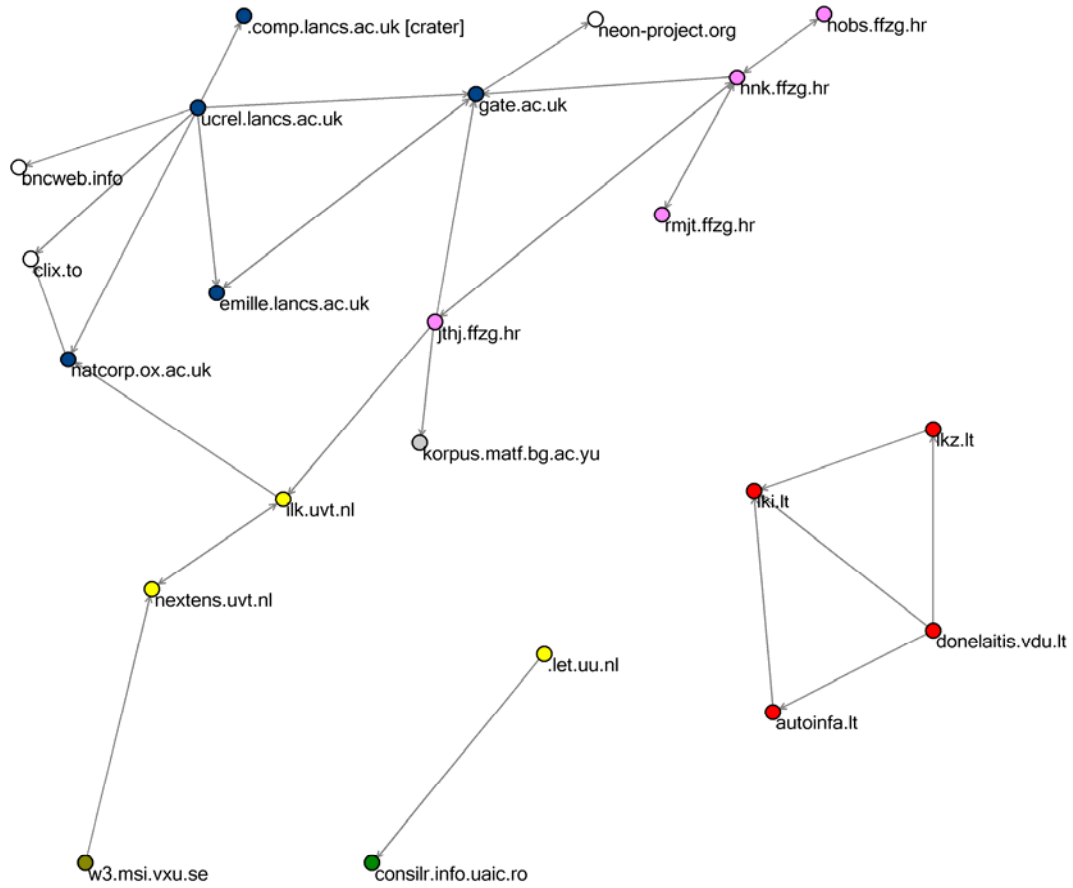
and Wavesurfer are domain independent. They have a broad application, rather than being bound to a domain-specific research problem, and this could be a contributing factor to their comparative popularity within the linguistics and language resources communities and beyond.

The rate of interlinking between resources and tools in the CLARIN seed set was low, with only two percent of the potential number of inlinks being represented in the results.

References

- Ackland, R., Fry, J., and Schroeder, R. (2007) Scoping the Online Visibility of e-Research by Means of e-Research Tools. *Proceedings of the 3rd International e-Social Science Conference*, Ann Arbor, Michigan, 7-9 October 2007.
- Ackland, R., and Antony, R. (2007) Developing e-Research Tools for the Analysis of Large-Scale Web Crawl Data. . *Proceedings of the 3rd International e-Social Science Conference*, Ann Arbor, Michigan, 7-9 October 2007.
- Cummings, J. and Kiesler, S. (2007) Who Works with Whom? Collaborative Tie Strength in Distributed Interdisciplinary Projects. . *Proceedings of the 3rd International e-Social Science Conference*, Ann Arbor, Michigan, 7-9 October 2007.
- Fry, J. (2006) Scholarly Research and Information Practices: A Domain Analytic Approach. *Information Processing and Management*. 42, pp. 299-316.
- Hey, T. and Trefethen, A. (2003) The Data Deluge: An e-Science Perspective *In*: F. Berman, A. Hey, and G. Fox, editors, *Grid Computing - Making the Global Infrastructure a Reality*. John Wiley & Sons. Chapter 36; 809–824.
- Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature*, 400(6740), 107-109.
- Mullins, N.C. (1972). The development of a scientific specialty: The Phage Group and origins of molecular biology. *Minerva* 10(1), pp. 51-82.
- Rogers, E. M. (1995). (4th Ed.) *Diffusion of innovations*. New York, The Free Press.
- Thelwall, M. (2002). Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university web sites. *Journal of American Society for Information Science and Technology*, 53(12), 995-1005.
- Thelwall, M. (2008). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology*, 59(1), 38-50.
- Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, 39, 81-135.
- Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., and Koskenniemi (2008). CLARIN: Common Language Resources and Technology Infrastructure. *Proceedings of the sixth international conference on Language Resources and Evaluation (LREC) 2008*, Marrakesh, Morocco, 28-30 May 2008.

