The ELRA Newsletter

January - March 2003

Vol.8 n. 1

Contents

Signed articles represent the views of their authors and do not necessarily reflect the position of the Editors, or the official policy of the ELRA Board/ELDA staff.

Editor in Chief:
Khalid Choukri

Editors:
Khalid Choukri
Valérie Mapelli
Magali Jeanmaire

Layout:
Magali Jeanmaire

Contributors:
Bidyut Baran Chaudhuri
Pat Hall
Durgesh Rao
NEMLAR Consortium

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
55-57, rue Brillat Savarin
75013 Paris - France
Tel: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail: choukri@elda.fr
Web sites:
http://www.elra.info or
http://www.elda.fr

Dear Colleagues,

We will present in this first letter for the year 2003 our plans and the strategic tasks ELRA and its agency, ELDA, will focus on for the coming months: the production and the validation of LR, the evaluation of HLT, and our activities in the area of terminology. During our annual General Assembly, the ELRA Board will elaborate more on these axes.

First, concerning the production of new LR, we have participated in a number of projects the objective of which is to produce LR, mainly Spoken Language Resources, for various purposes. For example, within the European Speecon project, which aims at launching voice-driven interfaces for consumer applications, 18 speech databases in total were to be created; at ELDA, we have produced, or supervised the production of the French, Italian and Swedish ones. For these speech databases, 550 adults and 50 children speakers distributed in 6 dialectal regions for each country were recorded in 5 different environments. The databases are to be validated and will be distributed by ELDA.

We will produce other speech databases, notably in the framework of the European OrienTel project, for the launch of multilingual interactive communication services in Mediterranean countries. We are in charge of collecting the speech data in Morocco and Tunisia, in Modern Standard Arabic, Modern Colloquial Arabic and French. The recordings in each country will cover 3 linguistic variants.

Other projects for the production of new LR which will continue are Network-DC, C-ORAL-ROM, for the production of spoken corpora in the 4 main Romance languages (Portuguese, Spanish, Italian, and French), SALA II, for the production of speech databases in North and South American languages, etc.

Regarding our activities in the area of LR validation, we offer to the HLT community since the beginning of 2002 some services aiming at ensuring the quality of our resources. We work in cooperation with SPEX (the SPeech EXpertise center, located in the Netherlands) for the validation of the SLR. In addition to the bug report service which was set up last year and which allows the users of SLR to report the bugs and discrepancies they may find in the SLR they purchased from ELDA, our partners have produced at SPEX a number of Quick Quality Check reports, which describe the quality of the concerned resource, and which may be made available from our web site.

For 2003, ELRA and its validation committee, VCom, plan to launch the same kind of services for the WLR, thanks to its newly selected network of VC_WLR centres, headed by CST (Center for SprogTeknologi, Denmark). They will establish the methodology for WLR quality, provide the standards for the validation of WLR, and apply these to the existing resources at ELRA. A bug report service for WLR will also be launched, similar to the one for SLR. We intend to have in the near future most of our resources, either spoken or written, checked or/and validated.These activities are visible through the web pages which have been set up by SPEX, CST, and ELRA, where more information about our validation process can be found.

Evaluation of HLT is another strategic task. In order to better reflect our involvement in the field, ELDA changed its name to “Evaluations and Language resources Distribution Agency”, and we started setting up a team fully dedicated to this activity. Our participation in projects and campaigns at the French and European levels has allowed us over the past months to get more deeply involved in evaluation work. We have played an active role in Technolangue, the French programme for language technologies which consists of 4 sections (development and reinforcement of LR - both data and tools, creation of an infrastructure for the evaluation of HLT, better accessibility to norms and standards, and set up of an intelligence watch network in HLT), and notably coordinate within Technolangue the EVALDA project, which includes itself 8 sub-campaigns for the evaluation of HLT for the French language (parsers, machine translation systems, multilingual text alignment, terminology extraction, information query, speech synthesis, broadcast news transcription and indexing, and oral dialog).

Other projects related to the evaluation of HLT we have been participating in are: CLEF (cross-language evaluation forum), and Aurora, with the distribution of speech databases for evaluation purposes. In addition, we will collaborate for the set up of a European evaluation infrastructure, the first one of its kind, to be launched within the 6th Framework Programme of the European Commission, and capitalising on ELRA experiences for the management of LR, acting as a European clearing house.

ELRA intends to stress its presence in the field of terminology, and to develop its activities related to terminological resources. We plan to conduct a survey aiming at identifying the users of terminological resources, to whom we will address a questionnaire to identify their needs. ELRA will be able then to better adapt its offer to the needs, improving its involvement in the field.

You are kindly invited to contact us if you would like to obtain more information about our activities in the areas mentioned above, or if you would like us to help you on the matter, through e.g. customized production, validation of specific resources, evaluation of technologies, etc.

We will take care during the coming months of the organisation of LangTech 2003, to be held in Paris (France) next November. LangTech 2003 is the second edition of the European Forum for Language Technologies. ELRA was involved in the organisation of LangTech 2002 in Berlin, in which over 330 HLT specialists from both academic and industry participated, which appears as a very successful and fruitful event. We are now responsible, in cooperation with a number of partners, for the organisation of its second edition, which will take place in Paris on 24-25 November. Information about LangTech 2003, including the registration procedure, programme, etc., as well as an overview of LangTech 2002, are available at http://www.lang-tech.org. This web site will be updated regularly to reflect the latest news regarding the November 2003 event.

Further to the projects in which we have been participating for the last few months, we will be involved in new ones within the EC FP6 programme; here is a brief overview of two of these projects, NEMLAR and TC-STAR.

For the latter, we have actively participated in its preparatory phase, named TC-STAR_P, helping to draw the blueprint with the objectives of the project, the schedule, the budget, etc. TC-STAR (Technology and Corpora for Speech to Speech Translation) aims at making speech-to-speech translation real, bringing together industrial partners and researchers to define first the requirements, develop the appropriate systems, produce the suitable LR, and provide the technologies. You will find more about TC-STAR at http://www.tc-star.org.

As for NEMLAR (Network for Euro-Mediterranean LAguage Resources), the goal is to establish a network of partner centres dedicated to NLP and LR in Arabic and other southern Mediterranean languages, located in 6 countries, namely Jordan, Morocco, Egypt, Lebanon, Tunisia, West Bank, Gaza Strip. A paper about NEMLAR is available in this issue, and more detailed information at http://www.nemlar.org.

