An archive for all of Europe: the TRACTOR initiative

Martin Wynne
Centre for Corpus Linguistics,
Department of English,
University of Birmingham,
Birmingham
UK – B15 2TT
martin@clg.bham.ac.uk

Abstract

TRACTOR is the TELRI Research Archive of Computational Tools and Resources. It features monolingual, bilingual, and multilingual corpora and lexicons in a wide variety of languages, as well as tools for language processing. TRACTOR is a key element of TELRI II, a pan-European alliance of focal national language technology institutions with the emphasis on Central and Eastern European and NIS countries. TRACTOR hopes to complement other archives by providing a service for languages and users who are currently under-represented in existing archives. TRACTOR's unique strength lies in the amount of resources provided by centres in Central and Eastern Europe, and its role at the hub of a network of resource creation, standardisation and distribution which links the EU and non-EU European research communities. The TRACTOR User Community brings together resource providers, academic users and industrial users in an ongoing relationship, which is designed to foster the emergence of joint research projects in language engineering.

The TRACTOR philosophy is to accept deposits of resources in any format, and to distribute them in the form in which they are received (with small changes if possible such as additional documentation, and putting a browsable version or sample online.) In addition, certain standards are recommended and help is offered to providers who wish to make their resources conformant with the standards. This lack of standardisation is not simply a pragmatic measure in the face of problems of heterogeneity, but is based on a profound scepticism towards current resource standardisation practice.

In the future, TRACTOR aims to build up particularly parallel corpora and tools for processing and extracting meaning from such resources.

1 What is TRACTOR?

TRACTOR (www.tractor.de) is the TELRI Research Archive of Computational Tools and Resources. It features monolingual, bilingual, and multilingual corpora and lexicons in a wide variety of languages, currently including Belarus, Bulgarian, Cornish, Croatian, Czech, Dutch, English, Estonian, French, German, Hungarian, Italian, Latvian, Lithuanian, Romanian, Russian, Serbian, Slovenian, Swedish, Turkish, Ukrainian and Uzbek. TRACTOR is also building up a portfolio of tools for language processing. While the original aim was to pool the resources of TELRI partners, TRACTOR also serves other institutions by making their resources and tools available for the wider community. The borders

of the TRACTOR User Community now stretch beyond Europe, with members in Africa, Asia, and America.

1.1 What is TELRI?

TRACTOR is a key element of TELRI II. The EC Concerted Action TELRI II is a pan-European alliance of currently 28 focal national language technology institutions with the emphasis on Central and Eastern European and NIS countries.
TELRI II´s primary objectives are:

- To strengthen the pan-European infrastructure for the multilingual language research and development community;
- To collect, promote, and make available monolingual and multilingual language resources and tools for the extraction of language data and linguistic knowledge;
- To offer a customized comprehensive service to academic and industrial users;
- To prepare and organize research and development projects focusing on translation aids, multilingual authoring systems, and information retrieval;
- To provide a forum where experts from academia and industry share and assess tools and resources, assess software, evaluate new trends, investigate alternative approaches, and engage in joint activities;
- To make available the expertise of its partner institutions to the research community, to the public, and to language industry.

TELRI II implements these objectives in the following activities:

- Networking: This work package includes liaising with related infrastructure activities, centres, and institutions, promoting TELRI activities (newsletter, webpage, TELRI list), and strengthening the permanent infrastructure of the TELRI Association.
- TELRI Seminars: The series of successful TELRI seminars continues with annual seminars in CEE/NIS countries, this year to be held in Bulgaria.
- TRACTOR Service: This work package comprises promotion, support, and availability of customized service to the TRACTOR User Community.
- TRACTOR Tools and Resources: This work package focuses on the acquisition of attractive tools and resources for TRACTOR, the TELRI Research Archive of Computational Tools and Resources.
- Organizing Joint Research: TELRI partners will prepare European R&D projects with strong industrial involvement focusing on multilingual language and terminology issues.

2 TRACTOR's Role

TRACTOR is not the only language resource archive, as participants in this workshop are well aware. It does not aim to become the only focal point for language resource distribution on the web. Rather it hopes to complement other archives by providing a service for languages and users who are currently under-represented in existing archives. TRACTOR's unique strength lies in the amount of resources provided by centres in Central and Eastern Europe, and its role at the hub of a network of resource creation, standardisation and distribution which links the EU and non-EU European research communities. The TRACTOR User Community brings together resource providers, academic users and industrial users in an ongoing relationship, which is designed to foster the emergence of joint research projects in language engineering.
The resources are in many languages, with several different character sets and widely differing storage and markup conventions, and variable degrees of documentation, which is also in various formats. This raises particular challenges in terms of standardisation and evaluation of the resources. Since most of the resources have been collected over a period of several years, and from countries with varying

degrees of familiarity with the standards of the western European human language technology community standards, they do not in general conform to one, or in some cases to any, commonly accepted standards of digital storage, character encoding, textual or linguistic annotation. There are exceptions to this, such as the highly standardised corpus resources collected within the ELAN framework, now made available through the TRACTOR archive.

The TRACTOR philosophy is to accept deposits of resources in any format, and to distribute them in the form in which they are received (with small changes if possible such as additional documentation, and putting a browsable version or sample online.) In addition, certain standards are recommended and help is offered to providers who wish to make their resources conformant with the standards.

