

Language Resources Primer

Mark Liberman and Ron Cole

Introduction

What are "Language Resources"?
Who and What are Language Resources for?
Data-centered Research Communities
Existing Models for Creation and Distribution
of Resources
Intellectual Property Issues
Standards Issues

Introduction

For participants in the Skamania Language Resources Workshop, sponsored by the Interactive Systems Program of NSF's IRIS division, we've prepared a short annotated tour of web sites offering background information on the workshop topic.

The goal is to ensure that the workshop participants, with their diverse backgrounds, are all familiar to some extent with a common core of information about the present state of language resources and their use. We are not trying for complete coverage, though we welcome additions and corrections. All opinions expressed, and any factual mistakes, are of course the responsibility of the authors.

You should read this document in whatever way is most comfortable for you. However, we've designed it to work best if you go through the whole thing once rapidly, without following the hyperlinks, and then take come back and take a more leisurely tour, with as many side trips as you have time and inclination for.

What are "Language Resources"?

Here is a definition from the recent call for the First International Conference On Language Resources And Evaluation. It's a little wordy but covers the ground well:

The term language resources (LR) refers to sets of language data and descriptions in machine readable form, used specifically for building, improving or evaluating natural language and speech algorithms or systems, and in general, as core resources for the software localization and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users. Examples of linguistic resources are written and spoken corpora, computational lexicons, grammars, terminology databases, basic software tools for the acquisition, preparation, collection, management, customization and use of these and other resources.

The relevance of evaluation in Language Engineering is increasingly recognized. This involves assessment of the state-of-the-art for a given technology, measuring the progress achieved within a program, comparing different approaches to a given problem and choosing the best solution, knowing its advantages and drawbacks, assessment of the availability of technologies for a given application, and finally product benchmarking.

It accompanies research and development in Human Language Technologies, and has driven important advances in the recent past in various aspects of both written and spoken language processing. Although the evaluation paradigm has been studied and used in large national and international programs, including the US ARPA HLT program, EU Language Engineering projects, the Francophone Aupelf-Uref program and others, particularly in the localization industry (LISA and LRC), it is still subject to substantial unresolved basic research problems.

We support this broad definition of language resources, which encompasses all corpora, lexicons, tools and standards that specifically support research, development and evaluation of human language technologies. Moreover, we support a broad definition of human language technologies, which includes the range of technologies needed to enable interactive systems using natural communication skills. Examples of the diverse natural communication skills that might be used in interactive systems are lip reading, facial movements and expressions, gestures and "body language," handwriting, American Sign Language, and Braille. In this context, we might extend the definition of language resources to overlap with the resources and technologies needed to interpret images.

Who and What are Language Resources For?

Our primary focus is on resources needed for "building, improving or evaluating natural language and speech algorithms or systems." The Survey of the State of the Art in Human Language Technology gives an overview of this area of engineering. Some additional information can also be found in the on-line first draft of the EAGLES Handbook on Spoken Language Systems.

It is important to recognize that there are also scientific and educational issues at stake. For instance, there are hardly any accessible tools that can be used for creating and experimenting with language technologies in the curricula of high schools, community colleges, and even most four-year undergraduate institutions. Initial efforts to develop and disseminate such tools are in their infancy.

The varied humanistic and scientific disciplines that deal with human language and its use have a mutually-beneficial relationship with the development of linguistic technology. In addition to the obvious connections with linguistics, psychology, and so forth, specific applications in fields such as history and anthropology also warrant special consideration.

Data-centered Research Communities

These are groups of researchers tied together by a shared body of data that provides raw material for developing and testing scientific theories or basic technological capabilities. There are examples in a variety of fields, from seismology, genetics and child language development, to astrophysics and particle physics. These data-centered scientific communities generally arise where data are expensive to get, or definitive once acquired, or both.

There is a long tradition of such communities in the humanities, where scholars have traditionally clustered around particular collections of documents. This tradition has become electronic in the case of ancient Greek, Latin, Egyptian and Middle English texts, French and English literature, Sumerian texts, and many other things as well. The oldest effort to collect and disseminate such digital resources for scholars in the humanities is the Oxford Text Archive, which has been in operation for more than 20 years.

It is worth mentioning as a special case the existence of government-funded institutes dedicated to documenting national languages. Typically, these institutions are responsible for particular work products such as dictionaries. There is a trend towards making some of the accumulated resources available (at least for remote search and retrieval) outside the responsible institution, but it remains more common for them to be closely held. Examples can be found for Czech, French, German, Korean, Swedish, and quite a few other countries/languages.