Appendix I – Diagram showing the Interlinking patterns between resources and tools within the CLARIN seed set of URLs



Appendix II – Number of websites linking to the CLARIN seed set of URLs (in descending order of URL most highly linked to)

Title of site or page	Base URL	Pages linking to URL	Sites linking to URL
The British National Corpus (BNC)	http://www.natcorp.ox.ac.uk/	1505	929
Wavesurfer	http://www.speech.kth.se/wavesurfer/	740	516
Modern Lithuanian Dictionary	http://www.autoinfo.lt/webdic	268	223
GATE (General Architecture for Text Engineering)	http://www.gate.ac.uk/	166	138
NeXTeNS Dutch text-to-speech conversion	http://nextens.uvt.nl/	111	105
ISO/TC 37/SC 4 - Language Resources Mgt.	http://www.tc37sc4.org/	91	65
CRATER Multilingual Aligned Annotated Corpus	http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html	59	47
NeOn Toolkit	http://www.neon-project.org/web-content/	58	44
Log likelihood calculator	http://ucrel.lancs.ac.uk/lwizard.html	42	32
Croatian National Corpus	http://hnk.ffzg.hr	32	29
BNC web index	http://clix.to/davidlee00	36	27
Sámi language technology project	http://giellatekno.uit.no/	36	25
Building a Virtual Research Environment for the Humanities	http://bvreh.humanities.ox.ac.uk/	30	25
Corpora of South Asian languages	http://www.emille.lancs.ac.uk/	29	23
Dictionary of place-names (without translation in English)	http://lkz.mch.mii.lt/Vietovardziai/	22	18
Lancaster Corpus of Mandarin Chinese	http://bowland-files.lancs.ac.uk/corplang/lcmc/	22	17
TiMBL Tilburg Memory Based Learning	http://ilk.uvt.nl/timbl	22	17
Database of Lithuanian Office language - without translation in English	http://lkz.mch.mii.lt/Kanceliarinis/	19	16
Dictionary of the Lithuanian Language (0,5m headwords)	http://www.lkz.lt/en/dze.htm	16	13
CLAWS part of speech tagger for English	http://ucrel.lancs.ac.uk/claws/	14	13
Database of language advices - without translation in English	http://www.lki.lt/lki/kkb/kkb.php	14	12
MBT Memory-based tagger-generator	http://ilk.uvt.nl/mbt	14	11
Talbanken05	http://w3.msi.vxu.se/~nivre/research/Talbanken05.html	11	11
BNC web	http://www.bncweb.info/	12	9
Wmatrix corpus analysis and comparison tool	http://ucrel.lancs.ac.uk/wmatrix/	11	8
METER Project	http://www.dcs.shef.ac.uk/nlp/meter/Metercorpus/metercorpus.htm	8	7
Sámi language technology project	http://giellatekno.uit.no/ipk.html	6	6

