Digital Tools for Linguistics

A Methods Network Working Paper

Computational Linguistics

The field of linguistics draws heavily on computational approaches and is a field where programming skills and high levels of technical expertise are common. Broadly speaking, early work in the field centred on the problems of natural language processing and the development of techniques to enable sophisticated human-machine interaction in a variety of ways. A significant amount of work was carried out to try and make progress in areas such as machine translation, speech recognition and automated question/answer systems (artificial intelligence) and a lot of this work was based on advances made in information theory by figures such as Claude Shannon who as early as 1948 wrote the influential paper, 'A Mathematical Theory of Communication'.

The long-established relationship between Computer Science and Linguistics is indicative of the centrality of digital tools development to the discipline and as a result, there is a prodigious amount of software available to researchers to carry out a wide variety of sometimes quite specific functions. Scholars engaging with linguistics have no choice but to embrace digital tools because their research is often predicated on the detailed and quantitative analysis of large amounts of digitized text. However, one of the unifying conclusions to many articles and essays concerning the subject is that quantitative research methods have to be mediated with qualitative analysis. As Marilyn Deegan states in her rapporteur's report relating to the Methods Network expert seminar on linguistics, 'there is no such thing as bias-free research or intuition-free linguistics'

Stochastic Methods

The early development of stochastic (mathematical and statistical) methods of analysing information have had a profound influence on more recent initiatives to deal with the prodigious amount of information that is now available to researchers, but it is plausibly as much to do with developments outside of the field of linguistics that stochastic techniques are now very firmly back on the agenda again after falling out of favour in the 1960's. With the exponential growth of available data over the World Wide Web and the increasing availability of corpora and treebanks (parsed corpora), it makes logical sense that such methods are now a standard way of dealing with the vast amount of information that is available to researchers and that a great deal of focus is now being put on probability-based models and statistical analysis. An additional external influence on this trend is, of course, the explosion in processing and storage capacity that has occurred in relation to computing since the 1980's, which has allowed the analysis and exploitation of data to take place outside of high specification machine-rooms and away from dedicated servers.

Acknowledgement is also required of advances made within the field however, one very specific example being the work based on the 'hidden Markov model'[1] which uses likelihood and probability to discover unknown values based on related visible parameters. Developed in the late 1960's and early 70's, this had particular relevance to speech recognition techniques and went onto have broad cross disciplinary relevance, particularly in the field of bioinformatics. More generally though, some of the rapid advances in linguistics research over the last two decades can be ascribed to the fact that in many cases, experimental practice can be measured against real-world data and benchmarks can be established that will indicate whether the process that is under scrutiny is actually capable of delivering useful results.

In the case of part of speech (POS) tagging for instance, a subset of the tagged information can be analysed against a manually tagged portion of the data and the automated process can then be evaluated

[1] Rabiner (1989)

to see how accurately it has managed to mimic the very labour-intensive but (theoretically) 100% accurate reference source. The kind of figures that can be obtained from these comparisons are extremely useful in defining what the capabilities of the currently available tools and methods are and provide researchers with tangible goals and challenges to try and aim for in subsequent development phases. In the context of the arts and humanities, this is an unusual working model and reinforces the slightly anomalous status of linguistics research in comparison with the more orthodox interpretative and critical approaches that are normally associated with other arts and humanities subject areas.

By way of example, table 1 refers to the different levels of word accuracy rate that might reasonably be expected from current speech recognition systems within four different contexts – word accuracy being defined as the relatively simple measure of how many words the system misses or confuses when trying to automatically transcribe speech from a variety of sources

Word Accuracy Rate	Source of speech
95%	Closely mounted microphone, quiet room, speaker adapted
75 – 90%	Broadcast News
30 – 90%	Telephone Speech
25 – 45%	Multi-speaker spontaneous speech

Table 1 –Automatic Speech Recognition Capabilities[2]

The figures in table 1 require qualification on a number of points, particularly with reference to the parameters built into the various systems that have provided the performance metrics. As of 2004, Hajič states that the latest speech recognizers could handle vocabularies of 100,000 or so words and that there were examples of research systems which contained one million word vocabularies. Accuracy rates will very much depend on the domain of speech that is being analysed and the relative sizes of the reference vocabularies available to respective systems. As such, the above table is mostly illustrative in its intent.

Corpus Linguistics

In many cases, discussion of tools development within the context of linguistics takes for granted the principle that those tools will have suitable datasets to interact with, process and/or analyze. The key activity that underpins much of this work is the process of corpus construction. The aggregation of machine-readable text from a variety of sources including transcriptions of spoken language, literature, specialist publications, mass circulation periodicals and all manner of other sources, has given rise to the sub-discipline of corpus linguistics, a field of research that goes back to the early 1960's with the creation of the first modern, electronically readable corpus, the Brown Corpus of Standard American English.[3] Forty years later, there is a substantial collection of corpora available in a wide range of specialist and non-specialist areas and many of these collections of text are freely available or available via enquiry or subscription, for researchers to carry out analysis on or to use as reference corpora when comparisons are required with a larger body of language.