As you can see from following some of the links above (and those below as well), these activities are very diverse in origin, structure and result.

Some efforts are unfunded, volunteer and bottom-up, while others are elaborately organized with government, private or mixed funding. In some cases the resulting data is in the public domain, while in other cases it is copyrighted by one or more owners. Orthogonally to the question of ownership, the data may be distributed for free, or licensed by sale, or restricted to members of a consortium. We'll examine these distinctions in more detail below.

Data-centered Human Language Research

Over the past decade, a data-centered mode of organization has become increasingly important in the development of linguistic technologies.

This is partly because most contemporary approaches to "language engineering" rely on analysis of large corpora, reference to large lexicons, and use of other resources expensive enough that it is logical to share them. Another motivation has been the development of evaluation-based research management in speech technology, textual information retrieval, message understanding, and other areas as well.

This approach was pioneered by a series of DARPA projects whose current embodiment is the Human Language Systems program. The approach has been taken up to one extent or another by others, for example AUPELF-UREF, an international organization of Francophone universities and research groups, which has launched a set of "Concerted Research Actions" (ARC) in the areas of spoken dictation, spoken dialogue, and speech synthesis. These actions are explicitly inspired by the example of DARPA's human language technology efforts, but also present some interesting innovations.

These examples have been initiated by government agencies, but there are also some similar initiatives that have arisen in the private sector. One notable example is Unipen, which grew out of an ad hoc committee of the International Association for Pattern Recognition. Unipen created a large body of data and tools for open competition in the area of on-line handwriting recognition, entirely by volunteer efforts. More than forty institutions donated samples totaling more than five million characters from more than 2200 writers.

As this on-line inquiry (about an upcoming Unipen evaluation) indicates, such unfunded and volunteer benchmarking efforts can be hard to keep on track over extended periods of time.

Perhaps for this reason, and also perhaps because formal benchmarks are not the only appropriate mode of self-organization in science and technology, data-centered research communities are more often focused on sharing basic resources. A crucial step is often the process of forging agreement about what these resources should be like. A good example of this process in progress is the Discourse Resource Initiative, which provides a clearinghouse for documents, corpora and software in support of discourse research, and a focus for discussions to define the content and form of various types of discourse annotation.

Another (somewhat more top-down) example of the same type is the ToBI effort to develop standards for intonational transcription.

Another interesting experiment is the Kamusi Internet Living Swahili Dictionary project at Yale, an attempt to create a dictionary by collective effort.

Brian MacWhinney's CHILDES (Child Language Data Exchange) project is one of the most mature (also influential and effective) examples of a non-evaluation-based data-centered

community of language researchers. It combines standards and tools with an ever-growing body of contributed data.

Existing Models for Creation and Distribution of resources

Bottom-up volunteer initiatives

An important role has been played by initiatives such as the ACL Data Collection Initiative, the European Corpus Initiative, and Project Gutenberg. These begin as volunteer efforts, though they may acquire funding and even a semblance of formal organization later on. Such initiatives tend to depend on impetus from a few key people (sometimes only one person), and continue to thrive as long as they are willing or able to maintain the effort. Of the three cited above, which all effectively started in the late 1980's, only Project Gutenberg has continued as an active enterprise, due to the single-minded dedication of Michael Hart.

Another on-going (apparently) volunteer effort is Jim Breen's Edict Japanese dictionary project at Monash.

Enterprises of this kind are especially important in seeding new areas and in creating free or low-cost resources, which can be very widely distributed and used.

Individual sponsored projects

A closely-related category is the traditional individual-investigator sponsored project, which may produce, by plan or as a by-product, some resources that come to be distributed more-or-less widely. The granddaddy of these is the Brown Corpus, which was produced with U.S. Government funding by Kucera and Francis some three decades ago at Brown University, and is still in widespread use. Other very widely-used resources of this type include linguistic data and software from the WordNet project at Princeton, the XTag project at Penn, the JUMAN morphological analyzer for Japanese at Kyoto, and Brill's POS tagger, originally created at Penn. The CHILDES project is another example of this type.

Unfortunately, for each case like Brown or WordNet, there are dozens of resource-creating sponsored projects where the underlying resources have never gone outside the PI's lab or office, although of course many scientific or scholarly publications may have been based on them. In many relevant disciplines (e.g. sociolinguistics), this is still the rule rather than the exception. Aside from investigator retentiveness, there are often real reasons for this: it takes a lot of work (and a certain amount of cash money) to prepare resources for publication and distribution. There are often also problems with privacy, subject releases, copyright, and so on. If the initial hurdles are overcome and distribution begins, there have often been problems with long-term maintenance of the data and of the distribution process.

