The language belongs to the People!

Cornelis H.A. Koster
Computing Science Institute
University of Nijmegen
The Netherlands
kees@cs.kun.nl

Stefan Gradmann
Regionales Rechenzentrum der
Universität Hamburg
Germany
stefan.gradmann@rrz.uni-hamburg.de

Abstract

We defend the thesis that basic linguistic resources, especially when developed largely with public money, should be made freely and openly available to the public. We draw an analogy with software world, which is well ahead of the linguistic world in this respect.

> Os baldios são do Povo!
> Portugal, around 1980.

1 Introduction

For every natural language, computer-readable basic resources like grammars and linguistic lexica are increasingly needed, for educational, scientific and economical reasons. Especially in countries with a strong and modern economy, enormous efforts have already been invested in developing such resources, but often without common purpose and synergy. Property rights tend to be yealously guarded by industrial and academical developers alike. Enormous amounts of monetary support are wasted on projects that perforce must start by reproducing the work of others, since they can't use the previous results, and whose results in their turn either remain hidden or just evaporate. We appear not to be standing on the shoulders of our predecessors but rather on their toes.

The present situation of linguistic resources resembles that of software systems and applications in the eighties: on the one hand many people try in vain to make some money from their intellectual work, hampering free exchange and collaboration, on the other hand monopolies threaten to corner the market. The resemblance is large enough to suggest the same solution: *basic linguistic resources which have been developed with public support should be freely and openly available to the public* at essentially no cost for their use, not even a small one, for commercial as well as scientific purposes, and in the form of open source texts.

We are not proposing this in order to criticize or deprecate organizations like LREC/ELDA, ANC, BNC etc., who are doing their best to make linguistic resources available at a reasonable price, but to foster discussion and to ask for more generosity and openhandedness from developers, copyright owners and funding agencies alike.

2 Basic Linguistic Resources

First of all we should make clear that it is not our contention that *all* linguistic resources should be free, and that different opinions are possible about what constitutes a basic linguistic resource (BLR). By a BLR we mean a collection of data about an individual natural language, which have been empirically derived by observing the language itself or its use, and software directly based on these data, which together form an essential starting point for linguistic applications: word lists, frequency lists, tag sets, lexica, taggers, grammars and parsers, corpora, treebanks and ontologies. What resources should be considered basic is in fact both temporally and culturally determined, with older and simpler resources coming to be considered as basic when more sophisticated ones are developed, as long as they are of sufficiently general interest and do not overmuch depend on a single linguistic theory or ideology.

Second, our proposal holds most strongly for those languages in which the whole superstructure of linguistic resources still has to be built up, that is to say, all languages with the exception of English and a few others. But even those privileged languages could benefit from more openness.

Thirdly, the modalities of BLR distribution should also be open, without fees or bureaucratic hurdles: in a generic form, as textfiles and tables or as open source code, which are freely downloadable, without using proprietary formalisms and not tied to the functionality of any commercial product. A closed database which can be queried may be very useful to research and education, but it cannot be incorporated as a component in more general linguistic applications.

The concept of BLR may be further clarified by an analogy to scientific publications. Considering the relation between journal articles and monographs, in most of the "hard" sciences it could be maintained that articles, being still part of a discussion context, and thus in a pre-aggregated state, need to circulate as flexibly and informally as possible, whereas monographs, which represent stabilized knowledge and are not subject to communicative dynamics any more, can be given the status of consumer goods without doing any harm to scientific exchange of information.

BLR would then have to be considered as raw, pre-aggregated (or at most semi-aggregated) material,

very much like journal articles – as opposed to fully developed and stabilized products that can be sold on the market of scientific information.

3 Three Case Studies

We shall highlight by small case studies the predicaments faced by linguistic academic researchers, research consortia, traditional publishers and application developers, in order to prove that the prevailing restrictive attitude is counterproductive.