In addition to the paper giving an overview of NEMLAR, this first newsletter for the year 2003 includes three articles dealing with NLP for Indian languages. First, B.B. Chaudhuri contributed an article dealing with OCR systems for Indian languages. An overview of machine translation in India was prepared by Durgesh Rao. Finally, Pat Hall agreed to write a short paper reporting on the SCALLA project, for which a more detailed document will be available in our next issue. The descriptions of the new resources added to our catalogue during the first quarter 2003 are available at the end of this bulletin.

Joseph Mariani, President

Khalid Choukri, CEO

A Survey of Indian OCR Systems

Bidyut Baran Chaudhuri

Introduction

Optical Character Recognition (OCR) is a process of automatic computer recognition of characters from optically scanned and digitized pages of text. OCR is one of the most fascinating and challenging areas of pattern recognition, with various practical applications. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications. Some practical application potentials of OCR system are as follows: (1) reading aid for the blind, (2) automatic text entry into the computer, desktop publication, library cataloguing, ledgering, etc., (3) automatic reading for sorting of postal mail, bank cheques and other documents, (4) document data compression (from document image to ASCII format), (5) language processing, (6) multi-media system design, etc. [14].

At present, sophisticated optical readers of Roman, Chinese, Japanese and Arabic texts are available. OCR systems in south Asian systems are still not very mature. The comment holds for Indian scripts as well. However, some important work has been carried out during the last ten years. This article is a brief survey of such activities.

Properties of Indian Scripts

In India, there are eighteen official (Indian constitution accepted) languages, namely Assamese, Bangla, English, Gujarati, Hindi, Konkanai, Kannada, Kashmiri, Malayalam, Marathi, Nepali, Oriya, Panjabi, Rajasthani, Sanskrit, Tamil, Telugu and Urdu. Among them, Hindi and Bangla are the first and second most popular languages (ranked 4th and 7th respectively in the world). Twelve different scripts are used for writing these languages (see Fig.1). Most Indian scripts originated from Brahmi, through various transformations. Of these, the Devnagari script is used to write Hindi, Marathi, Rajasthani, Sanskrit and Nepali languages, while the Bangla script is used to write Assamese and Bangla (Bengali) languages.

Apart from vowel and consonant characters, called basic characters, there are compound characters in most Indian script

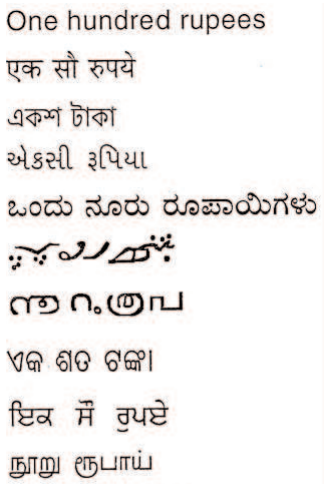

Fig. 1 Examples of 12 Indian scripts meaning "One hundred rupees": from top to bottom, English, Devnagari, Bangla, Gujarati, Kannada, Kashmiri, Malayalam, Oriya, Panjabi (Gurumukhi), Tamil, Telugu and Urdu.

alphabets (except for the Urdu, Tamil and

Gurumukhi scripts). These are formed by combining two or more basic characters. The shape of a compound character is usually more complex than the constituent basic characters. In some scripts, a vowel following a consonant may take a modified shape, which, depending on the vowel, is placed to the left, right, top, or bottom of the consonant. They are called modified characters. In most scripts, the total number of all these shapes is about 350.

In some script alphabets (like Devnagari, Bangla and Gurumukhi, etc.), it is noted that many characters have a horizontal line at the upper part.

In Bangla, this line is called *matra*, while in Devnagari it is called *sirorekha*. However, in this paper, we shall call it *head-line* (see Fig.2). When two or more characters sit side by side to form a word in the language, the head-line portions touch one another and generate a big head-line. Because of these, character segmentation is necessary for the OCR of these scripts.

Work on Indian Language Character Recognition

There are not sufficient numbers of published studies on Indian language character recognition. Most of the existing studies focus on Devnagari and Bangla, the two most popular scripts in India. A few pieces of work also report on the recognition of Tamil, Telugu, Oriya, Kanada, Panjabi, and Gujrathi scripts. Structural and topological features-based tree classifiers, support vector machines and neural network classifiers are primarily used for the recognition of Indian scripts. The Ministry of Information Technology, Government of India, has initiated the TDIL (*Technology Development on Indian Languages*) project, under which OCR system development for most of the official Indian language scripts has been taken up by different labs and academic institutions.

A script-wise study of Indian OCR systems is given below.

a-Devnagari character recognition

OCR work on Devnagari script started in the early seventies. Within earlier studies, a syntactic pattern analysis system and its application to Devnagari script recognition was discussed in the doctoral thesis of

Fig.2 Different zones of (a) English and (b) Devnagari text lines.

Sinha[31]. The work was later extended in [32]. Only recognition of basic characters was considered. He proposed a Plang language for the purpose and claimed that it would work for handwritten character as well. Sinha also demonstrated how the spatial relationship among the constituent symbols of Devnagri script plays an important role in the interpretation of Devnagari words [30].

Sethi and Chatterjee [28, 29] also made earlier studies on Devnagari character recognition. On the basis of presence and absence of some basic primitives, namely horizontal line segment, vertical line segment, left and right slant, D-curve, C-curve, etc., as well as their positions and interconnections, they presented a Devnagari numeral and alphanumeric recognition system based on binary decision tree classifier.

As stated, all these studies are about recognition of isolated characters, and they have not shown results of scanning real document pages. The first report in a complete Devnagari OCR system is due to Palit and Chaudhuri [33]. It was latter extended by Chaudhuri and Pal [8, 21].

The system is a very robust one and gives 97% accuracy for a variety of fonts. It uses a stroke based tree classifier for initial grouping of characters. The characters in each group are classified using crossing count based normalized feature vectors. Post processing and error correction are also used to improve the accuracy. The system has already been commercialized under a Technology Transfer agreement between ISI and C-DAC.

b-Bangla character recognition

Though research on Bangla character recognition started in the late seventies [37], no significant work was reported before the mid-nineties. Recently, several pieces of work on Bangla OCR have been published [8, 20, 22]and some commercial OCR system will be marketed.

Ray and Chatterjee [27] presented a nearest neighbour classifier employing features extracted by using a string connectivity criterion for Bangla character recognition. Dutta presented a generalized formal approach for generation and analysis of the Bangla and Devnagari characters.

The initial Bangla character recognition system developed by Chaudhuri and Pal [8] is capable of handling the character recognition of machine-printed multisizes Bangla script. In this system, preprocessing involves skew correction, followed by noise removal and preliminary segmentation of the input image into lines, zones and letters. A combination of feature and template matching is employed for recognition.