Archives and distribution centers

Some help in this area has been provided for decades by humanities-oriented institutions like the Oxford Text Archive, and the International Computer Archive of Modern and Medieval English (ICAME). More recently, technology-oriented institutions for managing the process of sharing linguistic resources have been created in the USA---the Linguistic Data Consortium (LDC)---and in Europe---the European Language Resources Association (ELRA).

LDC and ELRA were seeded with public funds, but are carrying on as self-supporting enterprises. They have played an extremely important role in fostering and managing the distribution of resources that would otherwise have remained more closely held, and also in promoting a general movement in the direction of sharing pre-competitive resources. Their legal structures and general operating models are quite different, although their goals are very similar. A considerable amount of information is available on their web sites, and in the LDC's report to the ISGW, so we will not discuss them more extensively here.

Efforts to create a counterpart institution in Japan have not yet succeeded. TELRI is an embryonic effort to extend the concept to Eastern Europe.

An earlier attempt to create such an organization for lexical resources, the Consortium for Lexical Research, failed due to lack of money---public funding ran out and financial self-sufficiency had not been established.

Diverse mixed cases
With the development of convenient methods of CD-ROM publication, and (especially!) the growth of the Internet, it has become much easier to set up distribution for language-related tools and data. As the cost/performance of computers improves, the "market" for such resources is increasing rapidly. As a result, an increasing number of institutions distribute language resources that they have created, or that they themselves have gotten from others. Quite a few examples have been cited earlier in this document.

We will give a only a small (but we hope representative) additional sample of the diverse universe of such sites. Some are all-volunteer, though most enjoy some mixture of public and private funding; some are particular to a specific resource or tool, while others may offer quite a wide variety; some resources are in the public domain, while others are copyrighted and licensed either implicitly or explicitly; some are free, while others are available for purchase or to members of some sort of subscriber group.

The Sleator/Temperley link-grammar parser at CMU;

the CHASEN

Japanese morphological analyzer at Nara;

the OGI CSLU corpora
and toolkit;

the Australian
Indigenous Languages Virtual Library;

the Computing
Resources site of the Summer Institute of Linguistics;

the IFCSS CNapps
site for Chinese computing resources;

and many others.

Closed government-led partnerships

In Europe and in Japan, governments have commonly put together ad hoc partnerships of companies and universities that cost-share (usually by in-kind contributions) in the production of resources that are then owned by the members of the joint venture. Notable examples include the British National Corpus and the Electronic Dictionary Research (EDR) project (here is a useful summary of the EDR project).

Distribution of the results outside of the initial group typically involves a time delay, a fairly high price (the EDR dictionaries cost about \$8,000 to universities, and about \$90,000 to companies, for a research-only license), and sometimes more-or-less complete exclusion (the British National Corpus is still unavailable to American researchers, as are many other European corpora funded with national or EU grants).

Some recent EC-funded projects such as MULTEXT and MULTEXT-East are producing open resources, and ELRA is now open to American members (though not all of its databases are available).

However, it is often difficult to tell from the available documentation just what the conditions on distribution will be. You may be interested in trying this exercise on a list of current "Language Engineering Resource" projects funded under the European Community's Telematics Applications Programme, and lists of projects funded earlier under LRE 1 and LRE 2. For instance, what will be the availability of the SPEECHDAT databases to American researchers?

There are many large, publicly funded projects around the world, specifically oriented towards the production of language resources, whose results are effectively unavailable to researchers outside of the institutions involved in the project. There are certainly dozens of such cases, and perhaps more than a hundred. If we were to count projects where language resources are produced as a by-product of other research, the number would be many times as large. The good news is that there are now hundreds of resource items that are available to the research community at large on some terms---ten years ago, there was essentially nothing other than the Brown corpus in this category.

Commercial sources

The "language industries" have of course progressed far enough that quite a few resources useful for research are available through ordinary commercial channels. LINGSOFT and inXight are good examples of tools and lexical resources in the textual area for sale to researchers and developers, while Entropic is a good example of speech-related tools for sale. The Agora languages marketplace, the dialect.com Japanese language site, and the World Language Resources site are examples more oriented to the general public, but where researchers and developers might still find some useful things.