[2] Figures from Hajič (2004)
[3] http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown /brown.html

At the Methods Network expert seminar on linguistics in 2005,[4] copious examples of using corpora for research were cited, mostly in the context of using word frequencies to illuminate aspects of particular texts. One of the significant tools mentioned in this forum was Wordsmith[5], developed by Mike Scott, which contains all the functionality one might anticipate from a corpus analysis tool including wordlist creation, concordancing, word clustering, collocation and lemmatization. This software has been in development since 1996 (the current iteration is version 4.0) and is in wide use by a large number of organisations and projects, including the Oxford University Press who use it for their own lexicographical work when building English and foreign language dictionaries - thereby taking advantage of Wordsmith's support for Unicode.

The way that a programme such as Wordsmith might be used - in conjunction with additional subsequent tools and methods - is exemplified by Paul Baker in a paper that analyses linguistic elements in transcriptions from a debate in the House of Commons about fox hunting.[6] By using the keyword list function in Wordsmith, Baker created two separate lists that related to speeches made by the pro and anti-hunt lobbies. By comparing the two lists and the occurrence of keywords in both, Baker was able to examine which words were more 'key' (i.e. more relevant) to one text rather than the other and then was able to discover the context of these words by using concordancing and collocational analysis tools. He cites an interestingly disproportionate use of the word 'criminal' by the pro-hunt lobby and draws out conclusions as to why this mode of speech might suit that particular agenda. He then describes comparing both lists against the FLOB Corpus[7] (1 million words of 1990's British English), in order to discover if both sides of the debate were using words that occur more often than one might expect in 'standard' British English usage. The most significant word highlighted by this stage of analysis turned out to be the word 'cruelty' - used with almost equal frequency by both lobbies and therefore not picked up by an analysis between the sub-texts.

In a further refinement of his analysis, which also usefully illustrates another common exploitation of corpus information, he then used a semantic tagger - in this case the UCREL Semantic Analysis System (USAS) – to examine how words with similar meanings might aggregate together to become significantly disproportionate in their usage.

1	he Bill makes illegal only the perfectly	reasonable	sensible and respectable occupations
2	continuation of hunting. I appeal to all	reasonable	hon. Members to support me in seeki
3	inal law rather than fiddle around in an	absurd	way with this absurd Minister on this
4	rmed roast. The debate has not shown a	rational	analysis of the facts: misplaced co
5	be justified by scientific evidence. The	ridiculous	new clause 13 wrecks it further, and i
6	this matter. Most people with common	sense	will say, "Why don't they reach a dea
7	eds your protection. Mr. Gray: Calm,	sensible	and rational people across Britain a
8	ss. Why not? That would be a logical,	sensible	and coherent approach. As I have to
9	method of control in that time is utterly	illogical	Mr. Gray: My hon. Friend makes an
10	ng-during that time. This ludicrous and	illogical	new clause is the result of a shabby d

Table 2 - (Taken from Baker (2005))

[4] Word Frequency and Keyword Extraction, Methods Network Expert Seminar on Linguistics, Lancaster University, 8 September 2005
[5] http://www.lexically.net/wordsmith/
[6] Baker (2005)
[7] Freiburg Lancaster-Oslo Bergen Corpus http://corp.hum.ou.dk/itwebsite/corpora/corp/page27.html

Certain low frequency words might be represented by a number of synonyms which if collated together might increase their 'keyness' to the point where they are a significant analytical component of the text. Additionally, antonyms are also pertinent to the analysis as words that are posed in direct opposition to each other are still conceptually connected, even if they represent two sides of the (same) argument. As can be seen in Table 2, this can be accommodated by the USAS system, which not only categorises the word 'reasonable' as being usefully related to the concept of being 'rational', but also characterises those words as having an antonymic relationship with words such as 'absurd and 'illogical'.

The corpus that Baker refers to in his paper contains just 129,798 words and is, relatively speaking, tiny in comparison to the types of corpora that researchers use to analyse more general usage of language. It is clear however that in this case, the corpus has been designed to address very specific research questions and it is logical that the size of the corpus that one constructs should be appropriate for, and representative of, the type of information that one is seeking to analyse. In relation to a limited study of textual information where the object of the exercise is to drill down into texts to reveal specific instances of word usage via frequency lists, concordances and citations, the easy-to-use and freely available tool TextSTAT[8] is a very useful starting point for constructing 'home-made' corpora. It even features an advanced query editing function which allows two terms (with wildcard functionality) to be entered with specifiable minimum and maximum word distances between them, thereby allowing collocational analysis of the texts.