For the OCR system for the Bangla and Devnagari scripts, Chaudhuri and Pal [6] proposed a new technique for skew estimation and correction. Also Garain and Chaudhuri [12] proposed a method which combines the positive aspects of feature-based and run number-based normalized template matching techniques.

Some pieces of work on Bangla handwritten text are also available. Using syntactic method, Parui et al [40] proposed a recognition scheme for isolated Bangla handwritten numerals. Rahman et al. [23] proposed a multistage classification scheme for handwritten Bangla character recognition.

Neural network approach is also used for the recognition of Bangla characters. Dutta and Chaudhuri [11] reported a work on recognition of isolated Bangla alphanumeric characters using neural networks. Concept of fuzzy sets is also used for Bangla script recognition. Sural and Das [35] defined fuzzy sets on Hough transform of character pattern pixels.

Work on on-line recognition of Bangla characters also exists. Garain and Chaudhuri [40] proposed an on-line handwriting recognition system for Bangla. The primary concern of the approach is the modelling of human motor functionality while writing the characters. This is achieved by looking at the pen trajectory where the time evaluation of the pen coordinates plays a crucial role. A low complexity classifier has been designed and the proposed similarity measure appears to be quite robust against wide variations in writing styles. For recognition of touching characters, see [13].

c-Tamil character recognition

Siromony et al. [36] described a method for recognition of machine printed Tamil characters using an encoded character string dictionary. The scheme employs string features extracted by row and column-wise scanning of character matrix. The features in each row (column) are encoded suitably depending upon the complexity of the script to be recognized. Chandrasekaran [5] used similar approach for constrained hand-printed Tamil character recognition.

Chinnuswamy and Krishnamoorthy [10] proposed an approach for hand-printed Tamil character recognition. Another work on on-line Tamil character recognition is reported by Sundaresan and Keerthi [41]. They used four types of features which are obtained from a sequence of directions and curvature, a sequence of angles, Fourier transform co-efficient and Wavelet features. The accuracy of the system is claimed to be about 96%.

d-Telugu character recognition

Some pieces of work on Telugu characters are published in the literature [19, 24, 25]. A two-stage recognition system is presented by Rajasekaran and Deekshatulu [24] for Telugu alphabet. In the first stage, they applied a knowledge-based search to recognize and remove the primitive shapes present in the input character. A directed curve-tracing method is used for the purpose. In the second stage, the pattern obtained after the removal of primitives is coded by tracing along points on it. On the basis of knowledge about primitives and basic characters in the input pattern, classification is achieved by a decision tree.

Sukhaswami et al. [38] presented a recognition system for printed Telugu characters by neural networks approach. Initially, they used Hopfield neural network model for the recognition purpose. Due to limitation in the storage capacity of the Hopfield neural network, they later propose a new scheme called MNNAM (*Multiple Neural Network Associative Memory*).

Recently, Negi et al. [19] presented a system for printed Telugu character recognition. A compositional approach using connected components and fringe distance template matching are used for the recognition. Fringe distances compare only the black pixels and their positions between the templates and the input images.

e-Oriya character recognition

Using Kohonen neural network Mohanti [18] proposed a system for alphabets recognition of Oriya script. In a system developed by Chaudhuri et al. [9] for the basic characters of Oriya script, the document image is first captured using a flat-bed scanner and then passed through different pre-processing modules like skew correction, line segmentation, zone detection, word and character segmentation, etc. These modules have been developed by combining some conventional techniques

with some newly proposed ones. Next, individual characters are recognized using a combination of stroke and run number-based features, along with features obtained from the concept of water overflow from a reservoir. The feature detection methods are simple and robust, and do not require pre-processing steps like thinning and pruning. The system has achieved about 96% accuracy.

f-Gurumukhi (Punjabi) character recognition

Lehal and Singh [17] developed a complete OCR system for Gurumukhi script, where connected components are first segmented using a thinning-based approach. In the recognition process, they have used two types of features' sets. In the primary features' set, the number of junctions, the number of loops and their position are used. In the secondary features' set, the number of endpoints and their location, the number of junctions and their location, and the nature of profiles of different directions are considered. A multi-stage classification scheme combined with binary tree and nearest neighbour classifier has been used for the purpose. The system has accuracy about 97.34%.

An OCR post-processor of Gurumukhi script has also been developed. Lehal and Singh [17] proposed a post-processor for Gurumukhi OCR, where statistical information of Punjabi language syllable combinations, corpora look-up and certain heuristics based on Punjabi grammar rules have been considered.

g-Gujrathi character recognition

To the best of our knowledge, only one paper reports on Gujrathi script. Antani and Agnihotri [1] describe the classification of a subset of printed digitized Gujrathi characters. For the classification, Euclidean minimum distance and K-nearest neighbour classifier were used with regular and invariant moments.

A Hamming distance classifier was also used. The recognition rate of the reported system is very low (about 67%).

h-Kannada character recognition

A few reports are available for Kannada characters recognition. A font and size independent OCR system for printed Kannada documents has been reported recently by Ashwin and Sastry[2]. The system first extracts words from the document image and then segments the words into sub-character level pieces. The segmentation algorithm is motivated by the structures of the script. A set of zoning features is extracted after the normalization of the characters for recognition. The final recognition is achieved by employing a number of 2-class classifiers based on the SVM (*Support Vector Machine*).

An on-line system for Kannada characters is described by Rao and Samuel [26]. The described system extracts Wavelet features from the contour of the characters. The convolutional feed forward multi-layer neural network is used as the classifier.

Conclusion

A brief survey of OCR activities in major Indian scripts is presented. We did not come across any work done in scripts other than those reported here. Perhaps some techniques of Arabic OCR system are usable for Urdu, since both scripts are similar in shape. Among the studies reported here, the systems on Devnagari and Bangla scripts are the most advanced and robust, followed by Tamil and Punjabi systems, which are less complex because of the absence of compound characters. Among the developed systems, some have been taken by software houses for commercialization. Indian Statistical Institute has transferred the largest number of technologies to organization like C-DAC, Pune (Devnagari and Bangla OCR), ER & DCI, Noida (Devnagari OCR), OCAC, Bhubaneswar (Oriya OCR) and IIT, Guwahati (Assamese OCR). C-DAC has already brought out the Devnagari OCR, named as Chitrankan, in the market. It is hoped that multilingual OCR systems in Indian languages will soon be made available.

References

1. S. Antani and Lalitha Agnihotri, "Gujrathi character recognition", In Proc. 5th Int. Conference on Document Analysis and Recognition, pp.418-421, 1999.

