LANGUAGE RESOURCES, TELRI AND MULTILINGUAL LEXICAL SEMANTICS

Wolfgang Teubert
Institut für deutsche Sprache
Postfach 10 16 21, D-68016 Mannheim
Tel/Fax +0049-621-1581-415
wolfgang.teubert@ids-mannheim.de

ABSTRACT

In the emerging European civil society, all citizens must be able to communicate freely and easily with each other and with the public institutions serving them, without being restricted by language barriers. There must be more and better language instruction, and there must be a multilingual language technology helping everyone to retrieve information regardless of the source language, to translate and to write texts in foreign languages. Most classical machine translation systems and other translation aids use a concept-based approach developed in AI research on a cognitive linguistics foundation. This approach works for controlled, that is, quasi-formal languages, but not for general language. There we need an abundance of language data extracted from monolingual and multilingual resources and processed for translations platforms. Instead of language-independent concepts, these platforms work with translation units and their equivalents, as they can be found in parallel corpora. TELRI brings together focal research centers in Europe and promotes multilingual research, creation of such resources and of corpus-derived linguistic knowledge.

1 CIVIL SOCIETY IN MULTILINGUAL EUROPE

The Slovenian language prospered in spite of many obstacles and became consolidated under the domination of the Habsburg Empire, where German was given priority as the language of administration and also in other fields. Slovenian continued to prosper during the Yugoslavian Kingdom and later during the Republic of Yugoslavia in spite of the pre-eminence attributed to Serbo-Croatian. Now that Slovenia is independent at last, will the Slovenian language withstand the onslaught of English, the global *lingua franca*? Will Slovenian have equal status among the national languages spoken in Europe, together with its more than thirty counterparts?

Or will Slovenia, along with many other smaller countries, defer her claim to use her national language in the multilingual dialogue between European countries? Will the European discourse take place in English? Most European conferences do. English has become the *lingua franca* in Europe, in science and research as well as in trade and industry. In the arts, we still find some insistence on cultural and linguistic identity by people who are not afraid to be reproached for provincialism. But news wire services sell their news abroad increasingly in English, movies produced for the European or world market are often produced in English, and the WWW, it seems, uses languages other than English only to encrypt information, to keep it hidden from the rest of the world.

On the other hand, there are strong reasons to believe that all national languages in Europe (and most of the languages enjoying regional autonomy) will survive. In spite of the unchallenged rate of English as the global *lingua franca*, people whose mother tongue is not English rarely read English newspapers or books, and very few watch English TV programmes. English is used in professional contexts, is used abroad mostly for travel arrangements, and is used increasingly for information retrieval from the internet. But even organized tourism is primarily monolingual. Hotels located in frequented vacation resorts focus on clients from a particular language area. The quality of foreign language instruction that the majority of people are subjected to is still so poor that we do not have to be worried about the future of natural languages.

What is more worrisome is that the majority of people cannot fully participate in the European civil society because they have neither the necessary language skills nor the technical aids to overcome these deficiencies. What could be a solution to this problem? Should we opt for the American tradition and choose English as the sole European language? This would probably lead to a situation where English becomes the language of the upper classes or the elites, while national languages become reduced in status to the role of local vernaculars. From a democratic point of view, this seems to be a questionable aspiration. But are there alternatives?

Europe is larger than the European Union. The countries of Central and Eastern Europe and also the Newly Independent States (the former Soviet Union countries) are a genuine and essential part of Europe. A common European identity goes hand in hand with a pan-European common market. Europe stands for diversity in unity. By choice and tradition, Europe's identity is founded on cultural and linguistic plurality. The new, united Europe that is gradually evolving is strongly determined to

cherish this tradition. What we are hoping to experience is the emergence of an integrated and, at the same time, multilingual society.

Therefore, we must promote skills to overcome language barriers, not just for the select few, but for everyone. We must provide access to information, regardless of its source language, and we must enable people to communicate with each other, regardless of their native languages.