Towards the other end of the scale is what is known as 'mega-corpora', also known as second-generation corpora to distinguish them from the collections put together in the 1960's,70's and 80's.[9] The British National Corpus (BNC) is a notable example of this type of undertaking and is 'designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written.'[10] This monumental resource featuring 100 million words provides researchers with a wealth of information extracted from a wide variety of sources,

> The written part of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text.[11]

The spoken component comes from equally diverse sources and all of this information is annotated according to Text Encoding Initiative (TEI) guidelines, which provides contextual information about the content in the form of metadata, as well as allowing data about the structural properties of the text to be included, such as part of speech analysis – carried out by the CLAWS tagging system (the Constituent Likelihood Automatic Word-tagging System) developed by the University of Lancaster.[12]

The BNC can be queried using the original query system, SARA,[13] which provides users with word and phrase search functions, concordancing and collocation features but doesn't allow searches by parts of speech or allow outputs in the form of charts. The updated query system, XAIRA,[14] is a fully functional general purpose XML search engine with full Unicode support that can be used on any corpus of well formed XML documents but has principally been designed for use with the BNC-Baby and the BNC-Sampler corpora.[15] A new edition of the BNC features the whole corpus in XML format which replaces the original SGML annotation and allows for greater interoperability with most software and closer alignment

[8] http://www.niederlandistik.fu-berlin.de/textstat/software-en.html
[9] For comprehensive corpus related info, see David Lee's website http://devoted.to/corpora
[10] http://www.natcorp.ox.ac.uk/corpus/index.xml
[11] http://www.natcorp.ox.ac.uk/corpus/index.xml
[12] http://www.comp.lancs.ac.uk/ucrel/claws/
[13] http://www.natcorp.ox.ac.uk/tools/SARA_search.xml
[14] http://www.oucs.ox.ac.uk/rts/xaira/
[15] BNC-Sampler is a general collection of one million written words, one million spoken. BNC-Baby consists of four one-million word samples from four different genres

with current development methods e.g. the use of DTD's (Document Type Definitions) and XSLT (Extensible Stylesheet Language Transformation).

The BNC supports a wide range of activities that includes reference book publishing, linguistic research, artificial intelligence, natural language processing, English language teaching and in at least one instance, a work of Internet Art that uses the relative word frequencies to rank 86,800 tokens from the BNC in descending order beginning with the most popular word in the corpus (which happens to be 'the') down to the least used as represented in Wordcount (which, for interest, is 'conquistador')[16]. At the scholarly level, tools developed to work with the BNC include the VIEW system[17] (Variation in English Words and Phrases), developed by Mark Davies at Brigham Young University, which offers users an enhanced search interface onto the BNC corpus, allowing - amongst other things – the specification of search terms in and across specific registers (i.e. spoken, academic, poetry, medical, etc). In addition to the very fast searching mechanism facilitated by the use of linked database tables and SQL queries, it also provides the possibility of conducting searches using synonyms and semantically related terms, the latter made possible in conjunction with Wordnet,[18] a semantically-organized lexicon of English, freely available and hosted at Princeton.

Other mega-corpora include COBUILD (a.k.a. the Bank of English) which is defined as a 'monitor corpus' in that it is designed to continue growing in size to reflect the condition of the English language as new words appear and others fall out of favour. As of December 2004, the size of this corpus was reported to have reached 524 million words. Another significant project is the International Corpus of English (ICE) which is a collection of initiatives to document the English language as it is spoken in different countries around the world. ICE-GB (the British component of ICE) is distributed with the retrieval software ICECUP (International Corpus of English Corpus Utility Program) which facilitates the querying of parsed corpora. One further important corpus that is currently in development is the American National Corpus (ANC) which aims to emulate the BNC in terms of its size and scope and will presumably become as influential a research tool as its British counterpart. Links to comprehensive listings of many other corpora including non-English, parsed, historical, subject specific, spoken and specialised examples can be found in the appendix section at the end of this paper.

Knowledge-Based Systems

If the construction of corpora is the main method by which statistical methods can be applied to the various problems of linguistics research; there is clearly a need to also examine a knowledge-based approach, which in practical terms means the attachment of encoded or categorised data and the construction of ontological classification systems to assist with a whole host of research questions relating to grammatical, syntactical, semantic, phonological, morphological, lexical and diachronic (i.e. historical) data issues. This is not to say that the two approaches are necessarily discreet; on the contrary - as has already been shown with reference to Paul Baker's study – research is often a blend of methods and approaches and an accumulation of information through phased querying.

