Survey of Language Engineering needs: a Language Resources perspective

Jeffrey Allen, Khalid Choukri

European Language Resources Association (ELRA) & European Language resources - Distribution Agency (ELDA) 55-57, rue Brillat-Savarin 75013 Paris France {jeff, choukri}@elda.fr

Abstract

This paper describes the current state of an on-going survey that aims at determining the needs of users with respect to available and potentially available Language Resources (LRs). Following market monitoring strategies that have been outlined within the Language Resources- Packaging and Production project (LRsP&P LE4-8335), the main objective of this survey is to provide concrete figures for developing a more reliable and workable business plan for the European Language Resources Association (ELRA) and its Distribution Agency (ELDA), and to determine investment plans for sponsoring the production of new resources.

1. Introduction

This paper provides results of questionnaires on user needs of a survey that has been conducted in 1999 and 2000 by the European Language Resources Association (ELRA) and its Distribution Agency (ELDA) within the LE4-8335 European Commission funded project. This is a follow-up of initial survey work conducted by ELRA/ELDA in 1997 (Nilsson, 1997a; Nilsson, 1997b, 1998) within the LE1-1019 project. Taking a multi-tier approach for gathering information on user needs, these 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. We describe herein the approach and procedures of the recent survey and provide anonymous results that have been obtained. These results are allowing ELRA/ELDA to streamline its approach for future marketing monitoring work, for the identification, collection and distribution activities of Language Resources (LRs), and to better plan new LR investments.

The questionnaire analyzed in this paper was sent mainly to respondents who are not ELRA members. It was sent directly to potential respondents during the first stage of the survey. The second stage of the survey included sending reminder requests to those who had not responded earlier. We provide statistics on various areas of Language Resource activities including: speech systems; speech evaluation and assessment; text processing; text processing systems; authoring and translation environments; information processing systems; multimedia and multi-modal LRs; languages needed, LR domains/fields; and regional areas of respondents.

2. Survey Methodology

After receiving a low amount of responses to the 1997 ELRA/ELDA LR User Needs Survey (Nilsson, 1998), ELDA staff revised the survey methodology and redesigned the questionnaire. Unlike the 1997 Survey that contained many open questions, the new questionnaire has aimed at providing questions with binary yes/no and check the box options that would limit the amount of time necessary for a participant to complete the questionnaire. Also, the new questionnaire was sent in personalized messages to all of the respondents. This questionnaire was designed to only take 10-15 minutes of time to complete.

The new questionnaire was sent out in 667 personalized messages during the month of August 1999 to individuals in the general field of language engineering and human language technologies that are listed in one of the contact databases at the ELDA office.

Of the nearly 670 questionnaires sent out to these language engineering specialists, 17.5% of the messages returned as bad addresses. After discounting the invalid addresses, the 90 respondents who returned a completed questionnaire to us represented 16.4% of the total number of valid addresses of potential respondents. The preliminary results obtained from the first 90 respondents are considered to be the first stage of this survey and have already been published (Allen, 1999c). We will not repeat the details in this paper but will rather include them in the cumulated statistics. Given that the first stage of sending out the new version of the questionnaire had a very successful response rate, we proceeded with extending the coverage of the potential respondents for the second stage of this LR survey work that was conducted in October 1999 through January 2000.

For the second stage of the survey, a clean-up procedure was conducted to correct and/or remove the invalid addresses in order to improve the response rate. Of the 460 valid addresses that did not respond to the first stage of this survey in August 1999, 367 questionnaire reminders were sent out (76% of non-responding valid addresses of the first stage). We did not recontact all 460 potential addresses because it was estimated that many general delivery e-mail addresses for companies did not yield good enough results, so more effort was placed on locating e-mail addresses for specific people. We were able to better target potential respondents by contacting people from other databases available at the ELDA office. We also contacted people whom ELDA staff had met at several conferences in 1999. Authors of papers in various conference proceedings were also contacted. E-mail list job advertisements appearing in September to December 1999 were also a source of potential respondents. In addition, we contacted people with whom we have corresponded by e-mail in 1999 on different issues. A total of 916 questionnaires were sent out during the second stage of this survey, including reminder messages and new contacts. Those who responded to the first stage of the survey were not recontacted with the second stage questionnaire – which was nearly identical to the questionnaire of the first stage.