2 COPING WITH A MULTILINGUAL EUROPE

What can we do to overcome language barriers? More extensive and improved foreign language instruction is the most important task. There is no reason not to begin foreign language instruction in kindergarten. The later a language is introduced, the more difficult it becomes to acquire it. Also, there is no reason why English should be the first language. Indeed, almost everyone will be exposed to English to some degree and, therefore, will somehow learn it. So the first foreign language could be a language spoken in one of the neighboring countries. And the third foreign language could be one of the more prevalent European languages with regional standing, like French, Russian, and maybe German. Three foreign languages is not a utopian goal. Many people in Switzerland, in the Netherlands, and in Belgium speak three foreign languages to some degree and shame barely bilingual linguists like myself.

But we must also accommodate ourselves to the fact that teaching foreign languages is not sufficient to overcome all language barriers in Europe. Even extensive language training will not ensure that I can extract all the information I am looking for from foreign language texts. It will not necessarily let me properly translate texts into the target language, and I will still have difficulties writing texts in a foreign language. Translators and interpreters certainly are members of one of the oldest professions and will remain indispensable. But the average private citizen usually cannot afford their services. Therefore, the only alternative are technical aids, as they have become available with the advent of the computer. If we had operational machine translation (and even interpretation) systems, language barriers would cease to be obstacles for communication.

3 LIMITATIONS OF LANGUAGE PROCESSING

As it turned out, however, there are numerous applications for which machine translation is not very effective. It works best with text types such as highly specialized and formalized technical language, maintenance manuals, for instance, or, to mention an early successful example, weather reports. Here we increasingly find what is called controlled language: controlled language supposedly eliminates the fuzziness and ambiguity which is an essential feature of natural languages. Controlled languages presuppose restricted domains, a high degree of unambiguous terminology, and not more than one coordinated structure per sentence. Ideally, and to some extent also in reality, documents written in controlled language can be mechanically processed like strings of formal language, without human intervention. General language cannot be reduced to controlled documentation language. Whenever we talk about social constructs, education, the arts, administration, the judicial system, and also when we talk about scientific innovation, the creation of new paradigms, and new research topics, it becomes impossible to reduce this language to a kind of formal algorithm that can be processed, that is, translated, automatically. (For an unbiased view of the predicament of machine translation, cf. Melby 95) To translate a text written in natural language, a human agent is always needed, someone who understands the meaning of the source text and who can judge if the target text indeed conveys this meaning. All of us who have tried know that it is impossible even with the best dictionaries to translate a text one does not understand. Computers do not understand. The world of semantics is closed to them. Meaning cannot be processed; meaning can be understood, summarized, interpreted, paraphrased, and translated. These are actions, not processes, and an essential feature of actions is intentionality, something computers do not have. (cf. Dennett 91, Searle 92) Computers can process syntactically well-formed strings according to algorithms, to algorithmically formulated grammar rules. They can turn a declarative sentence into a question, an active sentence into a passive one, a singular into a plural, but they will never understand their meanings. Only humans can do that. This is why machine translation will not solve the problems of a multilingual Europe.

4 THE NEED FOR LANGUAGE DATA

Perhaps, then, we should be less ambitious. We should develop translation aids for human translators, professionals who usually know source and target language quite well, and translations aids for those users who know only one of the languages really well (because it is their native language) and who have a restricted knowledge of the other language. These translation aids have to tackle real natural language, not with controlled documentation language. Their main objective is to offer choices for the most probable translation equivalents to the users. Recent experience has shown that we still do not know enough about language, in general, or about lexical semantics, in particular, to develop really powerful translation aids. More monolingual, bilingual, and multilingual language data are needed and have to be processed to become the kind of linguistic "knowledge"

computers can deal with. This knowledge then will make translation platforms work. It will have to be represented in a way that computers can handle it, and this is different from the way lexical knowledge is represented in dictionaries. Dictionaries were compiled for human users who can understand what a word means, and they can make the right choice of the options presented to them because they understand what they mean. Computers lack this faculty. They have to base a hierarchy of options for a given translation unit on a few rather unreliable rules, on a fair amount of statistics, and mostly on huge lists of lexical units in context, that is, lexical units and the other lexical units they co-occur with.