In addition, an enormous range and volume of newspapers, magazines and broadcast materials are now available on line, some free and other on a subscription or pay-per-use basis. Commercial CD-ROM text collections are also increasingly common. This is a key step forward. However, it is important to keep in mind that in most cases, IPR restrictions stand in the way of the collection, transformation and sharing of these resources that would be required for most types of research use. The current trend is for such problems to get worse rather than better.

In some cases, industrial researchers collaborate with academics to produce freely available resources, as in the case of Jeffrey's Japanese Dictionary, which builds on Jim Breen's Edict work. In other cases, companies give tools or other resources away freely, such as an experimental Xerox part-of-speech tagger, or Microsoft's excellent SDK, which offers speech recognition, synthesis and NLP tools. However, of course the commonest arrangement is for resources to be offered for sale. This should become a more and more common state of affairs as the technologies and their markets mature.

Governmental information sources

Finally, many governments (and some transnational organizations) provide access to materials that may be useful for language researchers, even though these resources are generally presented for informational purposes. As usual, the Internet is not the only mechanism of distribution, but its role is increasing. A sampling of sites of this type includes FedWorld, the National Technical Information Service, [and others to come].

The case of the Foreign Broadcast

Information Service is worth taking a special look at. These extensive translations and summaries of news stories from around the world have been

prepared by the CIA for many years, for use within the U.S. government. Some time ago, NTIS began making FBIS materials available on the Internet. This service was very popular, but raised a chorus of protest from the publishers and broadcasters whose stories were featured. Now the online FBIS service itself is available only within the government, while NTIS offers a similar, royalty-paying product called World News Connection for a fee of about \$100/month. An FBIS CD-ROM is available within the U.S. government only.

Some discussions of various FBIS issues (mainly funding problems) from the Federation of American Scientists may be of interest.

Intellectual Property Rights issues

As in the FBIS case, provision of language resources for research often runs up against IPR issues. Several U.S. Government projects have been threatened with legal action over IPR violations involved in handing out electronic copies of copyrighted text. Embarrassingly, one of these cases involved the state news agency of a foreign government that was being urged, at the time, to act more vigorously to block piracy of U.S. software and movies.

Using existing dictionaries effectively in the creation of lexical resources for research purposes has been especially full of problems and pitfalls, and has prevented a lot of useful stuff from being distributed more widely. In the case of dictionary resources for Chinese, for instance, much material that once was generally available has been withdrawn after protest from copyright holders, or because of commercialization plans on the part of compilers.

For general background information about copyright and other IPR issues as they affect electronic data resources, see Bitlaw, or The Copyright Website.

One particularly important and controversial issue today is the IPR status of facts and electronic collections of facts. Two key aspects of this are the European Database Directive and the proposed WIPO Database Treaty, which create new sui generis property rights for collections of facts. Here is Bitlaw's calm take on the subject. Very much more negative views come from the AAAS, the Government Printing Office, and the Electronic Frontier Foundation, among others.

James Love argues that such laws would make sports scores into property, so that they they could not be reported or perhaps even discussed electronically without payment of royalties. Even without new laws endowing databases with sui generis property rights, there have

been some recent U.S. court rulings in area of sports statistics that lead in the same direction.

Already, re-distribution of the video portion of broadcast news is made much more difficult by the embedded IPR issues associated with stock footage or clips from other sources. Typically, the broadcaster has the right to use this material, but not the right to give anyone else the right to use it. New sui generis property rights invested in facts, by laws like the European Database Directive and WIPO, would create very similar problems for nearly all journalistic material, and possibly even for some lists and tables derived from such sources.

For some general browsing on IPR issues from a public-interest perspective, take a look at the CPT's Intellectual Property Page, or the UPD page on copyright and neighboring rights. For a much more property-oriented perspective, take a look at infringaTek's internet monitoring service. There is a new industry based on patrolling the internet for IPR violations!

Another key issue is the standing of author's rights as opposed to copyright. Historically, British and American law have treated copyright as a property right, which can be sold outright, while mainland Europe has treated creators' rights as a basic human right, and thus inalienable, subject only to a sort of rental. This has posed practical problems for obtaining rights for corpus distribution in continental Europe, since (for instance) it seems to require dealing with all the individual journalists who have written for a newspaper or news agency. This issues has gained considerable force with the development of new electronic media; the International Federation of Journalists has launched its 1997 Author's Rights Campaign to promulgate the continental perspective. More advocacy on the topic comes from the Creator's Copyright Coalition.