As an example of an approach that is widely used across a range of disciplines, the abovementioned Text Encoding Initiative (TEI) is a rarity that does credit to the conception behind the encoding system that it describes and also highlights the obvious and widespread requirement that exists for a standardised textual markup framework. Unsurprisingly, linguistics is one of the subject areas that has benefited from being able to reference the TEI guidelines, allowing as they do the description of a variety of language features including the separation of elements of spoken discourse and the segmentation of text into sentences, phrases, words, morphemes and graphemes. In addition to the more formal aspects of language analysis, it also allows the description of identities of speakers, the context of textual sources, the

[16] http://www.wordcount.org/about.html
[17] http://view.byu.edu/
[18] http://wordnet.princeton.edu/

inclusion of temporal information, the description of methodological principles … and a great deal of other broadly applicable information that is encapsulated by the definitions available within the framework .[19] However, in response to needs expressed by researchers in the natural language processing and language engineering fields, it was felt that further refinement and a deeper level of encoding was required within the framework of the TEI, and the CES (Corpus Encoding Standard)[20] utilizes the TEI modular DTD and the TEI customization mechanisms to allow for the description of elements that are specifically appropriate to corpus encoding. An XML compliant version called XCES has been in development since 2000[21] and like its predecessor, it is particularly suited to those corpora which shed light on problems associated with language processing and engineering.

A more recent initiative (2003) to develop methods of describing linguistic resources has been proposed by the GOLD (General Ontology for Linguistic Description) community[22], whose objectives are to:

- to promote best practice as suggested by the E-MELD project;
- to encourage data interoperability through the use of ontologies
- to encourage the re-use of software
- to facilitate search across disparate data sets
- to create a forum for data providers and consumers

GOLD work closely with those involved with the E-MELD[23] initiative (who promote best practices in Digital Language Documentation) and OLAC (Open Language Archives Community)[24] who are also engaged with defining and describing linguistic resources using Dublin Core and other metadata frameworks. The GOLD ontology is an attempt to provide a way of mapping information from disparate sources, about different languages, from different theoretical perspectives, onto a common semantic resource. This resource, referred to as a set of 'descriptive profiles' was originally based on the detailed and very useful glossary of terms provided by SIL[25] (initially known as the Summer Institute of Linguistics), which have been substantially added to by members of the GOLD community.

In the wider context of knowledge-based systems, the use of externally defined ontologies, taxonomies and thesauri is also widespread and has already been mentioned in the context of USAS (UCREL Semantic Annotation System) and Wordnet. The latter is an online lexical reference system where nouns, verbs, adjectives and adverbs are organised into synonym sets which additionally link with other sets in various ways and are also incorporated into the Suggested Upper Merged Ontology (SUMO) which refers to itself as 'the largest formal public ontology in existence today'.[26]

The USAS category system (see table 3) currently has 21 top level fields, each of which is assigned with a capital letter, which then subdivides into 232 category labels (designated by numbers) which then allow for further hierarchically structured mapping of discrete concepts. Antonymity of conceptual classifications can be indicated by plus or minus markers within the tags (eg. A5.1+ = good; A5.1- = bad) and multiple possible semantic domains can be incorporated by the use of slash tags (e.g. sportswear may come under the category of both clothing and sport – B5/K5.1).

[19] For the latest (P5) TEI guidelines see: http://www.tei-c.org.uk/release/doc/tei-p5-doc/html/html/
[20] http://www.cs.vassar.edu/CES/
[21] http://www.cs.vassar.edu/XCES/
[22] http://www.linguistics-ontology.org/
[23] http://emeld.org/school/index.html
[24] http://www.language-archives.org/
[25] http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/
[26] http://www.ontologyportal.org/

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

Table 3 - The USAS top level categories (taken from UCREL website)[27]

One further example of a knowledge-based system – and one that refers specifically to historical word forms is the Historical Thesaurus of English.[28] The data in the thesaurus (650,000 meanings in 26 major semantic fields) is taken from the Oxford English Dictionary and includes the first - and where relevant the last - recorded dates of usage for each word along with a total of 29 fields of metadata including broad categorical style and status descriptions and information pertaining to part of speech. Data in this format is demonstrably useful to researchers working in a number of areas but for those involved with the diachronic study of linguistics it is a hugely valuable resource, not only for gauging the relative prominence or insignificance of lexical items from the period defined as OE = Old English (c.700 – 1150 A.D.) through to the present day, but also in disambiguating historical terminology for semantically identical words (see table 4).

