Language engineering - a champion for European culture

Conference or Workshop Item

Language Engineering - A Champion for European Culture

by

Dr. Neil Simpkins and John Diver MBCS

ABSTRACT

Language is key to culture. It is a direct cultural medium as well as a means of recording and providing access to non-lingual elements of culture. Language is also fundamental to a sense of cultural identity. For this reason, it is vital, in a changing Europe, that we preserve the multi-lingual character of our society in order to move successfully towards closer co-operation at a political, economic, and social level.

Language engineering is the application of knowledge of language to the development of computer software which can recognise, understand, interpret, and generate human language in all its forms.

The paper provides a high level view of the 'state of the art' in Language engineering and indicates ways in which it will have a profound impact on our culture in the future. It shows how advances in language engineering are an important aid in maintaining cultural diversity in a multi-lingual European society, while enabling the development of social cohesion across cultural and national divides. It addresses issues raised by the prospect of the Multi-lingual Information Society, including education, human communication with technology, and information management, as well as aspects of digital cities such as tele-presence in digital libraries, virtual art galleries, and electronic museums. The paper raises the issue of language as a factor in cultural domination, showing the contribution that language engineering can make towards countering it.

The paper also raises a number of controversial issues concerning the likely benefits arising from the ways in which language engineering is likely to influence the culture of Europe.

Introduction

The purpose of this paper is to show how language engineering is an important technology which will have a dramatic impact on the benefits to be delivered by the information revolution. Language engineering is defined, its essential components described, and some significant areas of its application outlined. The paper highlights the importance of language in a cultural context, and suggests ways in which language engineering will effect our daily lives and intercultural relationships. The paper concludes with a summary of the major benefits that the authors believe will result.

Language is the natural means of human communication. It the most effective way we have to express ourselves. We use language in a host of different ways and in many different types of interaction. For most of us language is fundamental to all aspects of our lives.

At the present time language is both the basis for communication and a barrier . In the main, it is only used in communications between humans, not between humans and machines, and then understanding is usually limited to those groups who share a common language. However , a change is beginning to take place which will revolutionise our use of language and greatly enhance the value of language in every aspect of communication. This change is the result of developments in language engineering.

Language engineering provides ways in which we can extend our use of language to make it a more effective tool. It is based on a vast amount of knowledge which has been developed through research, about the threory and structure of language and the way it is used, in practice. It uses language resources, such as electronic dictionaries and grammars, terminology banks and corpora, which have been accumulated over time. Research tells us the things we need to know about language and develops the models and techniques needed to represent, understand and manipulate it. Resources represent the knowledge base needed to recognise, validate, understand and manipulate language using the power of computers, and to train systems based on statistical models, such as those used in speech recognition. By applying our knowledge of language we can develop new ways to help solve problems across the political, social, and economic spectrum.

Language Engineering is a computer based technology which uses our knowledge of language to enhance our application of language in systems:

- improving the way we interface with machines for learning, work and leisure;
- analysing, selecting, using, and presenting information more effectively;
- providing facilities to improve communication through human language generation and translation.

Making Language work for us

The ability to develop our use of language is one of the keys to the future of a cohesive European society in the information age.

Successful developments in language engineering will enable us to:

- access information efficiently, focusing precisely on the information we need, avoiding the problem of being overloaded with information;
- talk to computer based systems, at home as well as at work, in our cars, and in public places where we need information or assistance;
- teach ourselves other languages and improve our use of our own, and doing it at our convenience: in our own time; at our own pace; and in our own place;
- do business efficiently over the telephone by interacting reliably and directly with voice operated computer systems, even instructing your PC or other computerised appliance to carry out transactions on your behalf;
- operate more effectively internationally, in business, in administration, in political activities and as citizens and consumers;
- provide a wider range of better services to the maximum number of fellow citizens, colleagues and customers.

The Importance of Language

General

Language is a means of effective, efficient communication. It is also a medium for recording information; in practice, the most convenient way of representing most of the information we need. Language is integral to our culture. It helps each of us to define our identity.