2. T V Ashwin and P S Sastry, "A font and size independent OCR system for printed Kannada documents using support vector machines", Sadhana, vol.27, pp.35-58, 2002.

3. V. Bansal and R M K. Sihna, "Integrating knowledge sources in Devnagari text recognition system", IEEE Transactions on Systems, Man, & Cybernetics Part A: Systems & Humans. Vol. 30, p 500-505, 2000.

4. U. Bhattacharya, T. K. Das, A. Datta, S. K. Parui and B. B. Chaudhuri, "Recognition of Handprinted Bangla Numerals using Neural Network Models", Advances in Soft Computing - AFSS 2002, Springer Verlag Lecture Notes on Artificial Intelligence, Eds. N. R. Pal and M. Sugeno, LNAI-2275,228-235, 2002.

5. M. Chandrasekaran, Machine recognition of the Tamil script, Ph.D dissertations, University of Madras, 1982.

6. B. B. Chaudhuri and U. Pal, 'Skew angle detection of digitized Indian Script documents', IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 19, pp.182-186, 1997.

7. B. B. Chaudhuri and U. Pal, 'An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi)', In Proc. 4th Int. conf. on Document Analysis and Recognition, pp. 1011-1016, 1997.

8. B. B. Chaudhuri and U. Pal, 'A complete printed Bangla OCR system', Pattern Recognition, vol. 31, pp. 531-549, 1998.

9. B. B. Chaudhuri, U. Pal and M. Mitra, "Automatic Recognition of Printed Oriya Script", Sadhana, (a journal of Indian Academy of Sciences) vol.27, part 1. pp.23-34, 2002.

10. P. Chinnuswamy and S. G. Krishnamoorty, "Recognition of hand-printed Tamil characters", Pattern Recognition, vol. 12, pp. 141-152, 1980.

11. A. K. Dutta and S. Chaudhuri, "Bengali alpha-numeric character recognition using curvature features", Pattern Recognition, vol. 26, pp. 1757-1770, 1993.

12. U. Garain, B. B. Chaudhuri 'Compound character recognition by run-number-based metric distances', *SPIE Proceedings* Vol. 3305 pp.90-97, 1996.

13. U. Garain and B. B. Chaudhuri, "Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts using Fuzzy Multifactorial Analysis", IEEE Transactions on Systems, Man and Cybernetics, Part B(2002). (accepted).

14. V. K. Govindan and A. P. Shivaprasad, "Character recognition - a survey", Pattern Recognition, vol. 23, pp. 671-683, 1990.

15. A K Goyal, G S Lehal and J Behal, Machine Printed Gurmukhi Script Character Recognition Using Neural Networks, Cognitive Systems Reviews and Previews, J. R. Isaac and K. Batra (Editors), Delhi, Phoenix Publishing House Pvt. Ltd, pp. 141-150, 1999.

16. G. S. Lehal and Chandan Singh, Feature extraction and classification for OCR of Gurmukhi script, Vivek, Vol. 12, No. 2, pp. 2-12, 1999.

17. G. S. Lehal and Chandan Singh, "A complete OCR system for Gurmukhi script, Structural, Syntactic and Statistical Pattern Recognition", T. Caelli, Amin, R.P.W. Duin, M. Kamel and D. de Ridder (Eds.), *Proceedings SSPR 2002*, Lecture Notes in Computer Science, Vol. 2396, Springer-Verlag, Germany, pp. 344-352, 2002.

18. S. Mohanti, "Pattern recognition in alphabets of Oriya Language using Kohonen Neural network", Int. Jour. of Pattern Recog. and Artificial Intelligence, vol.12, pp.1007-1015, 1998.

19. Atul Negi, Chakravarthy Bhagvati and B. Krishna, "An OCR system for Telugu", In Proc. 6th International Conference on Document Processing, pp. 1110-1114, 2001.
20. U. Pal, "On the development of an optical character recognition (OCR) system for printed Bangla script", Ph.D Thesis, 1997.
21. U. Pal and B. B. Chaudhuri, "Printed Devnagari script OCR system", Vivek, vol.10, pp.12-24, 1997.
22. U. Pal, P. K. Kundu and B. B. Chaudhuri "OCR error correction of an Inflectional Indian language using morphological parsing", Journal of Information Science and Engineering, Vol.. 16. No.6. pp. 903-922, 2000.
23. A.F.R. Rahman, R. Rahman and M.C. Fairhurst, "Recognition of Handwritten Bengali Characters: a novel multistage approach", Pattern Recognition, vol. 35, pp. 997-1006, 2002.
24. S. N. S. Rajasekaran and B. L. Deekshatulu, "Generation and recognition of Telugu characters", J. Inst. Electronics and Telecom. Engineering, vol. 4, 1974.
25. S. N. S. Rajasekaran and B. L. Deekshatulu, "Recognition of printed Telegu characters", Computer Graphics and Image Processing, vol. 6, pp. 335-360, 1977.
26. R. Srivivasa Rao and R. D. Sudhaker Samuel, "On-line character recognition for handwritten Kannada characters using Wavelet features and Neural classifier", IETE Journal of Research, vol.46, pp.387-392, 2000.
27. K. Ray and B. Chatterjee, "Design of a nearest neighbor classifier system for Bengali character recognition", J. Inst. Elec. Telecom. Engineering (India), vol. 30, pp. 226-229, 1984.
28. I. Sethi and B. Chatterjee, "Machine recognition of constrained hand-printed Devnagari numerals", J. Inst. Electronics and Telecom. Engineering, vol. 22, pp. 532-535, 1976.
29. I. Sethi and B. Chatterjee, "Machine recognition of constrained hand-printed Devnagari", Pattern Recognition, vol. 9, pp. 69-76, 1977.
30. R. M. K. Sinha, "Rule based contextual post-processing for Devnagari text recognition", Pattern Recognition, vol. 20, pp. 475-485, 1987.
31. R. M. K. Sinha, "A Syntactic pattern analysis system and its application to Devnagari script recognition", Ph.D Thesis, Electrical Engineering Dept., Indian Institute of Technology, India, 1973
32. R. M. K. Sinha and H. Mahabala, "Machine recognition of Devnagari script", IEEE Trans. on Systems, Man and Cybernetics, vol. 9, pp.435-441, 1979.
33. S. Palit and B. B. Chaudhuri, "A feature-based scheme for the machine recognition of printed Devanagari script", In Proc. Pattern Recognition, Image processing and Computer Vision, Ed. P. P. Das and B. N. Chatterjee, Narosa Publishing House, pp. 163-168, 1995.
34. U. Pal and B. B. Chaudhuri, "Printed Devnagari script OCR system", Vivek, vol.10, pp.12-24, 1997.
35. Shamik Sural and P. K. Das, "An MLP using Hough transform based fuzzy feature extraction for Bengali script recognition", Pattern Recognition Letters, vol.20, pp. 771-782, 1999.
36. G. Siromony, R. Chandrasekaran and M. Chandrasekaran, "Computer recognition of printed Tamil characters", Pattern Recognition, vol. 10, pp. 243-247, 1978.
37. Amal Som, "On Some nonparametric methods in Pattern Recognition", Ph.D Thesis, Jadavpur University, 1979.
38. Sukhaswami, P. Seetharamulu and A. K. Pujari, "Recognition of Telugu characters using Neural networks", International Journal of Neural systems, vol.6, no.3, pp. 317-357, 1995.
39. S. K. Parui, B. B. Chaudhuri & D. Dutta Majumder, "A procedure for recognition of connected handwritten numerals", Int. J. Systems Sciences., Vol. 13, No. 13, pp. 1019-1029, 1982.
40. U. Garain, B. B. Chaudhuri and T. T. Pal , "Online Handwritten Indian Script Recognition: A Human Motor Function based Framework", In Proc. 16th ICPR. vol.3, pp.164-167, 2002.
41. S. Sundaresan and S. S. Keerthi, "A study of representations for pen based handwriting recognition of Tamil characters", In Proc. 5th ICDAR, pp.422-423, 1999.
42. G. S. Lehal and Chandan Singh, "A Post Processor for Gurmukhi OCR", SADHANA Academy Proceedings in Engineering Sciences, Vol. 27, Part 1, pp. 99-111, 2002.