Synonyms in category 03.01.01.03.03. ./ 02. . . . -n -(.grandfather)					
Main Heading	Word	POS	Code	Start Date	End Date
Grandparent	ealdafæder	n	OE		
Grandparent	ieldrafæder	n	OE		
Grandparent	eldfather	n	OE	-1460	
	< ealdefæder				
Grandparent	grandsire	n		c 1290	-1876 ai&dl
Grandparent	aiel	n		1377	-1502
Grandparent	belsire	n		1377	-a 1631
Grandparent	grandfather	n		1424	

Table 4 – Synonyms for Grandfather Listed in the Historical Thesaurus of English (taken from Kay (2005))

It is a serious problem for researchers that the further one goes back in time, the more words are prone to be variously and inconsistently spelt and it is increasingly through probabilistic analysis and tagging methods, based on accurate sets of manually created sample data, that large corpora can now be automatically tagged with likely variant spelling definitions using a tool like VARD (Variant Detector Tool)

[27] http://www.comp.lancs.ac.uk/ucrel/usas/
[28] http://www.arts.gla.ac.uk/SESLI/EngLang/thesaur/homepage.htm
Test interface at: http://leo.englang.arts.gla.ac.uk/historicalthesaurus/menu1.html

developed at Lancaster University and described by Dawn Archer at the Methods Network workshop on Historical Text Mining.[29]

Developer Tools and Environments

Spending any time doing research into linguistic tools reveals that an enormous amount of computational work is being carried out in many areas of the discipline, and much of this effort seems to be coming from a community of practitioners who are familiar with various programming languages and capable of working with complex mathematical models. For those looking for an accessible route into this kind of activity, the programming language that is widely referenced as being particularly suited to developing linguistics resources is Perl. As well as having a wealth of pattern-matching and string-handling constructs that complement this kind of research, its structure is also 'comparatively transparent and logical'[30] making it an attractive choice for those new to programming. (Online exercises are available on pages maintained by Paul Bennett at the University of Manchester).[31]

Web pages maintained by Dan Melamed, Assistant Professor of Computing Science at New York University feature links to almost 300 different linguistics software tools[32] - written mostly in Perl by him, his colleagues and his students - all of which are available under a GNU General Public License (GPL). There is an assumption on this site (also in evidence elsewhere) that those wishing to carry out development work in this area will choose to use a UNIX platform.

Another widely used and very influential development environment is the General Architecture for Text Engineering (GATE), supported by the Natural Language Processing Group at the University of Sheffield.[33] Styled as the 'Eclipse of Natural Language Engineering' it consists of three main elements:

- An architecture describing how language processing systems are made up of components.
- A framework (or class library, or SDK), written in Java and tested on Linux, Windows and Solaris.
- A graphical development environment built on the framework.[34]

Extensive documentation is available on the website including information about support for a diverse number of languages using the Java Multilingual Unicode Text Toolkit (JMUTT).[35]

For researchers and developers working in the field of open source natural language processing software, the OpenNLP website acts as a central reference point for project listings and also hosts the OpenMaxent NLP machine learning package. A variety of java-based NLP tools use this resource for processes such as sentence detection, tokenization, pos-tagging, chunking and parsing, named-entity detection, and coreference, and there are links to dozens of other tools, API's and models written in a variety of languages including Perl, Python and C++.[36]

Field Linguistics

There is also an abundance of resources available for those involved with using and developing 'field linguistics' tools, which generally means that the resources are geared towards the capture, annotation and analysis of spoken data. This practice has particular value for groups associated with the recording and

[29] http://ucrel.lancs.ac.uk/events/htm06/
[30] http://www.llc.manchester.ac.uk/SubjectAreas/LinguisticsEnglishLanguage/Staff/PaulBennett/PerlforTextProcessing/
[31] See previous reference
[32] http://www.cs.nyu.edu/~melamed/software.html
[33] http://gate.ac.uk/
[34] http://gate.ac.uk/documentation.html
[35] http://gate.ac.uk/demos/unicode/index.html
[36] http://opennlp.sourceforge.net/projects.html

preservation of endangered languages and one of the prominent projects in this area is the Hans Rausing Endangered Languages Project.[37] A particularly useful site with links to tools for fieldwork is maintained by Stanford University,[38] which also contains links to a wealth of resources maintained by the Max Planck Institute for Psycholinguistics, based in the Netherlands[39]. There are descriptions of widely used tools such as Shoebox[40], a data management and analysis tool for field linguists, as well as more general format conversion, annotation and corpus management tools, all of which both enhance the choice and add to the plethora of software systems available on the Web for linguistics research.

