Mattias Agnesund (Göteborg University, Sweden)

*From Text to Theory - On Parallel Corpora as a Key to Understanding Translation*

The gap between translation as a philosophical problem and translation as a practical activity is an obstacle for a coherent treatment of the whole area. Large amounts of texts are being translated daily, with linguists and philosophers always lagging behind in their attempts to explain what is actually happening. Theories of translation are mostly based on related disciplines like logic and pragmatics, which provide the framework in which attempts to formulate the nature of translation are carried out. However, empirical evidence is rarely taken into account. References to actual translations are reduced to brief examples supporting certain positions. One reason for this neglect has been lack of translated material to study. Today, that excuse may no longer be valid. Parallel corpora - i.e. collections of source language texts paired with their respective translations - are already playing an important role in contrastive linguistics. It seems likely that they will also have an impact on our understanding of translation from the philosophical viewpoint.

This paper explores the possibilities of using large parallel corpora as the basis of a theory of translation. It discusses three computational methods for extracting the necessary information from the corpora - statistical analysis, parallel concordance software and meaning representation - and shows how this information might be related to philosophical notions like equivalence for example. The paper takes as its starting point an English-Swedish parallel corpus of fiction, which provides the empirical evidence in an ongoing study based on the above principles.

Victoria Alsina & Janet DeCesaris (Universitat Pompeu Fabra, Spain)

*Bilingual lexicography, overlapping polysemy and corpus use*

This paper discusses the structure of adjective entries in Spanish-English/English-Spanish and Catalan-English/English-Catalan bilingual dictionaries in relation to two topics: use of corpus-based data and overlapping polysemy. A long-standing problem in producing entries for highly frequent adjectives in bilingual dictionaries is the identification of the possible contexts in which the equivalent provided by the dictionary may be used. Many high frequency adjectives are used in both a literal sense as well as in many metaphorical contexts, but the best equivalent for this single adjective in another language does not necessarily occur in both sorts of context, or, alternatively, may be used in the literal sense and some of the same metaphorical contexts but not all of them. We suggest that this phenomenon, which we call overlapping polysemy, should play a significant role in the structuring of the equivalents provided by the dictionary because it makes use of a speaker’s knowledge of the source language and results in shorter dictionary entries, thus freeing up space for other uses. Overlapping polysemy is based on the lexicographer’s
linguistic analysis of the use of the adjective, and can benefit from consultation of a corpus in the identification of which metaphorical extensions are commonplace, although the actual analysis is the lexicographer’s, not the corpus’s. One way in which a monolingual corpus has been used in lexicography is to determine the order of senses listed, and this use of corpus data could be applied in bilingual lexicography: one could list the equivalents by order of frequency of meaning as identified in the corpus for the source word. Frequency data, however, do not necessarily correspond to the literal-metaphorical extension parameter that we have identified as important to the structuring of adjective entries. In this paper we look at four adjectives in English which display varying degrees of overlapping polysemy with their equivalents in Spanish and Catalan: high, cold, important, and odd. We analyze existing entries in bilingual dictionaries, and show how the articles might be reorganized if they were based on an English-language corpus. We then compare the reorganized entries with those based on our analysis of overlapping polysemy. Our goal is to identify the proper role for use of a monolingual corpus in the preparation of a general language bilingual dictionary.

Bengt Altenberg (University of Lund, Sweden)

*Causative constructions in English and Swedish. A corpus-based study*

In a recent study of the lexical and grammatical patterning of the verb make in the International Corpus of Learner English, Granger and Altenberg found that French-speaking and Swedish EFL learners deviated in interesting ways from native American students’ use of the verb. While both learner groups underused (and misused) ‘delexical’ make (e.g. *make a decision, make a point*), they were clearly differentiated in their treatment of causative make (e.g. *make sb happy, make sb believe sth*): the French-speaking learners significantly underused causative make with adjective and noun complements (e.g. *make sth possible, make sb a star*), whereas the Swedish learners revealed a significant overuse of causative make with adjective and verb complements (e.g. *make sth easier, make sb understand*). Another interesting finding was that the learners’ treatment of causative make seldom resulted in clear errors but in a number of rather clumsy constructions, suggesting that the learners tended to opt for a semantically and grammatically ‘analytical’ (or ‘decomposed’) *make* + object + complement pattern in cases where a native writer would prefer a ‘synthetic’ causative verb alternative (e.g. *make the air polluted* instead of *pollute the air*).