Dr. B. B. Chaudhuri
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute
203, B. T. Road
Kolkata - 700 108
E-mail: bbc@isical.ac.in
Fax: (91) (33) 577 23035
Tel.: (91) (33) 2578 1832 (direct)
(91) (33) 2577 8085 Ext. 2852

Machine Translation in India: a brief Survey

Durgesh Rao

Durgesh Rao is the founder of DR Systems, a company focused on creating innovative applications of language technology and IT. Earlier, he was a Research Scientist in the Knowledge Based Computing Systems group at the National Centre for Software Technology (NCST), Mumbai. There, he worked on MT, IR and other applications of NLP, and also designed and taught courses in Software Technology. Durgesh has an M.Tech in Computer Science from the Indian Institute of Technology (IIT), Bombay.

Disclaimer: This survey is based on the information available to the author, and is believed to be accurate at the time of writing (June 2002) to the best of his knowledge and belief. No legal claim is made to the accuracy of the information. For the latest information on these projects, the interested reader should write to the contact person mentioned for each project.

Background

Machine Translation (MT) is an important technology for localization, and is particularly relevant in a linguistically diverse country like India. In this document, we provide a brief survey of machine translation in India.

Human translation in India is a rich and ancient tradition. Works of philosophy, arts, mythology, religion, science and folklore have been translated among the ancient and modern Indian languages. Numerous classic works of ancient, medieval and modern art have also been translated from European into Indian languages, and vice-versa, since the 18th century. In the current era, human translation finds application mainly in the administration, media and education, and, to a lesser extent, in business, arts, and science and technology.

India has a linguistically rich area. It has 18 constitutional languages, which are written in 10 different scripts. Hindi is the official language of the Union. English is very widely used in the media, commerce, science and technology and education. Many of the states have their own regional language, which is either Hindi or one of the other constitutional languages. Only about 5% of the population speaks English. In such a situation, there is a big market for translation between English and the various Indian languages. Currently, the translation work is essentially manual. Use of automation is largely restricted to word processing. Two specific examples of high volume manual translation are translation of news from English into local languages, and translation of annual reports of government departments and public sector units among English, Hindi and the local language.

As is clear from above, the market is largest for translation from English into Indian languages, primarily Hindi. Hence, it is no surprise that a majority of the Indian machine translation systems are for English-Hindi translation. It is well known that natural language processing presents many challenges, of which the biggest is the inherent ambiguity of natural language. MT systems have to deal with ambiguity, and various other NL phenomena. In addition, the linguistic diversity between the source and target languages makes MT a bigger challenge. This is particularly true of widely divergent languages such as English and Indian languages. The major structural difference between English and Indian languages can be summarized as follows:

English is a highly positional language with rudimentary morphology, and default sentence structure as SVO. Indian languages are highly inflectional, with a rich morphology, relatively free word order, and default sentence structure as SOV. In addition, there are many stylistic differences. For example, it is common to see very long sentences in English, using abstract concepts as the subjects of sentences, and stringing several clauses together (as in this sentence!). Such constructions are not natural in Indian languages, and present major difficulties in producing good translations.

As recognized the world over, with the current state of art in MT, it is not possible to have fully automatic, high quality, and general-purpose machine translation. Practical systems need to handle ambiguity and other complexities of natural language processing, by relaxing one or more of the above dimensions.

Thus, we can have automatic high-quality 'sub-language' systems for specific domains, or automatic general-purpose systems giving rough translation, or interactive general-purpose systems with pre- or post-editing. As can be seen further in the document, Indian MT systems have also adopted one of these strategies.

Machine translation in India is relatively young. The earliest efforts date from the late 80s and early 90s. The prominent among these are the projects at IIT Kanpur, University of Hyderabad, NCST Mumbai and CDAC Pune. The Technology Development in Indian Languages (TDIL), an initiative of the Department of IT, Ministry of Communications and Information Technology, Government of India, has played an instrumental role by funding these projects.

Since the mid and late 90's, a few more projects have been initiated, at IIT Bombay, IIIT Hyderabad, AU-KBC Centre Chennai and Jadavpur University, Kolkata.

There are also a couple of efforts from the private sector, from Super Infosoft Pvt Ltd, and more recently, the IBM India Research Lab.

Major MT Projects in India

We now look at some of the major Indian MT projects in more detail. The parameters we look at are: language pair(s), formalism, strategy for handling complexity/ambiguity, and application domain(s), wherever this information is available (see the disclaimer at the top of this article).

a - Anglabharat (and Anubharati)

Anglabharati deals with machine translation from English to Indian languages, primarily Hindi, using a rule-based transfer approach. The primary strategy for handling ambiguity/complexity is post-editing; in case of ambiguity, the system retains all possible ambiguous constructs, and the user has to select the correct choices using a post-editing window to get the correct translation. The system's approach and lexicon is general-purpose, but has been applied mainly in the domain of public health. The project is primarily based at IITKanpur, in collaboration with ER&DCI, Noida, and has been funded by TDIL.