Let us take as an example the situation of a language about which we are all (more or less) equally ignorant: the Icelandic language. We apologize to the real Icelanders, please do not feel insulted.

3.1 The University of Reykjavik

We at UofR have years of experience in researching the Icelandic language, and we are proud to say that we are world leaders in this area. We have put our results into a vast database. Recently, its contents was put onto a CD, which for scientific and educational applications can be bought at a nominal fee. Industrial users are referred to the Transfer Bureau of the University of Reykjavik, which will charge enough money to keep our research going.

Our Transfer Bureau is actively seeking rich industrial partners, by waiting for them to find the website. No, we don't do maintenance, because that would not be academic.

3.2 Bergens Aktienbolaget

We are the largest publishing house on the island of Iceland. In the last 150 years we have collected all words of Icelandic that have ever been written, and put them into our corporate database. All dictionaries, terminologies, ontologies and spelling checkers of Icelandic are based on this. Anybody using our data either pays or infringes on our copyright.

We are now looking for a technique to charge royalties for every written Icelandic word. Only a small amount, just enough to achieve our goal of 20% growth per annum. For the spoken word we are awaiting improvements in chip technology.

3.3 The EUROLONG project

Our project has been subsidized by the EC for 30 years. We have developed corpora, lexica and grammars for all major and minor European languages, including Icelandic and all its dialects. If you want to use our data you have only the following to do:

- become a member of the club
- use our formalism, whose manual consists of 1200 pages
- buy a supercomputer running PROLOG with at least 256 processors.

For scientific applications you pay only 256 Euro per database, and you get a reduction of 10% on every dozen. For commercial applications: price upon request.

4 The Problem

Anyone who has invested a large period of his life in developing software or data collections is convinced that these must be worth a lot of money, since they represent a large amount of work, and reproducing them may cost the same amount of work. Such is human nature. On the other hand, their development has often wholly been paid for by public money, and the employers (some university or other organisation) would like to get a return on their investment.

- academics often have difficulties paying even a small amount of money (and have less qualms putting PhD students paid by some outside agency to Cinderella work)
- any academic undertaking with a practical side threatens to have commercial value at some later time, and therefore can not safely be based on resources that are free only for scientific use
- as a compromise, an important infrastructure is set up to guard intellectual properties and control distribution to academic users for a nominal fee, in the hope that commercial users will pay the full value. But in the mean time this organization may eat up all the income gathered.

It should be clear that it is impossible to ever earn back the (public) investment made in linguistic resources, but even the costs of distribution and incasso may tax the budgets of academics, while industries feel compelled to invest in developing their own proprietary resources rather than base themselves on resources made by academics who may want their pound of flesh later.

5 Free Software

The solution for the quandaries in which the linguistic community finds itself was pioneered in the area of computer software: to consider linguistic resources as Free Software as defined by the Free Software Foundation: *software that comes with permission for anyone to use, copy, and distribute* [it], *either verbatim or with modifications, either gratis or for a fee. In particular, this means that source code must be available.* Well-known examples are given by operating systems like Linux or free BSD, by GNU and the Open Software Foundation, but primarily by thousands of individuals who, working purely alone or in noisome groups and cabals, contribute to, and profit from, the open source community.

What motivates some people to give away the fruits of their work, freely and without restriction? Primarily, the urge to create something, and the logic of donating it to posterity in order to be remembered by others. More arguments concerning the motivations for participating in such an economy of gift can be found in the writings of (Eric S. Raymond (1999)).

But the world is not only comprized of idealistic individuals. Bergens Aktienbolaget, which has skillfully exploited public funding in developing its own

resources, may well ask: how can we survive economically if our main assets are thrown into the public domain? Software companies, especially in capitalistic America, have in fact invented many business models where the main product is (also) free software.

Linux can be downloaded for free from RedHat, but it is also possible to enter into a service contract. A good example is MySQL AB, developer of the world's most popular open source database. Its strategy is to have at any point in time a commercial version, with older releases being free, as well as pre-releases of newer versions. In this way, they are providing the community with well-tried software for free, and at the same inviting the open source community to help in testing, debugging and improving their next release.