Communication

Language is the natural means of human communication. It the most effective way we have to express ourselves. We use language in a host of different ways: to explain complex ideas and concepts; to request and impart information, to manipulate; to persuade; to negotiate, to make our needs known; to express our feelings; to narrate stories; and to create beauty in poetry and prose.

Language is fundamental to all aspects of our lives and for most of this is natural. For many disabled people, however, the use of language naturally is either impossible or restricted. They need to be given ways to communicate which enable them to overcome their disability.

Language is frequently a barrier to communication. This is obviously the case when correspondents do not share a common language. It often happens when they think that they do but because of levels of competence, variations in usage, and cultural differences serious misunderstandings occur. In a world where instantaneous connection is possible to any part of the globe; where thousands of multi-lingual political and commercial transactions take place every day it is imperative that we try to ensure that we understand each other perfectly.

Power and Influence

Proficiency in the use of language is a source of power and influence. Lack of proficiency is exclusive. In a mono-lingual society this is a social and educational issue concerning individuals but in a multi-lingual, international environment this becomes a serious issue of political and economic power. It is often said that only one or two languages are needed for international activities in business, administration, and politics. To a degree this is true and it is today's reality. However, it could never be entirely satisfactory and should not be acceptable. The dominance of a few languages is conducive to an imbalance of political and economic power. It is also a poor use of resources because it reduces significantly the number of people who can participate effectively in any activity and this is bound to exclude valuable contributions. In addition, it leads to discontent: a feeling of not being valued; of exclusion. In time, such an approach is also likely to marginalise the languages which are not used so widely, reducing still further the scope of their usage and inevitably

diminishing the variety and richness of our culture. In Europe, it would adversely affect not only our feeling for national, regional, and cultural identities but also our sense of belonging to a truly European society, not just tolerant of its minorities but supportive of them. Inevitably, this limited language approach would be likely also to weaken our resistance to the influence of invasive popular culture from an outside source.

This restrictive approach to language use would limit the availability of a wide range of important new services and facilities by denying many people access to computers in their native language.

Identity

For each one of us, our own language is fundamental to our national and cultural identity, providing a link to our traditions as well as being the foundation of our education, work, and entertainment. In Europe, we have the benefit of a diversity of languages and cultures, which means that we have the opportunity to find out a great deal about different cultures and to learn from this knowledge. This should become one of the bases for a cohesive European society. We can only identify with the Europe Union if our own cultural roots are secure. Our mutual respect for and recognition of the value of each other's culture is imperative. For this reason multi-lingualism must remain a feature of the European way of life and we must explore ways in which to overcome the barriers to communication and understanding which this inevitably causes, for the time being.

Cultural Record

Much of our culture is expressed through the medium of language. Language is obviously integral in literature, drama, and opera, for example.

As a medium for recording our culture language is clearly just as important. The record of our law, the descriptions of political, commercial and judicial processes. All these are made accessible to each of us in our daily lives through written or spoken language.

Without language very little of our culture could be transmitted from one generation to the next and without records even less would reliably survive across further generations.

Increasingly, the records of our culture are being made available in multi-media form, delivered on CD ROMS or through telematics services, and presented through television screens, personal computers or specialised terminals. Language is still the most effective medium for indexing, classifying, and selecting this information.

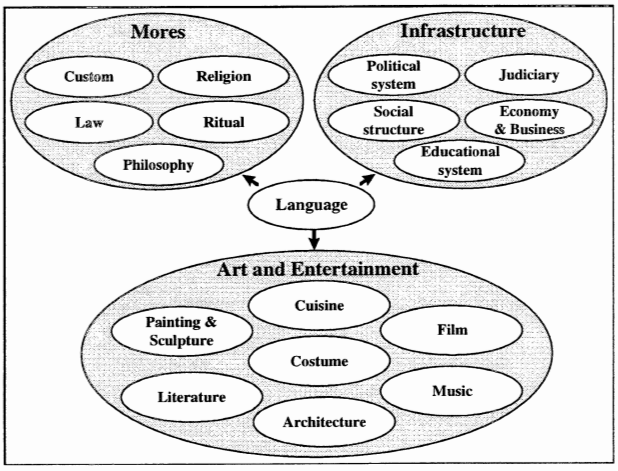

Figure 1 Elements of culture

What is Language Engineering ?