Conclusion

Towards the end of June 2006, the Digital Tools Summit in Linguistics was held in association with the Summer meeting of the Linguistic Society of America[41] and it would seem symptomatic of the need felt by both the humanities community more widely (members of which gathered for the similarly named Summit on Digital Tools for the Humanities in September 2005[42]) and the linguistics community in particular, to have meetings of this type to discuss viable futures for tools development and exploitation. A report based on the earlier humanities summit has been widely distributed and has informed a great deal of recent discussion about methods, strategies and funding approaches and it is reasonable to expect that the more recent meeting might also serve that purpose within the more bounded field of linguistics. A working paper written by Jeff Good of the Max Planck Institute was circulated before the linguistics summit which contains an enormous amount of detailed and interesting information about what he refers to as the 'ecology of documentary and descriptive linguistic work'.[43] As well as specific references to a number of resources and a discussion about the *concept* of tools, he also usefully analyses terms like 'interoperability' and 'community' which are often used as catch-all terms but can also obscure more sophisticated readings of the issues intrinsic to such ideas.

The central point of his working paper however is that any focus on tools and resources development (and indeed the use of them) needs to take place with an understanding of what else is available, so that everyone working in the field is pulling in the same direction, thereby maximising effort and minimising duplication. In broad terms, this is also one of the points that came out of the recent Historical Text Mining workshop at Lancaster. One of the questions posed towards the end of the session was 'How can we get the communities to talk to each other?', the communities in this case being all those groups identified as having a stake in linguistics research methods, namely: literary studies, computer science, information science, humanities, philosophy and so on. As a rallying call in the humanities, it is (unfortunately) rather a commonplace conclusion. What is perhaps less common is the prospect hovering beyond the realm of academia of large corporations and governmental agencies having strong vested interests in finding more efficient ways of extracting meaning and making sense of data collections that grow ever larger as storage capacity inexorably increases. As a discipline dealing in methods that are potentially very much in demand, there may be a case for remaining quite optimistic about the future of linguistics tools development.

Neil Grindley
Senior Project Officer
Methods Network
First draft - November 2006
Version control – 9 July 2007

[37] http://www.hrelp.org/
[38] http://www.stanford.edu/dept/linguistics/fieldwork/info/back.html
[39] http://www.mpi.nl/tools/
[40] http://www.sil.org/computing/shoebox/index.html
[41] http://www.ipsr.ku.edu/DTSL/links.html
[42] http://www.iath.virginia.edu/dtsummit/announcement.html
[43] http://emeld.org/workshop/2006/papers/ToolEcology-1.pdf

References

Baker, P., *"The Question is, how cruel is it?"Keywords, foxhunting and the House of Commons,* Word Frequency and Keyword Extraction, Methods Network Expert Seminar on Linguistics, Lancaster University, 8 September 2005
http://www.methodsnetwork.ac.uk/redist/pdf/baker.pdf

Hajič, J., *Linguistics Meets Exact Sciences*, in Schreibman, S., Siemens, R., Unsworth, J., (eds), A companion to Digital Humanities, (pp. 79 - 87), 2004

Kay, C., *Issues for Historical and Regional Corpora: First Catch Your Word*, Word Frequency and Keyword Extraction, Methods Network Expert Seminar on Linguistics, Lancaster University, 8 September 2005
http://www.methodsnetwork.ac.uk/redist/pdf/kay.pdf

McEnery, T., Xiao, R., *Character Encoding in Corpus Construction*, in Wynne, M. (ed), Developing Linguistic Corpora *A Guide to Good Practice*, AHDS Literature, Languages and Linguistics, Oxbow Books, 2005

Rabiner, L., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, vol.77, no.2, February 1989

Shannon, C., *A Mathematical Theory of Communication*, Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, 1948

APPENDIX

LINKS, RESOURCES AND FURTHER REFERENCES

The following resources are grouped under general headings but are otherwise in no specific order. They represent web resources visited in the course of researching this paper.

Contents

Organisations and Departments
Projects
Articles and Reports
Corpora and Treebanks
Resources and Listings
Tools and Listings
Developer Resources
Dictionaries, Ontologies, Glossaries

Organisations and Departments

Linguistic Society of America
http://www.lsadc.org/info/ling-fields-comp.cfm
Page on 'Computers and Language'

Association for Computational Linguistics
http://www.aclweb.org/

Association for Literary and Linguistic Computing
http://www.allc.org/index.html

UK Academic Departments
http://tangra.si.umich.edu/clair/universe-rk/html/u/db/acl/html/ACADEMIC/EUROPE/UK/
Linguistics and Natural Language Processing Departments in the UK

London Linguistics Circle
http://www.londonling.ucl.ac.uk/
Joint page of London related academic linguistics departments

GOLD
http://www.linguistics-ontology.org/
The GOLD Community is interested best-practice encoding of linguistic data and the mapping of legacy data onto a common ontology

E-MELD
http://emeld.org/school/index.html
Promoting best practices for digitizing language data.