A recent U.S. court decision came down in favor of publishers' rights to electronic re-publication of articles by free-lance writers.

For the most part, IPR concerns come down to worry about loss of revenue, even in the case of research and educational use where collection of substantial fees may seem unlikely. Publishers (or authors) worry that they may entirely lose control of their property (or their creation), once it starts to circulate in digital form.

It's worth noting that for audio and video materials and images, licensing fees tend to be higher than for text, and the tolerance for "fair use" much narrower. Here is an example of standard broadcast-industry ideas

about licensing

fees for distribution of news and public-affairs programming: thus for world-wide educational use, the price is \$20 per second of material used. At this cost (still much less than the charge to commercial customers), even allowing the 50% discount for use of ten minutes or more, licensing the 300 hours of DARPA Hub-4 material would have cost \$10.8 million.

Indeed, this was where last year's negotiations with broadcasters started for licensing DARPA Hub-4 material. In the end, the license fees were at most \$100 per hour, and in many cases were entirely waived. In order to accomplish this, the broadcasters had to be persuaded that their material would not simply escape into cyberspace.

One innovative and deservedly influential (though controversial!) approach to promoting free distribution of software through licensing restrictions is the Free Software Foundation's "copyleft" approach.

This mechanism has been used to govern the distribution of several important software packages, including GNU Emacs, the GNU C Library, and the gcc and gcc++ compilers. We are not aware of many language resources that are distributed under agreements of this general type. The licenses for WordNet and Edict

are notable examples that follow a similar style (although WordNet lacks, and Edict modifies, the controversial FSF stipulations about the openness of derived works). These two resources have been very widely used, in part because of their intrinsic quality, and in part because they are 'freeware' (though not in the public domain).

It is typical for researchers to restrict more closely the distribution of resources that they develop, including those that are government funded. In some cases the researchers may aim to commercialize the results themselves; in other cases, they want to use an "industrial affiliates" or similar program to derive some on-going support for their work. Neither of these approaches is forbidden by law or custom, and they do have significant advantages in fostering commercial development and in diversifying on-going project support. However, the FSF and similar models certainly also seem to have some advantages in establishing easily-available shared resources for the research community, and should be seriously considered, at least for cases in which external forces (such as licensing from publishers or broadcasters) do not forbid it. In general, we feel that a careful discussion of the range of IPR approaches for publicly-funded language resources is overdue.

For instance, it appears that the results of the EuroWordNet project will (as a practical matter) be rather expensive, rather than free, on the grounds that private property (from Novell and various dictionary publishers) is to be involved in its construction. This is exactly the sort of thing that the FSF licensing approach forbids, and thus is a good example of why such provisions are significant and consequential. We do not take a stand on these issues, except to urge fuller discussion and wider understanding of them among affected researchers.

By the way, a close reading of the EuroWordNet licensing conditions is a useful exercise. You should try to predict what research access will really be like, for various classes of users.

Standards issues

The National Institute of Standards and Technology(NIST) Spoken Language Processing Group and Natural Language Processing & Information Retrieval Group continue to play a key role in creating and implementing the common tasks that define many (U.S.) government-sponsored programs in the area of language technology.

The Expert Advisory

Group on Language Engineering Standards (EAGLES) is an effort by the European community to define a general framework.

Organizations such as the International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques for Speech Input/Output (COCOSDA) play a useful role in international information exchange. For instance, COCOSDA provided the forum for international discussion of the POLYPHONE framework for telephone speech database collection.

Among the many diverse ad-hoc standards initiatives that impinge on language technology, we mention here the development of various APIs (including SAPI, SRAPI, TSAPI and TAPI); the Text Encoding Initiative (TEI); and the Unicode consortium.

There are also standards activities involved in relevant pieces of the European COST structure, especially actions such 249 ("continuous speech recognition over the telephone"), 250 ("speaker recognition in telephony"), 219 ("future telecommunication and teleinformatics facilities for disabled and elderly people"), and 258 ("naturalness of synthetic speech").

A key role is also played by information and tools from diverse volunteer sources, such as Andrew Hunt's comp.speech FAQ or Lance Norskog's SoX software for the conversion of audio file formats.

Other useful or interesting links

Among the many examples we might pick to round out the discussion, we would like to cite the Edinburgh HCRC Language Technology Group helpdesk,

the Human Languages Page
, the ACL NLP/CL
Universe, the Verb
Sea, a gateway to information on facial
animation, the OLEADA
project (TIPSTER technology providing assistance to language teachers),
and John Hiese's
Akkadian site.