A Definition

Language engineering is the application of knowledge of language to the development of computer systems which can recognise, understand, interpret and generate human language in all its forms.

In practice, language engineering comprises a set of techniques and language resources. The former are implemented in computer software and the latter are a repository of knowledge which can be accessed by computer software.

Components of the Technology

The basic processes of language engineering are shown in Figure 1. These are broadly concerned with

- getting material into the computer, using speech, printed text or handwriting, or text either keyed in or introduced electronically;
- recognising the language of the material, distinguishing separate words, for example, recording it in symbolic form and validating it;
- building an understanding of the meaning of the material, to the appropriate level for the particular application;

- using this understanding in an application such as transformation (e.g. speech to text), information retrieval, or human language translation;
- generating the medium for presenting the results of the application;
- finally, presenting the results to human users via a display of some kind, a printer or a plotter, a loud speaker or the telephone.

Within this general model there are, of course, many different configurations, and depending on the applications for the technology not all these components are needed.

There are many techniques used in language engineering and some of these are described below.

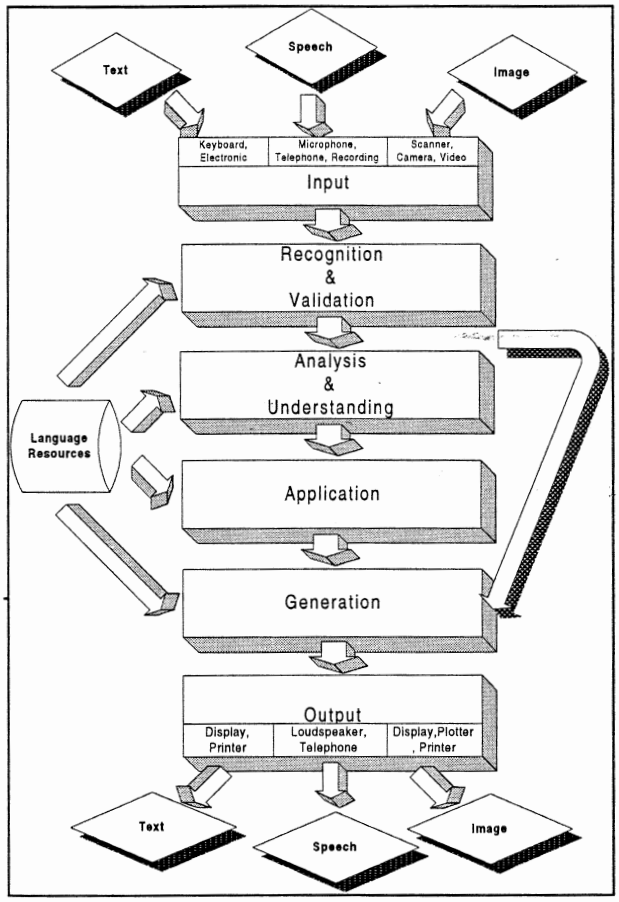

Figure 2 A model of language engineering

Speaker Identification and Verification

A human voice is as unique to an individual as a fingerprint. This makes it possible to identify a speaker and to use this identification as the basis for verifying that the individual is entitled to access a service or a resource. The types of problems which have to be overcome are, for example, recognising that the speech is not recorded, selecting the voice through noise (either in the environment or the transfer medium), and identifying correctly despite temporary changes (such as caused by illness).

Speech Recognition

The sound of speech is received by a computer in analogue wave forms which are analysed to identify the units of sound (phonemes) which make up words. Statistical models (Hidden Markov Models) of phonemes and words are used to recognise discrete or continuous speech input. The production of quality statistical models requires an extensive training sample and vast quantities of speech have been collected and continue to be collected for this purpose.