Of the nearly 1000 questionnaires (re)sent out individually to these language engineering specialists, only 130 (14%) bounced back as invalid addresses. In lowering the percentage of invalid addresses, we have shown that our survey methodology is progressively more effective. After again discounting the invalid addresses, there were 160 respondents who returned a completed questionnaire to us during the second stage of the survey. The 20.3% response rate of the second stage demonstrates nearly a 4% increase in responses between the first and second stages of the survey. A total of 1,234 addresses were contacted in both stages of the survey, but there is only a total of 987 valid addresses after discounting all invalid addresses of both stages. Of these 987 potential respondents, we have obtained a very successful response rate of 25.3% through the methodology mentioned above. The statistics provided below are based on the 250 completed questionnaires received in both stages of this survey of LR User Needs.

3. General Types of Data

The first important analysis of data obtained from the survey results concerns the types of LRs that the users indicate that they work with and/or are interested in. The four types of LRs presented in Figure 1 are speech databases, syntactic databases (including syntactically annotated data, data for parsers), lexical databases, and text databases. Each LR type has been divided into basic non-annotated data vs. annotated data. It can be noted that 30-40% of all respondents are interested in speech data and 50-60% are interested in textual data. Inbetween percentages are provided for syntactic and lexical databases. The overall percentages of both stages of the survey correspond closely with the results already published in Allen (1999) that contain statistics on the 90 respondents of the first stage. The cumulated results show that there is a lower number of users of speech data with respect to written data, although we did attempt to seek to contact more spoken LR users in the second stage of the survey. These figures also show that our survey work is reaching a high number of potential users of written LRs an objective of ELDA for 1999 and 2000.

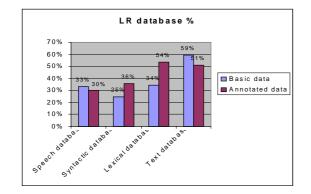


Figure 1. LR Type percentages

4. Speech processing

The second section of the questionnaire for this survey aims at gathering information about the type of work being conducted in the Speech domain. In this section and all subsequent sections, the LR users are divided into those who conduct research and those who develop products.

Figure 2 provides percentages on general speech categories including: speech recognition; speech synthesis; speech databases; speech analysis; speech coding; speech workstation software. From these results, we see that up to 1/3 of surveyed respondents are involved in speech recognition with the other areas of speech R&D trailing behind with slightly lower percentages.

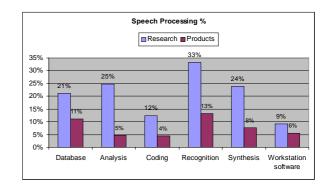


Figure 2. Speech Processing percentages

4.1. Speech systems

One subsection of the questionnaire specifically addresses different types of speech systems. Figure 3 provides the corresponding results on LR needs for the following types of systems: speech understanding; text-tospeech conversion; speaker identification; speaker verification; language identification; language verification; speech coding. These results again are based upon the total number of respondents of both stages of the survey.

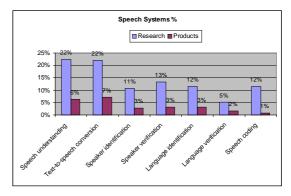


Figure 3. Speech System percentages

The percentages over 20% in Figure 3 reveal that a greater number of LRs users in this subfield are working with speech understanding (including speech recognition and speech dictation) and text-to-speech conversion systems. Most of the other speech systems each represent 10-15% of respondents.

4.1.1. Speech evaluation and assessment

Evaluation and assessment of speech recognition and speech synthesis are areas that were also surveyed. The results given in Figure 4 show that nearly 25% of all respondents need LRs for the evaluation and assessment of speech recognition systems for research purposes whereas there is a lower percentage for speech synthesis for research. Percentages are nearly equal of LR users conducting the evaluation and assessment of products.

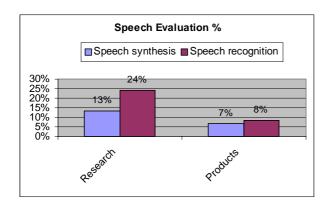


Figure 4. Speech Evaluation percentages

5. Natural Language Processing (NLP)

Some general areas of natural language processing (NLP) were identified and an excellent set of positive results have been obtained from this survey for the area of written LRs. The highest percentages were over 40% for research and over 15% for products and listed in order of highest to lowest are the following: text corpora; parsers; grammar development. The lower range of LRs users in this area included those working on automatic lexicon recognition, text/message understanding, dialogue management, and discourse understanding with percentages between 19-40% for research and 4-10% for products.

