JEFF ALLEN WRITES:

The ELRA Language Resources survey: languages needed


Jeff Allen

One of the growing areas within the field of language engineering and technologies over the past few years is the setting up of centers that specialize in language and speech databases for human language technology systems and applications. The European Language Resources Association (ELRA) and its Distribution Agency (ELDA) are a tangible result of such efforts by establishing a European center that is dedicated to the identification, collection, validation, distribution and production of Language Resource (LR) databases, and to disseminate general information in the field. LRs, as defined here, are essentially sets of language data and descriptions in machine-readable form, used specifically for building, improving or evaluating natural language and speech algorithms or systems. In general, they are core resources for the software localisation and language service industries, for language studies, electronic publishing, international transactions, subject-area specialists and end-users.

ELRA/ELDA have been involved in surveying user needs since 1997, initially within the LE1-1019 European Commission funded project (Nilsson, 1997; Nilsson, 1998), and more recently within the LE4-8335 Language Resources - Packaging and Production (LRs-PP) European Commission funded project (Allen, 1999a; Allen, 1999b; Allen and Choukri, 2000). These user needs surveys are longitudinal in nature, have evolved and improved over time, and thus provide an excellent barometer for measuring the recent past, present and future needs of LR users. By surveying the recent and current needs of users, it is possible more effectively map out future trends for language-oriented databases that are necessary for training and testing new human language technology systems and applications. The results of these surveys are allowing ELRA/ELDA to streamline its approach for future marketing monitoring work, to increase their activities in the identification, collection and distribution of Language Resources (LRs), and to better plan new LR investments.

The survey conducted in 1999 and 2000 covered the following areas: speech systems; speech evaluation and assessment; text processing systems; authoring and translation environments; information processing systems; multi-media and multi-modal LRs; languages needed; LR subject domains/fields; and regional areas of respondents. Throughout this survey, we sent an LR user needs questionnaire directly to a total of 1,234 email addresses. After discounting invalid e-mail addresses that were rejected, there were 987 potential respondents of which 250 responded to our survey between August 1999 and March 2000. We are grateful for a very successful response rate of 25.3% through the survey methodology carefully described in Allen and Choukri (2000). In this short article, I would like to focus on the results specifically related to the languages that are needed.

One of the sections of the questionnaire focused on the languages desired with regard to LR data. It was possible for LR users to tick more than one language box in the questionnaire. The statistics indicated in Figure 1 reflect languages that received 20 or more responses and Figure 2 those languages that each received less than 20 responses. The percentages presented in the charts are therefore based on the total number of individual language boxes that have been selected (i.e., 1,326 selected) as well as with regard to the total number of survey respondents (250 responses).

It is clear that English, French, German, Italian, and Spanish -- set apart on the left side of Figure 1 -- are currently the most desired languages for LR databases for those working in the areas of speech and text processing. The middle percentile group of responses, in alphabetical order on the right side of