There are a number of significant problems to be overcome if speech is to become a commonly used medium for dealing with a computer. The first of these is the ability to recognise continuous speech rather than speech which is deliberately delivered by the speaker as a series of discrete words separated by a pause. The next is to recognise any speaker, avoiding the need to train the system to recognise the speech of a particular individual. Finally there is the serious problem of the noise which can interfere with recognition, either from the environment in which the speaker uses the system or through noise introduced by the transmission medium, as in the case of the telephones, for example. Noise reduction, signal enhancement and key word spotting can be used to allow accurate and robust recognition in noisy environments or over telecommunications.

Character and document image recognition

Recognition of written or printed language requires that a symbolic representation of the language is accurately derived from its spatial form of graphical marks. For most languages this means recognising and transforming characters. There are two cases of character recognition: recognition of printed images, referred to as Optical Character Recognition (OCR), and recognising handwriting, usually known as Intelligent Character Recognition (ICR).

OCR from a single printed font family can achieve a very high degree of accuracy. Problems arise when the font is unknown or very decorative or when the quality of the print is poor. In these

difficult cases, and in the case of handwriting, good results can only be achieved by using ICR. This involves word recognition techniques which use language models, such as lexicons or statistical information about word sequences.

Document image analysis is closely associated with character recognition but involves the analysis of the document to determine firstly its make-up in terms of graphics, photographs, separating lines, and text and then the structure of the text to identify headings, sub-headings, captions etc. in order to be able to process the text effectively.

Natural Language Understanding

The understanding of language is obviously fundamental to many applications. However perfect understanding is not always a requirement. In fact, gaining a partial understanding is often a very useful preliminary step in the process because it makes it possible to be intelligently selective about taking the depth of understanding to further level.

Shallow or partial analysis of texts is used to obtain a robust initial classification of unrestricted texts efficiently. This initial analysis can then be used, for example, to focus on 'interesting' parts of a text for a deeper semantic analysis which determines the content of the text within a limited domain. It can also be used, in conjunction with statistical and linguistic knowledge, to identify linguistic features of unknown words automatically, which can then be added to the system's knowledge.

Semantic models are used to represent the meaning of language in terms of concepts and relationships between them. A semantic model can be used for example, to map an information request to an underlying meaning which is independent of the actual terminology or language in which the query was expressed. This supports multilingual access to information without a need to be familiar with the actual terminology or structuring used to index the information.

Combinations of analysis and generation with a semantic model allows texts to be translated. At the current stage of development, applications where this can be achieved need be limited in vocabulary and concepts so that adequate language engineering resources can be specified. Templates for document structure as well as common phrases with variable parts can be used to aid in generation of a high quality text.

Natural Language Generation

A semantic representation of a text can be used as the basis for generating language. An interpretation of basic data or the underlying meaning of a sentence or phrase can be mapped into a surface string in a selected fashion; either in a chosen language or according to stylistic specifications by a text planning system.

Speech Generation

Speech is generated from filled templates, by playing 'canned recordings' or concatenating units of speech (phonemes, words) together. Speech generated has to account for aspects such as intensity, duration and stress in order to produce a continuous and natural response.

Dialogue can be established by combining speech recognition with simple generation, either from concatenation of stored human speech components or synthesising speech using rules.

Providing a library of speech recognizers and generators together with a graphical tool for structuring their application allows some-one who is neither a speech expert nor a computer programmer to design a structured dialogue which can be used for example in automated handling of telephone calls.

Language Resources

Language resources are essential components of language engineering. They are one of the main ways of representing the knowledge of language which is used for the analytical work leading to recognition and understanding.

The work of producing and maintaining language resources is a huge task. Resources are produced, according to standard formats and protocols to enable access, in many EU languages, by resource production projects. These resources are made available through the European Language Resource Association (ELRA).