Language data of this kind are available only since the advent of corpora. They were never collected because human users do not need them. The lack of language data for computers, of lexical data, in particular, became apparent in the second half of the eighties, when projects on machine translation and also on machine-aided translation had failed due to the proverbial fuzziness and ambiguity of lexical units. Word sense disambiguation was needed in real language, and word sense disambiguation needed processed word lists and heaps of statistical data.

The first language corpora were compiled in the sixties, and they quickly became the favorite playground of empirical linguistics. These corpora were usually rather modest in size but large enough to account for most grammatical phenomena and to provide a few citations for the more common usages of the basic vocabularies. They certainly improved the quality of dictionaries for human users. But, in the late eighties, they turned out to be much too small for the kind of lexical data needed by translation tools. And, even then, for many languages, more and less prevalent ones, they were no corpora at all.

Early on, the European Commission acknowledged that the development of multilingual language technology was pivotal for a multilingual information society in Europe. It realized that machine translation provided an answer to only a very limited segment of the translation tasks and that it was necessary to encourage research on translation aids, aids for real natural language. When it became clear that the lack of adequate language data was the principal problem, the Commission began to encourage the creation of language resources. Language resources are corpora and other lexical data: electronic versions of dictionaries for human users and lexicons for language technology applications. A feasibility study, the *Network of European Reference Corpora,* discussed relevant aspects like user needs, corpus composition and size, text representation, and text annotation. (cf. Calzolari 96) The first PAROLE project provided a blueprint for corpus building, and in the second PAROLE project, comparable corpora of 20 million words each were set up for 12 European Union languages. The PAROLE project also produced uniform, standardized and comparable lexicons of 20000 words each for the languages involved. But, to a great extent, these lexical data were not extracted from corpora, and the lexical entries do not deal with the semantics of the lemmata. Unfortunately, this is true for the majority of lexical resources even today, and for all multilingual lexicons. There are a few corpus-based monolingual dictionaries, notably COBUILD, but no bilingual ones. (cf. COBUILD 87) The art to extract lexical data from corpora, including lexical semantics, and to generate monolingual, but also of bilingual and multilingual lexicons for computational purposes is still in its infancy.

5 TELRI: THE TRANS-EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE

For most Western European languages, we now have at least the textual resources that we need for development of the art of further lexical data extraction. This is not quite the case for the languages spoken in Central and Eastern Europe and in the former Soviet Union. Because the European Commission strives to promote pan-European co-operation in focal research and development areas, it decided back in 1994 to fund the Concerted Action TELRI, the Trans-European Language Resources Infrastructure. The overall objective of this project was to increase the awareness of language resources as the raw material for a new generation of sophisticated multilingual language technology. The project started in early 1995. (TELRI Final Report 98) Today, TELRI II, the follow-up project, is about to commence, again funded by the European Commission up to the end of 2001.

TELRI brings together more than forty renowned national language research and technology centers in the whole of Europe, two of them in Slovenia. One is the Josef-Stepan-Institute organizing this conference, with its impressive record in computational linguistics. The other is the Institute for the Slovenian Language "Fran Ramovs" in the Slovenian Academy for Sciences and Arts, the heart of Slovenian dictionary making. TELRI is a pan-European network promoting the creation of language resources and the research on the methodology for extracting language data from corpora and for processing them to be used in multilingual language applications. It is devoted to make language resources available to the academic and industrial sectors of the natural language processing community, and to disseminate the results of theoretical discussions and practical work in Europe and worldwide.

TELRI partners share an impressive record in corpus-based general language research, in developing techniques for processing language data, and in designing data-driven software for monolingual and multilingual applications. TELRI pools the resources, tools, specific

expertise, and general competence of its partners and of cooperating institutions all over the world.