5.1.1. Text processing

Text processing systems are those that basically deal with the production and use of textual material. Spell checking work represents the highest amount of LR users at 23% for research and 14% for commercial purposes. A middle range of percentages from 12-18% for research and 5-10% for research was indicated for style checkers, grammar checkers, controlled language checkers, computer-aided writing tools, and multilingual word processing. The lowest percentage was for desktop publishing.

5.1.2. Authoring and Translation environments

This subsection includes systems that are used in authoring and translation environments and includes

machine translation, translation memory, terminology management, controlled language implementation, and multilingual electronic dictionaries. Machine translation, terminology management, and multilingual electronic dictionaries occupy a significant spot with respect to LR needs with each having over 30% of respondents for research and over 15% of respondents for products as seen in Figure 5.

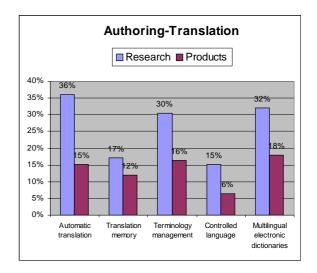


Figure 5. Authoring/Translation system percentages

5.1.3. Information Processing systems

This subsection of the questionnaire includes advanced techniques for the processing of written texts, including information retrieval, data mining, document indexing, text summarization, topic detection, and optical character recognition. Information Retrieval (IR) is opposed to Data Mining and Text Mining on the points of directed vs. undirected queries, ambiguous vs. unambiguous elements of information being sought, and already structured vs. unstructured data or information.

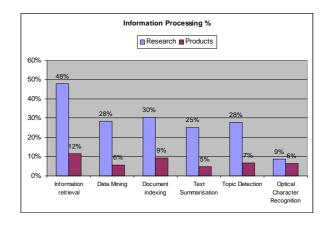


Figure 6. Information Processing percentages

As seen in Figure 6, there is currently a significant amount of research being conducted by 25-50% of respondents for the nearly all of these sub-areas that are a rapidly expanding part of current NLP work.

6. Multi-media and Multi-modal LRs

One of the most recent demands for LRs falls in the area of Multi-media and Multi-modal data. As for Multimodal Processing, the recent survey shows that 52% of all respondents are interested in Multi-media data and 35% are interested in Multi-modal data. The specific sub-areas of Multi-modal processing that have been identified and surveyed by this questionnaire include: face tracking, gesture recognition, facial analysis, eye-gaze tracking, face recognition, person identification, speech/lip reading, focus of attention, facial animation and multi-modal error recovery. From 5-10% of all respondents state that they want one of these several types of Multi-modal LRs for research, as shown in Figure 7. Product development is still low, but this is expected to grow quickly since this is a new area of research and development. ELRA/ELDA notice an overwhelming increase in Multi-modal LRs information since the 1997 Autumn/Fall Survey only indicated that 1/18th of the surveyed participants were interested in Multi-modal LRs.

Since researchers and developers in the Human Language Technology field are showing interest in this kind of data, it is important that ELRA/ELDA continue to closely monitor and survey this area in further detail in order to more adequately respond to this increasingly important area for new LRs.

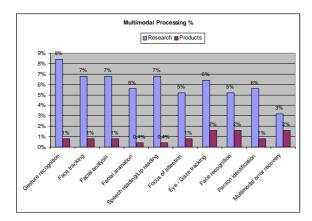


Figure 7. Multi-modal percentages

7. Languages needed

Another one of the questionnaire sections asks for the languages desired with regard to LR data. These statistics clearly help ELRA/ELDA understand language data needs, correlated with what is currently offered, and to see where there is a lack in what is being offered today. It was possible for LR users to tick more than one language box in the questionnaire. The statistics indicated in Figure 8 reflect languages that received 20 or more responses and Figure 9 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).

It is clear that English, French, German, Italian, and Spanish, set apart in the left of Figure 8, are currently the most desired languages for LRs. The middle percentile group of responses, to the right and in alphabetical order in Figure 8, containing the Asian Languages and some of the other European languages. The languages that receive less than 20 responses are for the most part Eastern European languages.