Lexicons

A lexicon is a repository of words and knowledge about those words. This knowledge may include details of the grammatical structure of each word (morphology), the sound structure (phonology),

the meaning of the word in different textual contexts, e.g. depending on the word or punctuation mark before or after it. A useful lexicon may have hundreds of thousands of entries. Lexicons are needed for every language of application.

Specialist Lexicons

There are a number of special cases which are usually researched and produced with separately from general purpose lexicons:

- *Proper names* - Dictionaries of proper names are essential to effective understanding of language, at least so that they can be recognised within their context as places, objects, or person, or maybe animals. They take on a special significance in many applications, however, where the name is key to the application such as in a voice operated navigation system or in a holiday reservations system or railway timetable information system based on automated call handling.
- *Terminology* - In today's complex technological environment there are a host of terminologies which need to be recorded, structured, and made available for language engineering applications. Many of the most cost-effective applications of language engineering, such as the technical document management and translation, depend on the availability of the appropriate terminology banks.
- *Wordnets* - A wordnet describes the relationships between words, for example, synonyms, antonyms, collective nouns, and so on. These can be invaluable in such applications as information retrieval, translator workbenches, and intelligent office automation facilities for authoring.

Grammars

A grammar describes the structure of a language at different levels: word (morphological grammar), phrase, sentence, etc.. A grammar can deal with structure both in terms of surface (syntax) and meaning (semantics and discourse).

Corpora

A corpus is a body of language, either text or speech, which provides the basis for:

- analysis of language to establish its characteristics;
- training a machine, usually to adapt its behaviour to particular circumstances;
- verifying empirically a theory concerning language;
- a test set for a language engineering technique or application to establish that it works in practice;

There are national corpora of hundreds of millions of words but there are also corpora which are constructed for particular purposes like, for example, a corpus comprising recordings of car drivers speaking to a simulation of a speech operated control system, designed to help establish the user requirements for a real control system.

The application of language engineering

In practice, language engineering is applied at two levels. At the first level there are a number of generic classes of application:

- human language translation;
- information management (multi-lingual);
- authoring (multi-lingual);
- human/machine interface (multi-lingual voice and text);

At the second level, the first level of applications are in turn applied to real world problems across the political, social and economic spectrum. So, for example, information management can be used in an information service as the basis for analysing requests for information and matching the request against a database of text or images to select the information accurately.

Authoring tools are typically used in word processing systems but can also be used to generate text such as business letters in foreign languages, as well as in conjunction with information management to provide document management facilities.

Human language translation is currently used to provide translator workbenches and automatic translation in limited domains.

Most applications can usefully be provided with natural language user interfaces, including speech, to improve there usability.

The Impact of Language Engineering

Language technologies can be applied to a wide range of problems in business and administration to produce better, more effective solutions. They can also be used in education, to help the disabled, and to bring new services both to organisations and to consumers. There are a number of areas where the impact is significant:

- accessibility and participation;
- improved educational opportunities;
- better information;
- effective communications;
- entertainment, leisure and creativity.

Accessibility and Participation

One of the most important ways in which language engineering will have a significant impact is in the use of natural language, especially speech, to interface to machines. This improves the useability of systems and services. It will also help to ensure that services can be used not just by the computer literate but by ordinary citizens without special training. This aspect of accessibility is fundamental to a democratic, open and equitable society in the Information Age. A good example of the type of service which will be available is an automated legal advice service. The accessibility of the justice system to all citizens is becoming a serious problem in many societies where the cost of legal expertise and the process of law prevents all but the very rich and those qualifying for legal aid from exercising there legal rights. It will be possible using language based techniques not only to provide advice which is based on an understanding of the problem and an analysis of the relevant body of law but also to understand a natural language, probably verbal, description of the problem and deliver the advice, as a human lawyer would have done, in spoken or printed form. Such a service could be made available through kiosks in court buildings or post offices, for example. This type of application can also be used to inform citizens of social security entitlements and job opportunities, as well as providing a useable, comprehensible interface to more open government.