6 THE TELRI AGENDA

TELRI offers:

- Annual seminars on multilingual issues
- Training in corpus linguistics
- Service for language industry
- Dissemination of linguistic data, knowledge, and ideas
- TRACTOR: the TELRI Research Archive of Computational Tools and Resources
- Internet information services and printed publications

♦ TRACTOR

TRACTOR is the TELRI Research Archive of Computational Tools and Resources. It features monolingual, bilingual, and multilingual corpora and lexica in these languages. Furthermore, TRACTOR offers a large variety of corpus- and lexicon-related software, for example, 3,000 Kb Bulgarian, 360,000 Kb Czech, 15,000 Kb Dutch, 49,400 Kb English, 30,000 Kb Estonian, 90,000 Kb French, 500,000 Kb German, 800,000 Kb Hungarian (NB includes much digital sound data), 700 Kb Icelandic, 30,000 Kb Italian, 130,000 Kb Latvian, 5,000 Kb Lithuanian, 700 Kb Norwegian, 40,000 Kb Rumanian, 37,000 Kb Russian, 60,000 Kb Serbian, 30,000 Kb Slovenian, 27,000 Kb Spanish, 25,000 Kb Swedish, 40,000 Kb Turkish, 300 Kb Uzbek. While the primary aim is to pool the resources of TELRI partners, TRACTOR also serves other institutions by making their resources and tools available.

While TRACTOR has been set up for the benefit of TELRI partners, the archive can also be accessed by the members of the TRACTOR User Community (TUC). Individuals and every academic, public, or industrial organization can join the TUC for a nominal annual fee of 50 ECU (Western European members) or 20 ECU (rest of Europe). Comparable conditions apply for members from outside Europe. Tools and resources can be used free of charge for research purposes only. For more information on TRACTOR, see: http://solaris3.ids-mannheim.de/~tractor. If you are interested in joining TUC, contact the TELRI office at telri@ids-mannheim.de.

For the next three years, TRACTOR will have top priority on the TELRI agenda. The spectre of resources and tools will be quickly expanded.

♦ TELRI Research Projects and Joint Ventures

TELRI is the framework for multilateral projects and for joint ventures with industrial partners. TELRI closely interacts with the European Commission and other public or private funding authorities to explore the feasibility of new projects. Two examples:

The Plato Project

More than 20 TELRI partners brought together and aligned 11 versions of Plato's *Republic* in a unique parallel corpus. In workshops and seminars, multiple approaches were explored for extracting multilingual lexical data and for testing hypotheses of translation equivalence. The aligned parallel texts and relevant query tools are available on a TELRI CD-ROM. For more information of the Plato Project, see: http://www.ids-mannheim.de/telri/cdrom.html.

The Bridge Joint Venture

More than ten TELRI partners set up a consortium with publishers with the objective to localize and globalize the *English COBUILD Student Dictionary*. A plentitude of bilingual versions (English/local language) will generate a multilingual lexicon to be used for the automatic generation of bilingual dictionaries of lesser-used languages (e.g., Estonian/Slovenian). For the localized version, our partners have contracted local dictionary publishers. For more information, see: http://www.ids-mannheim.de/telri/bridge.html.

♦ Partnership with Industry

TELRI offers comprehensive service, expert consultation, application-oriented language data, and customized software solutions with partners in many low-wage countries able to offer competitive rates. It will guarantee the quality of its service. With its global links, TELRI is the ideal partner of small- and medium-sized enterprises in language industry, both for the development of new applications and for the localization and globalization of existing software. For more information, contact: TELRI Association c/o IDS, Postfach 10 16 21, D-68016 Mannheim or log in at http://www.angelfire.com/biz/telri/.