A general conclusion to make from these language statistics is that ELRA/ELDA have been responding to the need for European languages. Some Eastern European LRs have been made available to meet such needs. There is however a need for more LRs for the main Asian languages (Chinese, Japanese and Korean) and for Arabic, as can be seen in Figure 8.

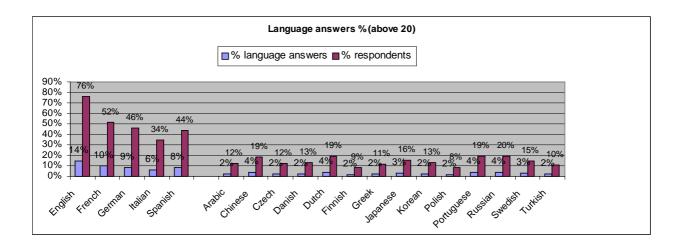


Figure 8. Over 20 responses per language

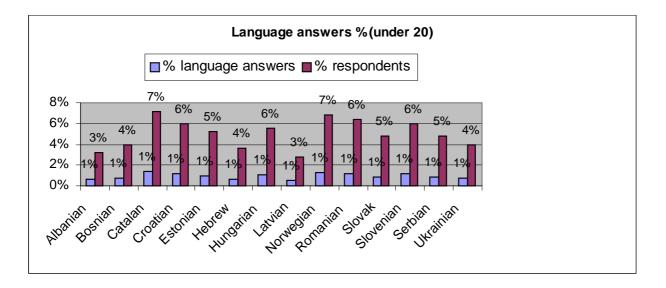


Figure 9. Under 20 responses per language

8. LR domains/fields

The questionnaire also contained a section on the domains/fields for LR data. The results are summarized below.

43% of the respondents indicated that they are interested in LRs from all domains.

The specific fields desired by at least 20% of respondents are: Computer Science; Telecommunications; Technology; Business; Economics and Finance; Administrative; Data Processing.

The fields desired by 10-19% of respondents are: Automobile; Medicine; Education/Pedagogy; Law; Electrical Engineering; Electronics; Health; Tourism; Mechanical Engineering; Pharmaceutical. The fields desired by under 10% of respondents are: Aeronautics; Heavy-machinery; Sports; Leisure; Chemistry; Geography; Biology; Agriculture; Navigation; Arts; Architecture/construction; Physics; Food Sciences; History; Psychology; Sociology; Geology.

9. Countries of respondents

The questionnaire also included a section asking for where the respondents' R&D labs are located in order to establish a demographic profile of users. More than 10 replies per country came from the USA, the UK, Germany, France, Japan, The Netherlands, Spain, Italy and Canada, as shown in Figure 10. There are several countries represented that have 0.4 to 3% of responses. These include: Korea, Austria, Denmark, Sweden, Switzerland, Australia, Belgium, China, Finland, Ireland, Portugal, Russia, Czech Republic, Norway, Poland, Taiwan, Romania, Slovenia, Bulgaria, Estonia, Greece, Hungary, Croatia, and Slovakia.

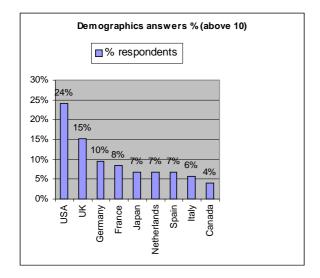


Figure 10. Demographics percentages

10. Conclusion

The statistics provided in this paper are taken from a recent two-stage LR User Needs Survey conducted by ELRA/ELDA in the summer 1999 for the first stage and in the fall of 1999 and beginning of 2000 for the second stage.

A total of 1,234 e-mail addresses were contacted for both stages of the survey with personalized messages containing a LR User Needs questionnaire. Of these 1.234 addresses, 987 addresses were found to be valid. We received 250 completed questionnaires during a time span beginning in September 1999 and ending at beginning of March 2000. The number of completed questionnaires received show a very successful overall response rate of 25.3% of which there was nearly a 4% increase in responses between the first and second stages of the survey. In general, as noted from on-going survey work conducted by ELRA/ELDA since 1997, approximately 1/3 of respondents are interested in speech LR data, and approximately 2/3 are interested in written LRs. As ELDA has been proactive in 1999 and 2000 for investigating issues with regard to the distribution of written LRs, this recent survey work confirms that a high percentage of written LR users does in fact exist across multiple sectors.