Berkeley
http://linguistics.berkeley.edu/~jcgood/bifocal/
Berkeley Initiative for Computer Assisted Linguistics

Projects

Hans Rausing Endangered Languages Project
http://www.hrelp.org/
SOAS initiative which focuses on the documentation and preservation of language

DoBeS project
http://www.mpi.nl/DOBES/
Documentation of endangered languages

Open NLP site
http://opennlp.sourceforge.net/
open source natural language processing project list

Revere Project
http://www.comp.lancs.ac.uk/computing/research/cseg/projects/revere/
Reverse engineering of requirements to support business process change

Articles and Reports

Digital Tools Summit in Linguistics
http://www.ipsr.ku.edu/DTSL/callforpapers.html
University of Kansas conference, June 2006

Working paper on Linguistics tools
http://emeld.org/workshop/2006/papers/ToolEcology-1.pdf
Jeff Good – Max Planck Institute – advance reading for the Tools Summit

Data Mining
http://www.dlib.org/dlib/march06/cohen/03cohen.html
Article on textual data mining of Large Digital Collections

Stephen Bird
ftp://ftp.cis.upenn.edu/pub/sb/papers/cp-intro/cp-intro.pdf
Introduction to computational phonology

Stephen Bird long list
http://www.ldc.upenn.edu/sb/home/publications.html#bk94
A variety of useful articles based on Stephen Bird's linguistics research

Evaluating Linguistic Tools
http://linguistics.berkeley.edu/~jcgood/bifocal/SoftwareQuestions.html
Article on evaluating linguistics tools

Conceptual Basis for Unicode
http://acharya.iitm.ac.in/multi_sys/unicode/uni.php?topic=concept_uni
Good description of the objectives and value of Unicode

RLG article on Automatic Extraction of Keywords
http://www.rlg.org/en/page.php?Page_ID=17068&Printable=1&Article_ID=991
Deegan, Short, Archer, McEnery, Baker and Rayson

German Fraktur Fonts
http://www.morscher.com/3r/fonts/fraktur.htm
Outlines the difficulty of OCR with old fonts.

Discussion slides from HTM workshop
http://ucrel.lancs.ac.uk/events/htm06/DiscussionHTM06.pdf
Outlines concerns and issues that face the community.

NLP History
http://nltk.sourceforge.net/tutorial/introduction/section-x65.html
Brief History of Natural Language Processing

Re-emergence of Statistical Methods
http://www.vinartus.net/spa/00a.pdf
Revival of stochastic methods

Hidden Markov Model (HMM)
http://www.informatik.uni-bremen.de/agki/www/ik98/prog/kursunterlagen/t2/node4.html
The prediction of unknown results based on related visible parameters

HMM for Bioinformatics
http://www.csse.monash.edu.au/~lloyd/tildeMML/Structured/HMM.html
Use of Hidden Markov Models in bioinformatics

Querying Linguistic Databases
http://projects.ldc.upenn.edu/QLDB/
Links to articles on the formation of linguistic queries

Using XML/Xpath
http://projects.ldc.upenn.edu/QLDB/cassidy-lrec.pdf
A use case analysis of using XQuery as an annotation query language

Explanation of Text Clustering
http://www2.parc.com/istl/projects/ia/sg-clustering.html
Non-scientific theoretical example of word clustering analysis

Common Ontology for Linguistic Concepts
http://www.emeld.org/documents/knowtech_paper.pdf
EMELD paper which discusses TEI, CES and XML

Corpora and Treebanks

List of Corpora
http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html
UCREL list of available corpora

Treebanks
http://www.ifi.unizh.ch/CL/volk/treebank_course/
Short Course at the University of Zurich with resources

SCRIBE
http://www.phon.ucl.ac.uk/resource/scribe/

UCL dept of phonetics and Linguistics corpus of spoken British English

AHDS respository
http://ota.ahds.ac.uk/search/search.perl?search=QUICK&misc=corpus
38 varied corpora

Corpora Listings Page
http://devoted.to/corpora
A comprehensive and very useful site relating to Corpus Linguistics

BNC
http://www.natcorp.ox.ac.uk/
British National Corpus

Stanford Listings
http://nlp.stanford.edu/links/statnlp.html
Very good listing of corpora, treebanks and tools

Resources and Listings

SAMPA at UCL
http://www.phon.ucl.ac.uk/home/sampa/
Computer Readable phonetic alphabet

XSLT description page
http://emeld.org/school/classroom/stylesheet/xsl-help3.html
What is XSLT and XPath?