♦ THE TELRI ASSOCIATION

The TELRI Association is registered in Mannheim, Germany, as a non-profit organization under German law. It is the legal framework in which the TELRI partners coordinate and carry out their activities. The TELRI Association is the independent pan-European voice of the multilingual research and development community, a devoted though impartial partner of the European language industry, and a respected consultant of the European Commission for the Multilingual Information Society.

7 CORPUS LINGUISTICS: METHODOLOGY OR THEORETICAL APPROACH

As we have said above, there are two kinds of language resources: corpora and lexica (or electronic versions of dictionaries). The latter are language resources in the full sense only if they are derived from corpora, using methods developed in corpus linguistics. Therefore, we

call corpora primary language resources and lexica secondary resources. We view corpora as the raw material containing all the language data empirical linguistics deals with. There are other kinds of linguistics that do not rely on empirical data in the form of corpora. This is the case for cognitive linguists, and cognitive linguistics still provides the theoretical background to most of the artificial intelligence research in our days.

Some advocates of cognitive linguistics oppose corpus linguistics on theoretical grounds, and this is true the other way around, as well. Even where both approaches are seen as complementary, there is no agreement on where to draw the dividing line. For the past decades, machine translation was primarily based on cognitive linguistics, and this is, to a lesser degree, also true of translation aids. The unsatisfactory performance of most systems, whenever general language is concerned, has to do with semantics, with the meaning of the text to be translated. Translation memories were the first applications that employed corpus linguistic notions. Their immediate success contributed to the paradigmatic change taking place these days. Corpus linguistics now is in a position to demonstrate whether it can compound cognitive linguistics in the treatment of semantics.

Corpus linguistics is the sub-discipline of linguistics that deals with extracting language data from corpora and processing them for various applications such as grammars for human users and for computers, dictionaries, and lexicons. Processing language data turns them either into linguistic knowledge for human users or into the kind of "knowledge" required by computers to perform certain operations involving algorithmic rules, extensionally defined lists, and statistical information on the occurrence of relevant phenomena, usually in the form of type-token rations or derived from them. Corpus linguistics began in the late sixties with the advent of the first corpora. At that time, most linguists used corpora to check, validate, and enlarge the kind of language data that always had been collected by linguists, grammarians, and lexicographers. Even today, the bulk of work done in corpus linguistics uses corpora as a larger and more balanced version of what used to be the individually collected citations on filing cards. Only gradually linguists begin to realize that corpus linguistics will change our perception of language in general, and the relationship between usage and meaning in particular. It was when corpus linguistics moved on to bilingual and multilingual research that this new view on lexical semantics began to take concrete shape. Corpus-based multilingual research needs comparable or parallel corpora to work with. These have not been available for a very long time.

In the beginning, corpora were mostly analyzed in view of grammatical phenomena and usage aspects of very frequent words, like function words. The size of these early corpora (like the Brown Corpus with one million words) was too small to account properly for even basic vocabularies of 5000 words. John Sinclair, the nestor of corpus linguistics, therefore, started in the seventies to build a corpus large enough to serve as a basis for a full-size monolingual dictionary of English. This was the famous COBUILD project, and his corpus grew quickly to a size of 32 million words which turned out to be just about sufficient to find citations for all the word senses that had to be recorded. Today, the original COBUILD corpus has evolved into the Bank of English with approximately 350 million words.

THE COBUILD dictionaries were corpus-based, and they introduced new definition types to describe the meanings. Otherwise, they did not differ much from traditional dictionaries. The basic lexical unit presented was still the single, isolated word, just as it had been the lexicographical custom from classic times on. But working with a corpus had shown that our traditional view of words denoting extralinguistic concepts of phenomena was appropriate only for a certain part of the vocabulary. It works for nouns denoting what is called *natural kinds* like apples and pears, and it works for the standardized terminologies in the various departments of science and engineering. To some extent, it also works for nouns naming cultural artifacts like bathtubs. But the classic view, that is, that words have a limited number of distinct senses, does not work for the words that we use when we talk about natural kinds or artifacts or social constructs or anything else that may bother us. Indeed, it does not work for most of the words that make up our general language vocabularies.

