Panel: Industrial Needs for Language Resources

Bente Maegaard[1], Khalid Choukri[2]

1) Center for Sprogteknologi, University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen S
bente@cst.dk

2) ELDA, 55-57 rue Brillat Savarin, F-75013 Paris
choukri@elda.fr

Purpose of the Panel

In this panel we want to discuss the needs for Language Resources as seen by industry. We also want to discuss the funding of such resources. How can industry and academia collaborate on language resources?
As a background we below give a short account of the results of investigations made in the European ENABLER network.

Motivation

LRs are needed for the language industry, for the translation industry, and in general for any content industry. Large companies produce their own resources for the languages for which a business can be made, but small companies cannot afford this, and often no resources are built for less spoken languages this way. Additionally, resources that are built by companies are normally not shareable.
As one consequence of these observations, an important goal of the ENABLER network was to map which type of resources and which type of LR format and content are needed and required by the European HLT industry, - or any other industry that wants to make tools for the European languages. The target group industries are the language industry, the translation industry, and relevant representatives of the content industries. 38 companies, both large and small industries, were interviewed.
The type of language resources needed by the companies, is of course determined by the language technology application. Companies involved in machine translation require domain-specific dictionaries and/or corpora, while companies developing search engines primarily demand semantic networks and/or thesauri covering technical sublanguages, and companies in the speech application field demand speech databases for a broader range of languages, tools to manipulate data, spontaneous speech etc.

Some results for the written area (WLR)

Acquisition

According to the investigation made, the language resources that are available are actually being bought by the companies. The providers of language resources can be either ELRA and LDC or other companies, or universities.
As modern language technology often builds on statistical models, huge amounts of text data are needed. Such amounts of data are mainly supplied by publishing companies. Acquisition of language resources from publishers for language technology purposes is normally not easy, as the publishers fear that the texts could be misused. In addition, the prices of the language resources from the publishing firms are considered as sometimes being very high. However, there seems to be a positive development in this area, with publishers getting a better understanding of the needs.
Another possible question is whether such data can be disseminated by organisations such as ELRA or LDC, or the publishers will always want to have direct control over the use of their data. This is still an open question, and a task for organisations like LDC and ELRA to pursue.

Quality

Regarding the quality most companies were satisfied, but a few were not. Data do not necessarily precisely fulfil what is needed in specific development tasks.
The large companies however, do acquire a good deal of language resources from external providers and make the conversion and updating necessary. This is a good sign as it shows that good quality language resources are valuable and worth buying even if they do not exactly meet the specific requirements.

Standards

With respect to how the data should be represented/stored (i.e. the format issues), the interchange formats, XML and SGML are indeed considered to be very useful and contribute to reducing the labour costs for conversion. This investigation showed no particular interest in standards as Martif, Geneter or OLIF.

Some results for the spoken area (SLR)

Most companies were involved in speech recognition. They develop speech databases and are involved in speech analysis and speaker verification. Fewer companies were involved in Speaker identification, Speech coding and Language verification.
Almost all companies were involved in the speech recognition assessment and evaluation, probably to keep an eye on the competition, whereas around half of them deal with the assessment and evaluation of TTS and dialog-based systems. Most of them are interested in Multimedia and Multimodal LRs.

Use of SLR

All the interviewed companies use internally produced LRs, and many also use LRs produced by specific contracted vendors and data centres. Most of the companies were satisfied with the acquired data. The major reason why companies were not purchasing data is that the data are not available. Among the LRs required the companies quoted: telephony speech/car/office databases, specific application oriented databases such as military-based LRs, close talk, GSM.

Languages

English and French are the most used languages. Some East European languages are still not used, like Albanian, Serbian and Ukrainian. Among the most needed languages we found German, Japanese and Spanish, English (including English spoken by non natives), French and Finnish. There is a need for various accents of languages like Egyptian/ Gulf or Maghreb countries Arabic, UK, US, Australian or Indian English, etc.

Type of LRs

All of the interviewed companies use or need Read speech (like the Speechdat family), most use Elicited speech and many spontaneous and prompted speech. Only a small part of them use or need Prepared speech. As to the bandwidth/condition of the acquisition, most mention In Car environment (both telephony and local microphones), many mention telephone speech and Wide band microphone and broadcast news. A small part mentions conversational telephony speech which is probably still an adequate resource for basic research activities.

Technical Information on formats and encodings

All the interviewed companies use standard telephony encodings. The most used file format is SAM and Wav, NIST/Sphere, files without header, and Au, AIFF formats were also mentioned at least once. Most companies use a sampling rate of 16 kHz, and many also 8kHz. As for the annotation standards, Sam labels and XML are the most used.

Validation of LRs

Most of the companies validate their data internally. Almost half of the companies have used external organisations to validate the data. Specific validation standards are followed in half of the cases and many, but not all, result in concrete validation reports.

Distribution of LRs

Close to half of the companies wish to make their resources available to others, but another half argue against distribution mostly for strategic and commercial reasons, while a few mention legal and technical reasons. Those who wish to make their data available are more eager to distribute them to end users and researchers than to agencies.

Conclusions

The ENABLER investigation confirmed that LRs are needed by industry: Not enough LRs are available, and the ones that exist do not always meet the requirements stated. This is especially the case for less used languages.

Less used languages are not very interesting seen from a commercial point of view, and hence there is a particular need for public support for the development of these resources. This fits very well with previous recommendations made by the EUROMAP project 2003 (*Benchmarking HLT progress in Europe*, 2003). Apart from this, the investigation also gives other ideas for the creation and provision of LRs.

Acknowledgements

We would like to thank our colleagues Claus Povlsen, CST, and Valérie Mapelli and Mahtab Nikkhou, ELDA, for their contribution to the ENABLER Industrial needs investigation.
The ENABLER network was supported by the European Commission.

References

Maegaard, B., K. Choukri, V. Mapelli, M. Nikkhou, C. Povlsen: *Language Resources – Industrial Needs*, ENABLER Deliverable 4.2, 2003, http://www.enabler-network.org/reports.htm