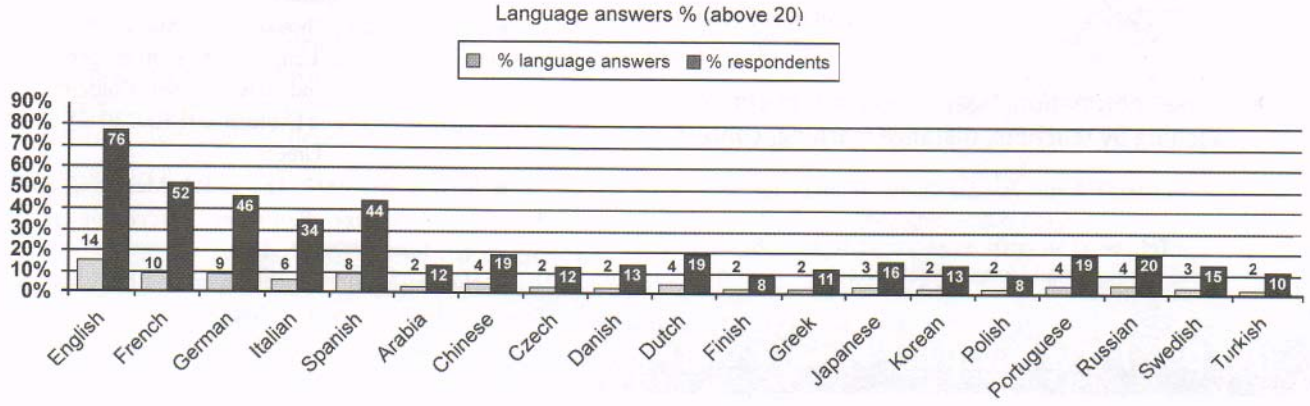

Figure 1. Over 20 responses per language

Figure 2.
Under 20 responses per language

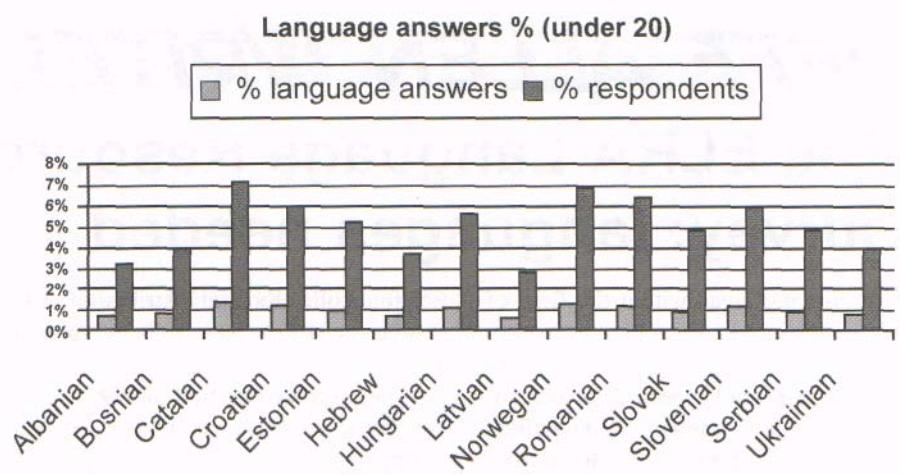

Figure 1 (i.e., Arabic, Chinese, Czech, Danish, Dutch, Finnish, Greek, Japanese, Korean, Polish, Portuguese, Russian, Swedish, Turkish), contain the Asian Languages and some of the other European languages. The languages that each received less than 20 responses include Albanian, Bosnian, Catalan, Croatian, Estonian, Hebrew, Hungarian, Latvian, Norwegian, Romanian, Slovak, Serbian, Ukrainian.

--In conclusion, these statistics have been taken from a survey conducted by ELRA/ELDA with 250 responses of which approximately 1/3 of respondents work in the area of speech technologies and approximately 2/3 in text processing. Although five European languages have been confirmed in this survey as having the highest amount of requests for LRs, this survey has revealed that other European languages, Asian languages, and some Middle-East languages should be focussed on in further LR collection, production and distribution efforts.

ELRA/ELDA plan to use the statistics obtained from this LR User Needs survey to focus on making more LRs available in its catalogue through further LR identification, collection, and production efforts and through launching future calls for proposals and tenders for LR production and packaging projects. If you are interested in taking part as a participant in this ongoing survey work, feel free to contact me at jeff@elda.fr or see the ELDA Web site (http://www.elda.fr). Future contributions to this journal will include other results that have been obtained from this survey work on LRs. ■

References:

- Allen, J., 1999a. Report on ELDA's Survey of Language Resource User needs. *European Language Resources Association (ELRA) Newsletter*, October-December 1999, 4.4:8-9.
- Allen, J., 1999b. Language Resources Go Digital: Update on the European Language Resources Association. *Language International magazine*, 11.6:38-39. Amsterdam: John Benjamins.
- Allen, J. and K. Choukri. 2000. Survey of Language Engineering needs: a Language Resources perspective. Presented at the Second International Conference on Language Resources and Evaluation (LREC2000), 31 May - 2 June 2000, Athens, Greece.
- Nilsson, M. 1997. The ELRA Marketing Survey. *European Language Resources Association (ELRA) Newsletter*, June 1997, 2.2:11.
- Nilsson, M. 1998. Final Report on ELRA Marketing Studies 1997 – short extract. *European Language Resources Association (ELRA) Newsletter*, May 1998, 3.2:3.