There are so many more examples: the Apache webserver is open source, Postgress offers free licensing, OpenOffice.org is free to use and distribute. Of course, some of the Open Source activity in this world is aimed at providing an alternative to Microsoft (to the relief of many customers, including the German Parliament which chose Linux technology for use within the Bundestag), but it also reflects good business thinking: distributing free software costs next to nothing, and it attracts a lot of business. The more important commercial customers do not wish to dabble in gratis software, they are quite willing to pay for maintenance, consultation, training and support. And a critical open source community keeps you on your toes.

The different forms of license which are available to provide a stable legal basis for open source software (such as GPL, LGPL and BSD) have been discussed at length by (Eric S. Raymond (1999)).

In the scientific publishing world, the Open Access movement is similarly gaining momentum; the major scientific funding organizations have expressed their support for sharing the fruits of scientific work rather than exposing it on the markets (the declarations of (Berlin declaration (2003)) and (Paris declaration (2004))).

6   A Modest Proposal

Why can't the same be applied to linguistic resources? There is a clear precedent: the Princeton WordNet (Miller et al (1993)), which in spite of its initial limitations and imperfections has enabled huge amounts of further research and experimentation, because it freed researchers from the necessity of first building their own ontology of English. As an other example, two parsers for English ((Sleator and Temperley (1993)) and (Sekine (1996))) which are publicly available have been used by many other researchers.

Free basic linguistic resources should be developed (with public money as necessary) for those languages for which they are lacking, and made generously available to everyone who wishes to make some productive use of them, using open formats. Users should be invited to learn from them, use them, and to join in their enhancement and completion, engendering a spirit of collaboration and synergy conducive to the best results. Competition should not take place on the basic linguistic resources, but on the *advanced* linguistic resources that developers can add to or derive from the BLR.

A free software approach is of benefit to all parties concerned:

- bringing such basic linguistic resources under an open license has many advantages for the resource developers:
    - developers are invited to improve and extend the work of their predecessors, rather that repeat it, and to share the results
    - electronic distribution of the most recent version is very cheap
    - the quality of the resources is continuously and voluntarily being scrutinized and improved
    - the maintenance problem is solved.

- not only academic researchers are liberated from the problems of using proprietary resources, but also commercial developers of applications:
    - they get a head start from using existing resources
    - they can afford to experiment at an early stage without large initial investments
    - at a relatively low cost they can ensure improvement and maintenance of the linguistic resources used
    - commercial use is possible without fuzz and (unpredictable) expenses.

- the funding agencies and their policy makers can act more effectively while spending much less money:
    - stimulating a few early developments has maximal effect
    - only incremental funding is needed for upgrading
    - they can leave the control of quality to the users
    - no need for Top-Down planning, no pressure from powerful lobbies
    - more synergy between users, more leverage of funding
    - de-facto standardization arises naturally, on a competitive basis, no need for costly top-down and often premature standardization efforts.

7 An Example: The AGFL System

The AGFL parser generator system for natural languages was developed at the University of Nijmegen in the Netherlands between 1991 and 1995 with support from the Netherlands Organisation for Scientific Research NWO. It threatened to wither into obscurity for lack of further funding, like so many other academic results, even though it was used successfully in linguistic research projects like TOSCA, (Oostdijk (1998)), AMAZON (Coppen (2002)) and ICE-GB (Oostdijk (2000)). After being consolidated with support from the NLnet Foundation, its copyright was transferred to the Free Software Foundation in 2001 (Koster and Verbruggen (2002)), making the AGFL system the first parser generator for Natural Languages under the GPL/LGPL. Everybody may freely use the system and distribute any parsers generated with it, even commercially (that's why some parts fall under the LGPL).