Anubharati is a recent project at IIT Kanpur, dealing with template-based machine translation from Hindi to English, using a variation of example-based machine translation. An early prototype has been developed and is being extended.

The contact person is Prof. RMK Sinha, <rmk@cse.iitk.ac.in>.

b - Anusaaraka

The focus in Anusaaraka is not mainly on machine translation, but on language access between Indian languages. Using principles of Paninian Grammar (PG), and exploiting the close similarity of Indian languages, an Anusaaraka essentially maps local word groups between the source and target languages. Where there are differences between the languages, the system introduces extra notation to preserve the information of the source language. Thus, the user needs some training to understand the output of the system. The project has developed language accessors from Punjabi, Bengali, Telugu, Kannada, and Marathi, into Hindi. The approach and lexicon is general, but the system has mainly been applied for children's stories. The project originated at IIT Kanpur, and later shifted mainly to CALTS (*Centre for Applied Linguistics and Translation Studies*), Department of Humanities and Social Sciences, University of Hyderabad. It was funded by TDIL.

Of late, LTRC (*Language Technology Research Centre*) at IIIT Hyderabad is attempting an English-Hindi Anusaaraka/MT system.

The contact persons are Prof. Rajeev Sangal, <sangal@iiit.net>, and Prof. G U Rao, <guraosh@uohyd.ernet.in>.

c - MaTra

MaTra is a human-assisted translation project for English to Indian languages, currently Hindi, essentially based on a transfer approach using a frame-like structured representation. The focus is on the innovative use of man-machine synergy; the user can visually inspect the analysis of the system, and provide disambiguation information using an intuitive GUI, allowing the system to produce a single correct translation. The system uses rule-bases and heuristics to resolve ambiguities to the possible extent – for example, a rule-base is used to map English prepositions into Hindi postpositions. The system can work in a fully automatic mode and produce rough translations for end users, but is primarily meant for translators, editors and content providers. Currently, it works for simple sentences, but work is on to extend the coverage to complex sentences. The MaTra lexicon and approach is general-purpose, but the system has been applied mainly in the domains of news, annual reports and technical phrases, and has been funded by TDIL.

The contact person is Durgesh Rao, <durgesh@ncst.ernet.in>, and the MaTra Team contact is, <matra@ncst.ernet.in>.

d - Mantra

The Mantra project is based on the TAG formalism from University of Pennsylvania. A sub-language English-Hindi MT system has been developed for the domain of gazette notifications pertaining to government appointments. In addition to translating the content, the system can also preserve the formatting of input Word documents across the translation. The Mantra approach is general, but the lexicon/grammar has been limited to the sub-language of the domain. Recently, work has been initiated on other language pairs such as Hindi-English and Hindi-Bengali, as well as on extending to the domain of parliament proceeding summaries. The project has been funded by TDIL, and later by the Department of Official Languages.

The contact person is Dr Hemant Darbari, <darbari@cdac.ernet.in>.

e - UCSG-based English-Kannada MT

The CS Department at the University of Hyderabad has worked on an English-Kannada MT system, using the UCSG (*Universal Clause Structure Grammar*) formalism, also invented there. This is essentially a transfer-based approach, and it has been applied to the domain of government circulars, and funded by the Karnataka government.

The contact person is Prof. K Narayana Murthy, <knmcs@uohyd.ernet.in>.

f - UNL-based MT between English, Hindi and Marathi

UNL, the *Universal Networking Language*, is an international project of the United Nations University, aiming at creating an Interlingua for all major human languages. IIT Bombay is the Indian participant in UNL. It is working on MT systems between English, Hindi and Marathi using the UNL formalism. This essentially uses an interlingual approach: the source language is converted into UNL using an '*enconverter*', and then converted into the target language using a '*deconverter*'.

The contact person is Prof. Pushpak Bhattacharya, <pb@cse.iitb.ac.in>.

g - Tamil-Hindi Anusaaraka and English-Tamil MT

The Anna University KB Chandrasekhar Research Centre at Chennai was established recently, and is active in the area of Tamil NLP. A Tamil-Hindi language accessor has been built using the Anusaaraka formalism described above. Recently, the group has begun to work on an English-Tamil MT system.

The contact person is Prof. CN Krishnan, <cnkrish@au-kbc.org>.

h - English-Hindi MAT for news sentence

The Jadavpur University at Kolkata has recently worked on a rule-based English-Hindi MAT for news sentences using the transfer approach.
The contact person is Prof. Sivaji Bandyopadhyay, <ilidju@cal2.vsnl.net.in>

i - Anuvadak English-Hindi software

Super Infosoft Pvt Ltd is one of the very few private sector efforts in MT in India. They have been working on a software called Anuvadak, which is a general-purpose English-Hindi translation tool that supports post-editing.

The contact person is Mrs. Anjali Rowchowdhury, <anjalir@del16.vsnl.net.in>

j - English-Hindi Statistical MT

The IBM India Research Lab at New Delhi has recently initiated work on statistical MT between English and Indian languages, building on IBM's existing work on statistical MT.

The url is:
http://www.research.ibm.com/irl/projects/translation.html

Conclusion

MT is relatively new in India – about a decade old. In comparison with MT efforts in Europe and Japan, which are at least 3 decades old, it would seem that Indian MT has a long way to go. However, this can also be an advantage, because Indian researchers can learn from the experience of their global counterparts. There are close to a dozen projects now, 6 of them being in advanced prototype or technology transfer stage, and the rest having been newly initiated.
So far, the Indian NLP/MT scene has been characterized by an acute scarcity of basic lexical resources such as corpora, MRDs, lexicons, thesauri and terminology banks. Also, the various MT groups have used different formalisms best suite to their specific applications, and hence there has been little sharing of resources among them. These issues are being addressed now. There are governmental as well as voluntary efforts under way to develop common lexical resources, and to create forums for consolidating and coordinating NLP and MT efforts. It appears that the exploratory phase of Indian MT is over, and the consolidation phase is about to begin, with the focus moving from proof-ofconcept prototypes to productionization, deployment, collaborative resource sharing and evaluation.