OLAC Linguistic Data Type Vocabulary
http://www.language-archives.org/REC/type-20060406.html
Ties in with Dublin Core and is about how to describe a linguistic resource

SIL links
http://www.sil.org/linguistics/computing.html
Linguistics Computing Resources on the Internet

TAPOR
http://taporware.mcmaster.ca/
Online text analysis portal

Linguistic Resources – University of Pennsylvania
http://www.ldc.upenn.edu/annotation/
Detailed list from end of 2001 of many tools and resources

Thai University linguistics department
http://pioneer.chula.ac.th/~awirote/ling/corpuslst.htm
Detailed listings of linguistics resources

Max Planck Institute for Evolutionary Anthropology
http://lingweb.eva.mpg.de/fieldtools/tools.htm
Lists tools for linguistics fieldwork

Unicode information
http://www.unicode.org/standard/principles.html
Introduction and technical information for Unicode encoding

Historical Thesaurus of English (HTE)
http://libra.englang.arts.gla.ac.uk/historicalthesaurus/menu1.html
Web interface for querying the thesaurus

Thesaurus of Old English
http://libra.englang.arts.gla.ac.uk/oethesaurus/
The offshoot project from HTE

Treebank Example
http://projects.ldc.upenn.edu/QLDB/data/wsj_0003.prd
With parts of speech tagged in trees

Non-major language resources
http://www.bmanuel.org/index.html
Covers the lesser studied languages and lists resources for them

CES
http://www.cs.vassar.edu/CES/
Corpus Encoding Standard

Perl Programming tutorials
http://www.llc.manchester.ac.uk/SubjectAreas/LinguisticsEnglishLanguage/Staff/PaulBennett/PerlforTextProcessing/
Paul Bennett at the University of Manchester

Tools and Listings

SYSTRAN
http://www.systransoft.com/index.html
Machine Translation developed and used by the European Commission

EXMARaLDA
http://www1.uni-hamburg.de/exmaralda/index-en.html
Hamburg developed system for computer assisted transcription and annotation of spoken language.

SIL Fieldworks
http://www.sil.org/computing/fieldworks/flex/overview.html
A suite of software tools to help language teams manage language and cultural data, with support for complex scripts.

Treeform Syntax Tree Drawing Software
http://www.ece.ubc.ca/%7Edonaldd/treeform.htm
Tool for analysing syntax

Max Planck Institute for Psycholinguistics, Nijmegan, Netherlands
http://www.mpi.nl/tools/
Short list of tools developed at the MPI

MPI Leipzig Glossing Tools
http://www.eva.mpg.de/lingua/files/morpheme.html
For adding annotations between lines (interlinear) in a standardised way

Wordcount
http://www.wordcount.org/main.php
Find word frequencies based on the BNC

Stanford Linguistics Fieldwork page
http://www.stanford.edu/dept/linguistics/fieldwork/info/back.html
A short list of tools for linguistics fieldwork

Shoebox
http://linguistics.berkeley.edu/~jcgood/bifocal/ShoeboxRev.html
A review of the Shoebox tool

Natural Language Toolkit (NLTK-Lite),
http://nltk.sourceforge.net/
Suite of tools – open source project

OpenMaxent
http://maxent.sourceforge.net/about.html
Advanced environment for java-based tools development

Open NLP Tools API
http://opennlp.sourceforge.net/api/index.html
List of parts of the Open NLP tool suite

Open NLP links
http://opennlp.sourceforge.net/links.html
Links to tools including translation software

Wordsmith
http://www.lexically.net/wordsmith/
Mike Scott's site for Wordsmith

WMatrix
http://www.comp.lancs.ac.uk/ucrel/wmatrix/
Software tool and web interface for USAS and CLAWS

CLAWS
http://www.comp.lancs.ac.uk/ucrel/claws/
(the Constituent Likelihood Automatic Word-tagging System)

Nora Text Mining Tool
http://www.noraproject.org/
Text Mining project

Next Generation Tools at UCL
http://www.ucl.ac.uk/english-usage/projects/next-gen/index.htm
Refers to a whole cycle experimental querying environment for parsed corpora

Developer Resources

UIMA
http://www.research.ibm.com/UIMA/
Unstructured Information Management Architecture

Corpus Rule Transformation Notation
http://crouton.sourceforge.net/
A small but fairly complete functional programming language for querying and transforming parsed manuscripts
Perl Programming Language
http://linguistlist.org/issues/14/14-1536.html
Book review of Perl Programming Handbook

Parsers
http://www.nyu.edu/pages/linguistics/parsers.html
Overview of parsing process

Dictionaries, Ontologies, Glossaries

Concise Oxford Companion to the English Language (online)
http://www.oxfordreference.com/views/BROWSE.html?subject=s8&book=t29
Oxford Reference Online version – accessible via ATHENS login

Glossary of Linguistic Terms
http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/
comprehensive and clear glossary of all terms

Suggested Upper Merged Ontology (SUMO)
http://www.ontologyportal.org/
General purpose upper level ontology