Systems with the capacity to communicate with their users interactively, through human language, available either through access points in public places or in the home, via the telephone network or TV cables, will make it possible to change the nature of our democracy. There will be a potential for participation in the decision making process through a far greater availability of information in understandable and 'objective' form and through opinion gathering on a very large scale.

Improved Education Opportunities

Distance learning has become an important part of the provision of education services. It is especially important to the concept of 'life-long learning' which is expected to become an important feature of life in the Information Age. Effective distance learning depends on telematics services and on the use of computer aided learning.

The quality and success of computer aided learning can be greatly enhanced by the use of language engineering techniques. If the computer aided learning package can understand the answers of students to questions, rather than simply recognise that the answer is right or wrong, students can be directed down a path which is more appropriate to their needs. In this way, students are likely to learn more effectively and to have a longer concentration span because a more sensitive package is inherently more comfortable to work with.

In future, in Europe, it will be essential in many walks of life to be cometent in more than one language. Of course, computer aided language learning (CALL) is an area of prime importance for the application of language engineering. The same knowledge that is essential to the ability of the machine to understand is also the basis for the interactive teaching process, providing quality diagnostics of student errors as well as illustrating correct usage.

Better Information

One of the key features of an information service is its ability to deliver information which meets the immediate, real needs of its client in a focused way. It is not sufficient to provide information which is broadly in the category requested, in such a way that the client must sift through it to extract what is useful. Equally, if the way that the information is extracted leads to important omissions then the results are at best inadequate and at worst they could be seriously misleading.

Information is also available throughout the world, on the World Wide Web, for example, in different languages In reality, however, it is only available to a client who can firstly request the information

in the language in which it is recorded and then understand the language in which the information is presented. Using automatic machine translation facilities the person seeking information will be able to complete an information request in his or her native language and receive the information in that same language, regardless of the language in which the information is recorded.

Language engineering can improve the quality of information services through the understanding and analysis techniques of semantic modelling, which are far superior to conventional indexing techniques, as well as through machine translation.

One of the major, direct benefits of the Information Society for the ordinary citizen will be the improvement in public service information. However, the wide accessibility of this information will dependent upon language engineering. People who are not familiar with the conventional user interface of a computer system will be able to request information by voice and the system will guide them through the possibilities. Those who want information about other countries, which may be held in a foreign language, will be able receive it in their own language. A good example of this is a service which is currently being developed by a project called TREE. This service will provide information about job opportunities across the European Union in the native language of the potential applicant. Obviously these are jobs where language skills are not significant. The service will be available on the Internet and it is also planned to have public booths where job seekers can use the service. In a monolingual pilot service run in Flanders, a surprising 26% of applications for jobs were received from applicants who had seen the details on the Internet.

Language engineering will make a contribution in a large number of public interest areas. Intelligence gathering for law enforcement is an interesting case. In detecting smuggling for example, there is a large amount of information available from public or commercial sources which, if collated and presented in the right way can give clear indications of suspicious activity. Details about ship movements, manifests, and company information can highlight abnormal profiles of activity. The ability of language based analysis to produce these profiles is an important aid.

Effective Communication

Communication is probably the most obvious use of language. On the other hand, language is also the most obvious barrier to communication. Across cultures and between nations problems arise all the time, not only because of the problem of translating accurately from one language to another but also because of the cultural connotations of word and phrases. A typical example in the European context is the word 'federal' which can mean a devolved form of government to someone who already lives in a federation but to someone living in a unitary sovereign state it is likely to mean the imposition of another level of more remote, centralised government. As the application of language knowledge enables us to support translators more effectively, with electronic dictionaries, thesauri, and other language resources, and eventually when machine translation becomes a reality so the barriers will be lowered. Maintaining a multi-lingual society will become a greater benefit than disadvantage. Agreements at all levels, whether political or commercial, will be better drafted more quickly in a variety of languages. International working will become more effective with a far wider range of individuals able to contribute.