Since then, the AGFL system has been freely distributed via the AGFL website, along with the EP4IR grammar and lexicon of English and the DUP4IR grammar and lexicon of Dutch, which were developed for Information Retrieval Applications in the European Projects DORO and PEKING (see the PEKING website). Similar grammars for Russian, German, Spanish and Modern Standard Arabic are under development by other groups.

For most users (not every user of language resources is a linguist) the most visible results are the parsers and the grammars and lexica from which they are generated. They go somewhat beyond the usual in containing subcategorization information for verbs, adjectives and nouns, but in this and other respects they are neither complete nor perfectly correct and consistent (like version 1.5 of WordNet); it is hoped that the growing user community will help in extending and refining them, adding more lexical and grammatical sophistication like frequencies, collocations, idioms and connecting the lexicon to WordNet-like ontologies. Furthermore, the grammars which were developed for generating phrasal document representations in Information Retrieval (Koster (2004)) may with some adaptation be used for question-answering applications and natural language interfaces.

The refinement process of such language resources is practically endless, since language is a fractal: the complexity of any part of its description (e.g. expressions of time and space, modality, semantical aspects) can be made as great as we wish.

The free and open availability of the AGFL system offers a solution to our problems of distribution and maintenance. It also means that the work on the AGFL project was not in vain, but may be of lasting value to the world at large.

8 Conclusion

Language was already invented long before us, and we can be thankful to our ancestors that they chose to put it in the public domain. All we can do is to investigate aspects of it, to discover its structure and collect data about its lexicon and its use. Nobody has a monopoly on language, on knowledge about it or on its uses. Nobody can claim any part of it as his own property because he found it first – it belongs to its natives, to its speakers and authors, its thinkers and poets.

For this reason alone all basic language resources properly belong in the public domain. But what we have tried to make clear is that there are also perfectly businesslike arguments showing that it is in the personal and mercurial interest of academic developers of linguistic resources, their bosses, the funding agencies and even the language industry that basic linguistic resources should be freely and openly available.

References

AGFL website: `http:www.cs.kun.nl/agfl`

Berlin declaration (2003): `www.zim.mpg.de/openaccess-berlin/berlindeclaration.html`

P.A. Coppen (2002), "Het geheim van de oude dame: de Nijmeegse parser AMAZON". Nederlandse Taalkunde, 7 (4), 312-334

Free Software Foundation: `http://www.fsf.org/`

C.H.A. Koster and E. Verbruggen (2002), "The AGFL Grammar Work Lab". Proceedings FREENIX/Usenix 2002, pp 13-18.

C.H.A. Koster (2004), "Head/Modifier Frames for Information Retrieval". Proceedings *Computational Linguistics and Intelligent Text Processing* (CICLing-2004). Springer LNCS 2945, pg 420-432.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, *Five Papers on WordNet*. Technical report, Cognitive Science Laboratory, Princeton University, August 1993.

NLnet Foundation: `www.nlnet.nl/`

Netherlands Organisation for Scientific Research NWO: `www.nwo.nl/`

N. Oostdijk (1998): "The TOSCA parsing system reviewed". In J. Carroll (ed.), Proceedings of the LREC workshop The Evaluation of Parsing Systems, Granada, May 1998. 63-69.

N. Oostdijk (2000): "Corpus-based linguistics at a cross-roads". In: English Studies, Vol. 81 No. 2: 127-141.

Paris declaration (2004): `www.itu.int/wsis/documents/`

PEKING website: `http:www.cs.kun.nl/peking`

E.S. Raymond (1999), "The Cathedral & the Bazaar", Beijing [etc.], O'Reilly.

S. Sekine (1996), "Manual of Apple Pie Parser", `citeseer.nj.nec.com/sekine96manual.html`

D. D. Sleator and D. Temperley (1993), "Parsing English with a link grammar", Third International Workshop on Parsing Technologies, `citeseer.nj.nec.com/sleator91parsing.html`