Again, it was John Sinclair who gave corpus linguistics a new push towards multilingual lexical semantics, in his seminal project *Multilingual Dictionary Experiment* (1990-1994), with partners from Britain, Germany, Italy, the Netherlands, Sweden, and additional countries as time went by. (cf. Sinclair 96) The objective was to establish translation equivalents and the conditions for their use, for a small list of rather frequent words, among them *time, man, diary, little/small, lend/borrow*. KWIC-indices of about 100 to 200 random citations each for the source and the target languages were extracted from comparable corpora and analyzed with the aim to establish the exact conditions for choosing the correct translation equivalent for these words which are highly fuzzy and polysemous in all the languages involved. Not unexpectedly, it turned out that the sense disambiguation provided by monolingual dictionaries was of no use. While bilingual dictionaries listed a fair amount of the different equivalents, they failed in detailing the conditions for choosing between them in such a way that they could be processed automatically, that is, without human intervention. Of course, classical bilingual dictionaries are compiled for human users able to

understand what a word means within a context. If they know the target language well enough, they have no problem in selecting the appropriate equivalent. Our goal was, however, to describe the selection conditions in terms that could be processed by a computer. To some extent, we were indeed able to identify the elements (or classes of elements) in the context that are decisive for a particular choice. Our descriptions of these conditions were very different from the definitions in monolingual or bilingual dictionaries. It seemed that the concept of meaning as it is used there is not at all useful for choosing the proper equivalent. Whether a thing called *diary* in English is a *Tagebuch* or a *Kalender* in German has relatively little to do with its being either "a daily record of events" or "a book for keeping" such a record. If you keep it under your pillow, it is a *Tagebuch*, and if it is on your desk, it is a *Kalender*. Thus, it is the context that determines the choice, rather than their meaning. (For a more detailed analysis of *borrow*/*lend*, see Teubert 96) Or, as J.R. Firth put it in his famous quotation: *Words shall be known by the company they keep.* Context analysis is the standard approach of word sense disambiguator. They are one tangible result of corpus linguistics.

We were only half successful in ascribing formal, that is, rule-based or list-based conditions for the proper choice of the translation equivalent. Even together with frequency-based criteria, we were only able to formulate the conditions without recurrence to human understanding in perhaps 50% to 80% of cases, depending on the word in question. Thus, the main result of our undertaking on the theoretical level was that it is primarily usage that determines which word is used in which context. This does not deny that words have their meanings in the sense of classical lexicography. These meanings account for our ability to understand sentences or texts, and to choose the proper translation equivalent. The computer's choice, however, will be based on usage, as expressed in the context in which the word occurs.

The recognition of the role of context was one of the contributions of corpus linguistics. It has led to a reconsideration of the basic unit of lexical semantics. In the European tradition, one focus had always been on the single word, that is, the character strings between blanks, and this concept of the nature of words has become the basis of most dictionary making. Dictionaries arranged isolated single words in alphabetical order, the *lemmata*, and each lemma was described in a lexical entry. But the concept of the single word does not correspond to the concept of a semantic unit. Semantic units include the relevant context features mentioned above. Some of these features are optional and some are contingent. But some elements of the context may be essential, that is, obligatory. This is the case with compounds, multiword units, collocations, and set phrases.

In the multilingual context, semantic units are equivalent with translation units. The comparison of any typical general language text (i.e., a newspaper article) with its translation will demonstrate that at least half of the translation units are larger than the single word. A translation unit is the entity which will be translated as a whole, rather than element by element. In the case of compounds and multiword units, all elements are essential; in the case of collocations, there may be additional, optional elements, and in the case of set phrases, some elements are essential while others may be variables or optional elements. These translation units and their equivalents in target languages can be extracted from parallel corpora. Again, the close analysis of a typical general language text with its translation will show that about one half of the translation equivalents found there are not recorded even in the larger bilingual dictionaries. It is true, however, that a fair amount of these equivalents are idiosyncrasies and poor examples which should not be repeated; but the larger part of them should be used again whenever we find the same semantic unit in a text to be translated. (The method is presented in Teubert 97.1)