An example of a project which is successfully helping to improve communications in Europe is Linguanet, a project which interconnects many of the police forces of northern Europe using a limited, controlled language which can be automatically translated, in real-time.

Entertainment, Leisure, and Creativity

The attraction of computer games to our children is a clear indication of the potential of the computer to affect our culture. Home entertainment can become more educational, while education can become more attractive, 'edutainment' as it has become known. The possibility of tele-presence in virtual environments such as museums, art galleries, and libraries will provide a rich cultural environment, available to a wide section of society in the comfort and convenience of their own homes. Virtual visits to such cultural stores will be aided by language technology enabling the research and selection of all forms of digitised, language based records, indexing and retrieval of images, dubbing and sub-titling of films, and providing translation of library and archive material.

For a wider range of people, writing can become a more exciting activity. Authoring tools will make it possible for them to achieve much higher quality results. The use of on-line dictionaries, and thesauri, for example, makes selection of the 'mot juste' more likely, and grammar can be checked. The result can be a far more satisfying experience for writers who are not naturally gifted or well educated but who want to express themselves effectively in their business or social correspondence.

Conclusions

Language engineering and culture

The impact of the application of language engineering on our culture is likely to be both far ranging and radical. Even where language is not the medium of a particular aspect of culture it is still the medium of our access to it and for the transmission of culture from one generation to the next. The greatest beneficial impact will be in the wider dissemination of information about contemporary events, in the availability and accessibility of services, such as professional advice services, and in the accessibility of the cultural inventory. Of course, there are likely to be dangers: less reason to leave home; potentially more interaction between humans and machines than normal social intercourse; perhaps, a significant substitution of virtual for real life.

A cohesive Europe of cultures

To embrace membership of a group such as the European Union, enthusiastically, people need to develop mutual respect for the other members and a genuine sense of belonging. Unfortunately, there are many examples where cultural or ethnic groups fit uneasily into larger, more formal political groupings: state, nation state and federation or union. The successful future of the European Union, if not Europe as a whole, depends on its ability to develop cohesion not only across national borders but also across cultures. The future of Europe then may depend upon its ability to retain its natural diversity and also to assimilate the cultures of its immigrant populations, to demonstrate that each component culture is valued and to actively support aspects of culture, such as language, which are at risk. As language is strongly linked to culture, and in many cases is a marker for a culture, the development of language knowledge and its application to protecting the continued use of individual languages must be a priority. Language engineering offers the possibility to achieve this and to do so in a most productive manner by making it possible to assimilate all European languages into the transactions of every aspect of modern living.

An Economic Perspective

Europe's position as a naturally multi-lingual community in a multi-lingual world can be used to our commercial advantage. As we endeavour to collaborate more closely, to develop the single market as our home market, we have a special incentive to develop solutions to the problems of a multi-lingual market place. In successfully supporting our own language needs, especially in business and administration , language engineering will help us to compete for business in the global marketplace. On the one hand, our businesses will have a competitive edge through their experience in using technology to service the needs of a multi-lingual marketplace. On the other hand, we shall also have language products to sell to the rest of the world.

Language engineering will improve the performance of business and administration. Products which are developed using language technology will revolutionise our systems and enhance the range of services available to business, government, and the public at large.

Bibliography

European Commission (1995) The Multi-lingual Information Society; Communication from the Commission to the Council for a decision COM(95) 486 final. Brussels

Edwards, J. (1985) Language Society and Identity. Oxford

Fishman, J. (1977) Language and Ethnicity. In: Gite H.(ed) Language ethnicity and inter-group relations. London

Saami Language as a marker of Ethnic Identity among Saami. Mikael Svonni. In: I Seurujärvi-Kari & U-M Kulonen (ed) Essays on Identity and rights. (1995) Helsinki

Language and Power. Norman Fairclough, (1989) New York ISBN 0 582 03133-8 CSD

Survey of the State of the Art in Human Language Technology. Giovanni Varile & Antonio Zampolli (ed) 1995