At a Glance: Summary of major MT projects in India

Project	Languages	Domain/ Main application	Approach/ Formalism	Strategy
Anglabharati	Eng-IL (Hindi)	General (health)	Transfer/rules	post-editing
Anusaaraka	IL-IL (5 IL-Hindi)	General (children)	LWG mapping/ PG	post-editing
MaTra	Eng-IL (Hindi)	General (news)	Transfer/frames	pre-editing
Mantra	Eng-IL (Hindi)	Govt. notifications	Transfer/XTAG	post-editing
UCSG MAT	Eng-IL (Kannada)	Govt. circulars	Transfer/UCSG	post-editing
UNL MT	Eng/IL (Hindi, Marathi)	General	Interlingua/ UNL	post-editing
Tamil Anusaa--raka	IL-IL (Tamil-Hindi)	General (children)	LWG mapping/ PG	post-editing
MAT	Eng-IL (Hindi)	News sentences	Transfer/rules	post-editing
Anuvadak	Eng-IL (Hindi)	General	N/A	post-editing
StatMT	Eng-IL	General	Statistical	post-editing

SCALLA - Sharing Capability in Localisation and Human Language Technologies

Pat Hall

SCALLA is a project funded by the EU under its Asia IT&C programme within EuropeAid - project ASI/B7-301/97/0126-05. The project partners in Europe are the Open University and Lancaster University in the UK, and ELRA in France; the project partners in India are the National Centre for Software Technology in Mumbai and the Indian Statistical Institute in Calcutta. Our objective is to help the development of language technologies and software localisation in both Europe and South Asia, through a series of conferences and other activities that will enhance collaboration between Europe and South Asia in this area. We have a particular interest in South Asian Languages, for their challenging differences from European languages, and because they are major world languages and important minority languages within Europe. Please visit the dedicated web site at: http://www.elda.fr/proj/scalla.html.

We launched the project in November 2001 with a conference in Bangalore, India. Seventeen leaders of language engineering within South Asia and seven experts from Europe attended, discussing at length the full range of language engineering issues from writing systems and software localisation through to translation and speech. It is clear that there is much potential for collaborations between South Asian research groups and groups in Europe, useful links were established, and have subsequently been built on. One excellent outcome triggered by discussions at the meeting was the development within NCST of a new input method for South Asian writing systems based upon the phonetic underpinnings of the scripts.

An area underrepresented at the Bangalore meeting was the pragmatic side of software localisation, and when the newly formed Indic Computing group arranged a meeting about this in September 2002 in Bangalore, we participated. Many localisation of Linux to Indic languages are underway, other applications are appearing, but there remain serious barriers such as availability of OpenType fonts.

Our second set of activities focussed on Europe. Originally, we had planned just a single event in Europe, but the richness conferences in Europe suggested that we facilitated participation from South Asia in a range of conferences. We had a small presence at LREC 2002, in Las Palmas, Canary Islands, Spain, in June 2002, and supported then two South Asian delegates at LangTech in Berlin, in late September 2002. Two other South Asian delegates could attend the localisation conference, LRC 2002, in Dublin in November. These visits have helped expose European language technologies to people from South Asia, and helped us within the project to better understand how we might expect language engineering to develop within South Asia. In Europe, there is a strong commercial and political imperative behind the research and development of language technologies, to help make Europe more cohesive. Could it be that such imperatives are not sufficiently strong in South Asia?

We are now about to hold our major meeting in Europe, a workshop at EACL 2003 in Budapest, supporting the travel of seven experts from South Asia to that meeting. See http://www.conferences.hu/EACL03/. The outcome of this workshop will be reported in a later newsletter.

Looking forward, our final activity will be back in South Asia, most likely in Calcutta, where we will support a general conference on language engineering. In this final conference, we must also pick up the social and cultural factors that are critically important for the development and deployment of language technologies, factors that were identified as important in that initial Bangalore conference, adding these to the commercial and political factors that seem to be pre-requisites for success.

The Asia IT&C programme includes several projects of interest to readers of this newsletter. Try looking at http://europa.eu.int/comm/europeaid/projects/asia-itc/html/main.htm

Professor Pat Hall
Computing Department, Open University
Milton Keynes MK7 6AA (UK)
Tel: 01908 652694 (work at OU)
01825 71 2661 (home and work)
07813 603376 (mobile)
Email: p.a.v.hall@btinternet.com

NEMLAR - Reaching out to the Mediterranean Countries

NEMLAR Consortium

NEMLAR (*Network for Euro-Mediterranean LAnguage Resources*) is a network project designed to consolidate knowledge about the state of art of Arabic and regional language resources, establish priority needs for industrial language technology organisations seeking to integrate Arabic and other languages into global networks, and support the development of basic resources for the partner countries and language forms.

The NEMLAR network covers recognised European centres and recognised partners in 6 of the Mediterranean countries covered by INCO-MED, namely Jordan, Morocco, Egypt, Lebanon, Tunisia, West Bank, Gaza Strip.

The Work Programme covers the following key tasks:

- Produce a comprehensive survey of organisations, people, projects and existing language resources for the project languages (forms of Arabic and other local languages where appropriate) and make the resulting information available as a web-based database.
- Produce, by consulting Europe and Mediterranean industry representatives in speech and text technologies, a survey of observed needs for language resources for the effective development of Arabic and local language systems, and establish on the basis of this survey a set of priorities for the development of Arabic and local language resources and tools.
- Establish a LR development programme for the region, based on the observed disparity between existing resources and

required priority resources, to develop a basic language resource kit covering all forms of Arabic and local languages in the region, and set realisable targets for the completion of these local tasks by NEMLAR partners, accompanied with training sessions to upgrade local personnel to LR management-readiness.
- Disseminate the results of surveys, analyses, evaluations and language resource development tasks to the human language technology community as a whole via the NEMLAR website. Hold an international conference on requirements and prospects for Arabic and other Mediterranean language resources.

NEMLAR is a joint initiative by CST (Copenhagen, coordinator), ELDA (Paris, technical coordinator) and ELSNET (Utrecht). You will find more information on, the dedicated web site, at http://www.nemlar.org.

This paper was prepared by the Consortium of the European NEMLAR project.

LangTech 2003

24-25 November 2003

Paris, Meridien Montparnasse Hotel

LANGTECH 2003 IS THE SECOND ANNUAL EDITION OF EUROPE'S FIRST DEDICATED FORUM FOR INDIVIDUAL AND ORGANISATIONS, INDUSTRIAL AND ACADEMIC, INVOLVED IN THE DEVELOPMENT, DEPLOYMENT AND EXPLOITATION OF SPOKEN AND WRITTEN LANGUAGE TECHNOLOGIES.

REGISTRATION IS NOW AVAILABLE!