This survey confirms earlier indicators of a significant interest in Information Processing systems at various R&D laboratories. This survey also indicates

that there has been a growing interest in the area of Multi-media and Multi-modal work since our 1997 survey, although we do acknowledge that publicly available LRs for these latter areas are underrepresented currently at present.

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-Eastern languages should be focussed on in further LR collection, production and distribution efforts.

Other than responses indicating an interest in all LR domains (43% of all respondents), the specific domains of greatest need for LRs are Computer Science, Telecommunications, Technology, Business, Economics and Finance, Administrative, and Data Processing. Two other lower percentile ranges of LR domains are also found from results of this survey.

Previous ELRA/ELDA surveys (Nilsson, 1997a; Nilsson, 1997b; Nilsson, 1998) have been used to make investment plans for funding the production of new LR projects (Allen, 1999a; Allen, 1999b; Allen, 1999d). ELRA/ELDA also plan to use the statistics obtained from this more recent LR User Needs survey to focus on making more LRs available in its catalogue through further LR identification, collection, and production efforts. Such surveys also provide valuable information on which ELRA/ELDA can base decisions for launching future calls for proposals and tenders for LR production and packaging projects.

Additional surveys are currently being conducted to further study the specific sections and subsections of the questionnaire analyzed in this paper. Responses from all ELRA/ELDA questionnaire are also analyzed to determine figures on market segmentation of Language Resources. This allows ELRA/ELDA to more adequately identify factors that contribute to a lack of LRs. By doing so, it is possible to provide a better service to LR customers.

We hope that the positive results obtained from this survey and presented in this paper will encourage more LR users to participate in ELRA/ELDA surveys. We continue to improve on the content of our surveys from valuable feedback and comments received from survey respondents. We also review our survey distribution strategy on a regular basis to determine how to better reach appropriate potential respondents. In addition to the 1999-2000 LR User Needs Survey, there are several other surveys currently underway: ELRA survey on evolution of speech products/languages; ELRA survey on evolution of translation software products/languages; ELRA survey on evolution of Information Retrieval / Text Mining products/languages. By responding to these questionnaires, for which all specific information is kept strictly confidential, respondents allow ELRA/ELDA to improve on its LR identification, collection, production and distribution activities for the Language Engineering and Human Language Technology community at large. For more information about ELRA/ELDA questionnaires that are currently available, please contact the ELDA office at the contact numbers and addresses provided in the header of this paper, or see our Web site (http://www.elda.fr).

11. Acknowledgements

We would like to thank the European Commission for funding the Language Resources – Packaging and Production project (LE4-8335) that has allowed ELDA to carry out this valuable survey work. Our gratitude is also extended to Emilie Marquois and Valérie Raymond for their assistance in analyzing some of the results from the questionnaires that were received.

12. References

- Allen, J., 1999a. ELRA 1999 Call for Proposals ELRA Commission Production of Language Resources. *European Language Resources Association (ELRA) Newsletter*, January-March 1999, 4.1:8-9.
- Allen, J., 1999b. Results of ELRA 1999 Call for Proposals – ELRA Commission Production of Language Resources. *European Language Resources Association (ELRA) Newsletter*, April-June 1999, 4.2:6-7.
- Allen, J., 1999c. 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., 1999d. Language Resources Go Digital: Update on the European Language Resources Association. *Language International magazine*, 11.6:38-39. Amsterdam: John Benjamins.
- Choukri, K., V. Mapelli, and J. Allen. 1999. New developments within the European Language Resources Association. Paper presented at Eurospeech99. Budapest, Hungary, 4-12 September 1999.
- Nilsson, M. 1997a. The ELRA Marketing Survey. European Language Resources Association (ELRA) Newsletter, June 1997, 2.2:11.
- Nilsson, M. 1997b. ELRA Market Segmentation Survey - update. *European Language Resources Association* (*ELRA*) Newsletter, October 1997, 2.3:10.
- Nilsson, M. 1998. Final Report on ELRA Marketing Studies 1997 – short extract. *European Language Resources Association (ELRA) Newsletter*, May 1998, 3.2:3.