This lack of standardisation is not simply a pragmatic measure in the face of problems of heterogeneity. It is also based on the idea that standardisation should not be a barrier to language resource distribution. Sometimes providers of resources do not have the resources or expertise to adapt them to strict standards. We also recognise that the resource providers and users do not have requirements in common which can easily be translated into standards. Highly standardised resources built according to carefully crafted and strict criteria, and intended to be of commercial value such as EuroWordnet, the SIMPLE ontology, the ELAN network have more resources and attention paid to the design than the quality or usefulness of the knowledge which they contain. Furthermore, they are often too narrow in their applicability and in many cases they are simply not being used. TRACTOR aims to produce a synergy between academic and industrial researchers whereby the knowledge and resources can be shared and developed at a pre-competitive stage. Such collaborations can then develop on a commercial basis when on terms suitable for all parties. This is thought to be a more productive approach than the attempting to identify global, common needs and then to deliver large, valuable resources of general applicability.

As far as documentation of the resources is concerned, we believe that standardisation is probably a good idea, and we are looking into participating in the OLAC initiative.

3 Practicalities

3.1 How does it work?

TRACTOR provides ongoing help and support for the User Community, a task which chiefly involves solving simple queries regarding the nature of the resources on offer, and how to download them, and referring more complex queries to experts in the User Community, or in the wider TELRI network. Comments and feedback from users on the service and the resources are encouraged. The helpdesk tries to put users with research interests in common in touch with each other.

Visits to the website and downloads from it are logged exhaustively, and feedback is given to the resource providers.

3.2 How do you join?

Fill in the form at http://www.tractor.de/docs/user.html and arrange for the Euro 50 annual fee to be paid (details on the form). There is a special reduced rate of Euro 20 for Central and East European countries outside of the EU, and for NIS countries.

If researchers also want to deposit some of your resources with TRACTOR for distribution, we will be very pleased to receive them. In this case the fee is waived. Once these formalities are completed, we send a user ID and password by email so that you can access the resources.

3.3 What is the TUC?

The TUC is the TRACTOR User Community. This is the name given to everyone involved in depositing and accessing resources in the archive.

3.4 Why do we charge a fee?

There is a small annual administrative fee charged for joining the TRACTOR User Community. This is waived in the case of members of the TELRI Association and users who also deposit resources.

The principal reason for doing this is in order to put membership on a formal basis. To put it bluntly, if researchers have to pay for access, then they usually have to alert their colleagues and masters to it and persuade them to authorise the payment. Also, having paid for something, they are more likely to use it. The fee is considerably less than what users have to pay for more commercially oriented archives such as ELRA.

TRACTOR needs an active User Community, and we would prefer to have a compact group of committed and active users, to a large number of people who register because it is free, and then take no active part in the activities. Having said that, we do hope a large number of people and institutions join, and membership is not conditional on taking part in any activities!

3.5 What sort of resources can I deposit?

Resources are gratefully received for the TRACTOR archive. We currently aim to build up the archive with language corpora, lexicons and software tools for processing language. If you have something different which would be of use to the human language technology community, then we would also be very pleased to hear from you!

3.6 What sorts of resources are there?

There are currently corpora, tools, and lexicons. Corpora can be subdivided as general monolingual (e.g. German Parole, Uzbek, Ukrainian, Slovenian), specialist monolingual (the CFRL collection of Russian literature); parallel corpora (the Plato and 1984 corpora).

An important recent acquisition for the archive is the special TRACTOR version of the Qwick corpus analysis program, which is available to the TUC with indexed monolingual corpora in many languages. This is considered an important development as it means that corpora can now be delivered ready to analyse and use without any further demands on the expertise or software which the user may or may not have.

It is intended to greatly increase the range and depth of holdings in the TRACTOR archive throughout 2001, as a key part of the work of the Centre for Corpus Linguistics in the English Department at the University of Birmingham. Particular efforts are being made to create and acquire parallel translation corpora, which are of particular interest for research purposes for many academic and industrial users of TRACTOR.

TRACTOR is not in competition with other archives such as the Oxford Text Archive, ELDA and the Linguistic Data Consortium. TRACTOR is unique in it's combination of the following features:

- TRACTOR is not just an archive; it is a network of researchers with interests in linguistic resources;
- all resources are available for research purposes only and not for commercial exploitation;
- the primary interest of the TRACTOR community is working with multilingual resources;
- the strength and focus of TRACTOR is on Central and Eastern Europe;
- the acquisition philosophy is not to seek to impose standards of textual markup or linguistic annotation on resource providers.

4 The future

TRACTOR aims to build up particularly parallel corpora and tools for processing and extracting meaning from such resources. Given the continuing failure of attempts to build useful multilingual software on the basis of conceptual ontologies and other artificial intelligence approaches, and the failure of translation approaches on the basis of translation equivalence at the word level, it is the belief of Wolfgang Teubert and other researchers in Birmingham and in the TRACTOR network that the most productive way forward now is to extract knowledge about equivalence from parallel corpora and apply it to automatic multilingual text processing tasks.

This is just one area of interest, and members are free to develop ideas and resources in a multitude of directions.

It is also intended that the User Community continues to grow and become more active, not only as a group of researchers accessing the archive, but also as a research network.