A **DISCOUNT** IS APPLICABLE IF YOU REGISTER **BEFORE 1ST AUGUST 2003.**

	Before 1 August 2003	*Normal registration*	*On-site registration*
Industrials	525 Euro	625 Euro	675 Euro
Academics	325 Euro	375 Euro	425 Euro

THE EXHIBITION AND ATTENDEE REGISTRATION FORMS ARE AVAILABLE ON-LINE AT

HTTP://WWW.LANG-TECH.ORG

CONTACT:
Khalid Choukri/Magali Jeanmaire
55-57, rue Brillat Savarin
75013 Paris (France)
E-mail: langtech2003@elda.fr

New Resources

ELRA-S0144 Italian SpeechDat-Car

The Italian SpeechDat-Car database contains the recordings of 300 Italian speakers (149 females, 151 males) recorded over the GSM telephone network, in a car. The Italian SpeechDat-Car database was produced by Alcatel Business Systems and ITC-IRST, with the collaboration of Fiat Research Centre CRF. This database is partitioned into 14 DVDs. The speech data files are in two formats. Four of the 5 microphones were recorded on the computer in the boot of the car. The speech data are stored as sequences of 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the cell phone, and was recorded on a remote machine. The data are stored as sequences of 8 kHz 8 bit A-law. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
This speech databases was validated by SPEX (the Netherlands) to assess its compliance with the SpeechDat-Car format and content specifications.
Each speaker uttered the following items:
- 2 voice activation keywords
- 1 sequence of 10 isolated digits
- 7 connected digits (1 sheet number -4+ digits, 1 spontaneous telephone number -9/11 digits, 3 read telephone numbers, 1 credit card number -16 digits, 1 PIN code -6 digits)
- 3 dates (1 spontaneous date e.g. birthday, 1 prompted date, 1 relative or general date expression)
- 2 word spotting phrases using an embedded application word
- 4 isolated digits
- 7 spelled words (1 spontaneous e.g. own forename or surname, 1 directory city name, 4 real word/name, 1 artificial name for coverage)
- 1 money amount
- 1 natural number
- 7 directory assistance names (1 spontaneous e.g. own forename or surname, 1 city of birth/growing up, 2 most frequent cities, 2 most frequent company/agency, 1 "forename surname")
- 9 phonetically rich sentences
- 2 time phrases (1 spontaneous time of day, 1word style time phrase)
- 4 phonetically rich words
- 67 application words (13 mobile phone application words, 22 IVR function keywords, 32 car products keywords)
- 2 additional language dependent keywords
- Prompts for spontaneous sentences

The following age distribution has been obtained: 134 speakers are between 16 and 30, 117 speakers are between 31 and 45, 46 speakers are between 46 and 60, and 3 speakers are over 60.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	90,000 Euro	120,000 Euro
For commercial use	90,000 Euro	120,000 Euro

AURORA Project Databases

The Aurora project was originally set up to establish a world wide standard for the feature extraction software which forms the core of the front-end of a DSR (Distributed Speech Recognition) system. ETSI formally adopted this activity as work items 007 and 008. The two work items within ETSI are:
- ETSI DES/STQ WI007: Distributed Speech Recognition - Front-End Feature Extraction Algorithm & Compression Algorithm
- ETSI DES/STQ WI008: Distributed Speech Recognition - Advanced Feature Extraction Algorithm.

AURORA Project Database - Subset of SpeechDat-Car - Italian database (AURORA/CD0003-05)

This database is a subset of the Italian SpeechDat-Car database which has been collected as part of the European Union funded SpeechDat-Car project. It contains contains 2200 Italian connected digit utterances divided into training and testing utterances in the following noise and driving conditions inside a car: high speed good road, low speed rough road, stopped with motor running, town traffic.

	ELRA members	Non-members
For research use only	1,000 Euro	1,000 Euro

AURORA Project Database - Aurora 4a

The Aurora project is now releasing a number of list files for performing the training and testing on the Wall Street Journal (WSJ0) data at two sampling rates -8 kHz and 16 kHz. The Aurora 4a database is based on the WSJ0 with artificial addition of noise over a range of signal to noise ratios. It contains both clean and multicondition training sets and 14 evaluation sets with different noise types and microphones.

	ELRA members	Non-members
For research use by an academic organisation	250 Euro	250 Euro
For research use by a commercial organisation	1,000 Euro	1,000 Euro

AURORA Project Database - Aurora 4b

An additional database, Aurora 4b, will be released later, that will contain noisy versions of the Nov'92 WSJ0 development set.

	ELRA members	Non-members
For research use only	250 Euro	250 Euro

ELRA-L0049 SCIPER-FR

This French monolingual dictionary contains around 90,000 lemmas (approximately 700,000 inflected forms), with their part of speech and some information related to their inflexion. The data are encoded in UTF-8 XML.

	ELRA members	Non-members
For research use	1,200 Euro	2,500 Euro
For commercial use	8,000 Euro	11,000 Euro

ELRA-L0050 SCIPER-AN

This English monolingual dictionary contains around 110,000 lemmas (approximately 218,000 inflected forms), with their part of speech and some information related to their inflexion. The data are encoded in UTF-8 XML.

	ELRA members	Non-members
For research use	1,200 Euro	2,500 Euro
For commercial use	8,000 Euro	11,000 Euro

ELRA-L0052 SCIPER-ES

This Spanish monolingual dictionary contains around 27,500 lemmas (approximately 50,000 inflected forms), with their part of speech and some information related to their inflexion. The data are encoded in UTF-8 XML.

	ELRA members	Non-members
For research use	1,200 Euro	2,500 Euro
For commercial use	8,000 Euro	11,000 Euro

ELRA-M0033 SCI-FRAN

This bilingual dictionary contains around 120,000 pairs of French-English terms, with their part of speech. The data are encoded in UTF-8 XML.

	ELRA members	Non-members
For research use	1,500 Euro	3,000 Euro
For commercial use	10,000 Euro	14,000 Euro

ELRA-M0035 SCI-FRES

This bilingual dictionary contains around 80,000 pairs of French-Spanish terms, with their part of speech. The data are encoded in UTF-8 XML.

	ELRA members	Non-members
For research use	1,500 Euro	3,000 Euro
For commercial use	10,000 Euro	14,000 Euro

ELRA-M0037 SCI-ANES

This bilingual dictionary contains around 60,000 pairs of English-Spanish terms, with their part of speech. The data are encoded in UTF-8 XML.

	ELRA members	Non-members
For research use	1,500 Euro	3,000 Euro
For commercial use	10,000 Euro	14,000 Euro

Discounts - SCIPER Resources (monolingual and multilingual lexicons)

2 dictionaries	10% discount
3 dictionaries	20% discount
> 3 dictionaries	25% discount
2 monolingual dictionaries + +1 bilingual dictionary	20% discount