Speech, Thought, and Writing Presentation	http://bowland-files.lancs.ac.uk/stwp/default.htm	6	5
Sámi language technology project	http://giellatekno.uit.no/fao.html	6	5
Dialect archive (metadata, audio files and JPEG) - without English translation	http://tarmes.mch.mii.lt/	5	5
Lancaster Newsbooks Corpus	http://bowland-files.lancs.ac.uk/newsbooks/	5	4
MaltParser	http://w3.msi.vxu.se/~jha/maltparser/	5	4
Tadpol, Dutch morpho-syntactic tagging	http://ilk.uvt.nl/tadpole	4	4
Open Mind Common Sense Dutch	http://commons.media.mit.edu/nl/	4	3
ConslLR	https://consilr.info.uaic.ro/edtlr/	4	3
The Chaos Project: Robustly Natural Language	http://ai-nlp.info.uniroma2.it/external/chaosproject/	3	3
"Eesti keele keeletehnoloogiline tugi (2006-2010)" (National Programme for Estonian Language Resources)	http://www.keeletehnoloogia.ee/	3	3
Distributed Access Management for Language Resources: Project description	http://www.dam-lr.eu	4	2
Institute for Applied Linguistics, Universitat Pompeu Fabra	http://www.iula.upf.edu/corpus/corpus.htm	4	2
Corpus of contemporary Serbian	http://korpus.matf.bg.ac.yu	3	2
Text Encoding Initiative	http://www.tei-c.org/Tools/index.xml	3	2
Sámi language technology project	http://giellatekno.uit.no/lex.en.html	2	2
Sámi language technology project	http://giellatekno.uit.no/text.en.html	2	2
Croatian Dependency Treebank	http://hobs.ffzg.hr/	2	2
Granska, Swedish Grammar Checker	http://www.csc.kth.se/tcs/projects/granska/	2	2
Corpora of old Lithuanian writings (concordances, original and transcribed texts) - without translation in English	http://www.lki.lt/seniejirastai/	2	2
ConslLR	http://consilr.info.uaic.ro/en/index.php?showpage=060103	2	1
UvT expert collection	http://ilk.uvt.nl/uvt-expert-collection/	2	1
Institute for Applied Linguistics, Universitat Pompeu Fabra	http://www.iula.upf.edu/recurs02uk.htm	2	1
Institute for Applied Linguistics, Universitat Pompeu Fabra	http://www.iula.upf.edu/recurs03uk.htm	2	1
Cyberletteratura	http://ai-nlp.info.uniroma2.it/cyberletteratura	1	1
ABRAXAS (Automating Ontology Learning for the Semantic Web)	http://nlp.shef.ac.uk/abraxas/resources.html	1	1
20th century corpora	http://ucrel.lancs.ac.uk/20thCenturyEnglish/	1	1
UCREL Semantic Analysis System	http://ucrel.lancs.ac.uk/usas/	1	1
Tools for Downloading (Granska Tagger, Inflector, GTA)	http://www.csc.kth.se/tcs/humanlang/tools.html	1	1
Institute for Applied Linguistics, Universitat Pompeu Fabra	http://www.iula.upf.edu/recurs01uk.htm	1	1
Institute for Applied Linguistics, Universitat Pompeu Fabra	http://www.iula.upf.edu/recurs04uk.htm	1	1

Institute for Applied Linguistics, Universitat Pompeu Fabra	http://www.iula.upf.edu/recurs05uk.htm	1	1
NeOn Technologies (links to a number of external URLs)	http://www.neon-project.org/web-content/index.php?option=com_productbook&func=viewcategory&Itemid=99999999&catid=1	1	1
ConsILR - Consortium for the Romanian Language: Resources & Tools	http://consilr.info.uaic.ro/en/index.php?showpage=060101	0	0
Sámi language technology project	http://giellatekno.uit.no/cgi/d-sme.nno.html	0	0
Development Tools, Morphological disambiguation	http://giellatekno.uit.no/doc/ling/docu-disambiguation.html	0	0
Development Tools, Morphological analysis	http://giellatekno.uit.no/doc/tools/tools.html	0	0
Sámi language technology project	http://giellatekno.uit.no/fao.nn.html	0	0
DAM-LR Acceptance test	http://imdi.inl.nl/damlr-test/test.html	0	0
IMIDI Language Resources	http://imdi.inl.nl/imdiportal/BC	0	0
Croatian Language Technologies portal (corpora)	http://jthj.ffzg.hr/corpora.htm	0	0
Croatian Language Technologies portal (tools)	http://jthj.ffzg.hr/tools.htm	0	0
Lancaster Corpus of Mandarin Chinese (available through OTA)	http://ota.ox.ac.uk/headers/2474.xml	0	0
Research programme Computational Linguistic Models and Language Technologies for Croation	http://rmjt.ffzg.hr	0	0
Methods and Tools for automatic grammar extraction	http://stp.lingfil.uu.se/~bea/gramex/home-en.html	0	0
Granska, Swedish Grammar Checker	http://www.csc.kth.se/tcs/projects/granska/index-en.html	0	0
ConsILR	http://www.let.uu.nl/lt4el/index.php?content=tools	0	0