Thus, multilingual corpus linguistics will generate an abundance of lexical translational data, for single words, and, even more important, for larger semantic units. It will not stop there. It will also provide the context data containing the conditions for selecting a particular translation equivalent from a set of many. They are contained in the parallel corpora, and the more there are the better will be the results. Good translation matches will be repeated, while poor choices will remain singular instances. More than 95% of the semantic units in our typical general language text have been translated before, because they also occurred on other texts and our parallel corpus will provide the repository of the translation equivalents for these units. The 5% rest usually will be neologisms, and to translate them, a human translator is indispensable.

As I said above, in the *Multilingual Dictionary Experiment* we had been only half successful in determining the formal criteria that govern the choice of the proper translation equivalent. The reason is now easy to see. We were working with comparable corpora, and we selected the matching translation equivalents from bilingual dictionaries where they were arranged according to traditional semantic categories, useful for human users but not appropriate for the kind of algorithmic processing we took for granted to be the only way in which to do multilingual natural language processing.

But translation is an action presupposing intentionality and, except for controlled documentation languages, it cannot be reduced to a merely syntactic manipulation of symbols. The use of parallel corpora makes the impossible possible. Parallel corpora are texts and their

translations, human translations, we should add, or at least translations supervised and post-edited by translators who understood both the source and the target text. These translations, broken down into translation units and their equivalents, together with relevant context, are the source data for information translation aids. Context information consists of the features that the contexts share in respect to a given equivalent and of those features that differ and thus identify the equivalents of a given translation unit. The traditional semantic categorization found in bilingual (or monolingual) dictionaries has, thus, become irrelevant. Parallel corpora are also a repository of larger translation units, like compounds, collocations, and set phrases. Unlike single word units, these units are only seldom ambiguous. Even large bilingual dictionaries list only a fraction of them due to space considerations and also due to the fact that, before corpus linguistics, there have been no attempts to collect them systematically and all-inclusively.

Multilingual corpus linguistics uses parallel corpora for translation. It develops the methodology to recover the linguistic and translational expertise of human translators and to re-use their products, broken down into translation units and their equivalents and processed for disambiguation. It can provide answers where the traditional, cognitive approach in machine translation fails. This is why multilingual language resources are so important.

8 MULTILINGUAL LEXICAL SEMANTICS

In cognitive linguistics, and hence in classical artificial intelligence, including machine translation, the words of natural languages correspond somehow to concepts. The "somehow" is where authors differ. So we can read: (cf. Teubert 97.2 and Fodor 98)

- concepts are expressed in words in a natural language
- concepts represent the abstract meaning of words
- concepts represent word meanings
- word meanings are just concepts

Concepts are thought to be pure whereas words are contaminated by the idiosyncrasies and contingencies we find in natural languages. In Cartesian dualism, they represent the mind, while the language-dependent expressions represent the body. Concepts can be mental fabric in three senses:

- Concepts can be the result of an explicit agreement between experts. They are, in principle, language independent or, at least, language neutral. This is the view we find in terminology. An alphabetic character is what the relevant ISO committee has defined.
- Concepts, at least primitive concepts, are innate, features of the mind/brain, so to speak, while they may be occasioned by the pertaining experiences. Complex concepts, in this view, are composed of primitive concepts. This view was endorsed by the early Jerry Fodor, when he was still a Radical Nativist (in *The Language of Thought*) and is still the standard paradigm for Stephen Pinker (e.g., in *The Language Instinct*). For him, concepts are the words of a universal mentalese. (Fodor 75, Pinker 94)
- Concepts exist independently of people having them. Concepts are the true meanings of things (regardless what we believe). They are the pure essence, the *idea* underlying the material. That is the Platonic view. Augustine also subscribed to it, and today, it is the view of the metaphysical realists dominating the American philosophy of language, notable Hilary Putnam. (Putnam 81)

The second and the third view of concepts is the standard picture held by the artificial intelligence community. For them, concepts are mental features and, at the same time, features of the world. In his most recent book, *Concepts,* Fodor tells us how he sees the link between the mind and the world. For him, having the concept 'doorknob' is just *having the property that minds like ours reliably lock to in consequence of experience with typical doorknobs*. Thus, the concept "DOORKNOB expresses a property that things have in virtue of their effects on us." (Fodor 98, p. 148) Now the phrase *in virtue of* is intentionally vague, and *lock to* seems to be a metaphor taken from our view of the immune system. Fodor's link is perhaps evocative but certainly not explanative for the dual aspect of concepts.

Corpus linguistics knows nothing about concepts. For corpus linguistics, words are symbols with two aspects, the aspect of expression and the aspect of meaning. The relationship between them is arbitrary. What a word means is a convention reached by the language community. This is the standard view of continental European structuralism, as it was developed by Ferdinand de Saussure. Word meaning is not immutable, and it is not necessarily fixed. In discourse, it can be discussed, questioned, modified, changed, and replaced. This discourse goes on continually. Therefore, meanings must be fuzzy and ambiguous. Since meanings are nothing but one of the two aspects words, as symbols, have, there cannot be something such as language-independent meaning. Therefore, it is not possible to arrive at the proper translation equivalent by the syntactic manipulation of symbols. Rather, only if we, as human beings, understand the sentence in question, we can paraphrase it in, that is, translate it into, a sentence in a different language, by choosing the translation equivalents which, in our reading (or interpretation) of the sentence, correspond best to the translation units. Parallel corpora capture the discourse of translation, the perennial endeavor of translators to find the proper translation equivalent in a world where meanings are fuzzy and

ambiguous. Parallel corpora contain the translational knowledge needed by general language translation aids.

9 REFERENCES

Melby, Alan K.(1995): T*he Possibility of Language*. Amsterdam (Benjamins)

Dennett Daniel C. (1991): *Consciousness Explained*. New York (Little, Brown)

Searle, John R. (1992): *The Rediscovery of Mind*. Cambridge, Mass. (MIT)

Calzolari, Nicoletta; Baker, Mona; Kruyt, Johanna G. (eds.) (1996): *Towards a Network of European Reference Corpora*. (=Linguistica Computazionale XI). Pisa

Collins *Cobuild English Language Dictionary*, ed. by John McH Sinclair (1987). London (HarperCollins)

TELRI Final Report, ed. by Wolfgang Teubert and Kirsten Plöger (1998). Birmingham/Mannheim (TWC)

Sinclair, John McH (1996): "An International Project in Multilingual Lexicography." *International Journal of Lexicography* Vol. 9, Nr. 3, pp. 179-196

Teubert Wolfgang (1996): "Comparable or Parallel Corpora?" *International Journal of Lexicography* Vol. 9, Nr. 3, pp. 218-237

Teubert, Wolfgang (1997.1): "Translation and the Corpus." *Language Applications for a Multilingual Europe*. (TELRI Proceedings of the Second European Seminar). Mannheim/Kaunas (IDS/VDU), pp. 147-164

Teubert, Wolfgang (1997.2): "Translation Ontologies for Multilingual Applications." Proceedings of the Workshop Corpus Use and Learning to Translate (Bertinoro/Forli): http://www.sslmit.unibo.it/cult.htm

Fodor, Jerry A. (1998): *Concepts: Where cognitive Science Went Wrong*. Oxford (Clarendon)

Putnam, Hilary (1981): *Reason, Truth and History*. Cambridge (CUP)

Fodor, Jerry A. (1975): *The Language of Thought*. New York (Crowell)

Pinker, Steven (1994): *The Language Instinct*. New York (William Morrow)