To get a better understanding of the Swedish learners’ overuse of causative *make* a contrastive study of causative constructions in English and Swedish was carried out on the basis of the English-Swedish Parallel Corpus. The starting-point was the main analytical constructions in the two languages: *make* + object + adjective/infinitive complement in English, and *göra* + object + adjective and *få* + object + infinitive in Swedish. For each construction, cross-linguistic correspondences were explored in two directions: from source texts to translations, and from translations to source texts.
The investigation shows that, in principle, English and Swedish have a similar range of resources for expressing causative relations. The main alternatives in both source texts and translations in both languages were:

- analytical constructions
- synthetic verbs
- other causative verbs
- miscellaneous other constructions

In both languages the two analytical constructions with make in English and göra/få in Swedish were the most common alternatives. However, in the corpus as a whole they represented less than 50% of the causative examples and their mutual translatability was moderate (51-55%). The competition from other alternatives is consequently strong in both languages. Yet, the frequency of the analytical constructions was consistently higher in the translations than in the source texts, which indicates that they were perceived as the natural or ‘unmarked’ translation equivalents by the translators of both languages.

The relative importance of the other alternatives varied somewhat in the two languages. Generally, the ‘miscellaneous’ category was the most common alternative in both languages, especially in the source texts. Synthetic verbs were more common as an alternative to analytic adjective constructions than to infinitive constructions, and they were more common in the English than in the Swedish texts. Other causative verbs were a more common alternative with infinitive complements than with adjective complements. The choice of construction was found to be determined by a range of language-specific and language-independent factors – lexical, grammatical and contextual.

Ludo Beheydt (Université catholique de Louvain, Belgium)

The A-index: a new measure of lexical diversity for corpus linguistics

Corpus linguistics is in need of reliable and simple measures of lexical diversity. Especially for contrastive purposes, it would be of great use to be able to apply the same measure of vocabulary density to samples of different length and of varying languages. A comparison of the lexical diversification of Learners’ Language in the foreign language and the lexical diversification in native language samples of the same subjects might yield a reliable measure of lexical diversity. Such a comparison might even eventually provide us with a yardstick for foreign language lexical development.

Two major problems have up till now impeded the development of such a reliable vocabulary measure.

First of all, the linguistic differences between languages tend to complicate comparisons. Although it seems possible to discern the word as an entity in each and every language,
the linguistic status of the word may greatly differ in typologically different languages. There is clearly no one-to-one lexical correspondence between languages. Analytic languages like French, Spanish or Italian tend to analyze semantic lexical entities into formally discrete lexical elements (e.g. French *machine à laver* = 3 word forms; *village natal* = 2 word forms; *championnat du monde* = 3 word forms). Synthetic languages like Dutch or German tend to keep semantic lexical entities in compound formal units (e.g. Dutch *wasmachine* = 1 word form; *geboortedorp* = 1 word form; *wereldkampioenschap* = 1 word form).

Lexical measures based on word counts (e.g. TTR) not taking account of such typological differences will distort comparisons inacceptably.

A second problem that has hitherto hampered comparisons of lexical richness, even on an intralingual basis, is the fact that due to the complex mathematical relation between the number of tokens (total number of words) and the number of types (number of different words) in texts, it has as yet not been possible to develop reliable measures of lexical complexity for texts of unequal length. A very popular measure like the Type Token Ratio, for instance, has the obvious disadvantage of getting smaller as the size of the sample gets bigger.

My presentation will offer a newly developed measure of lexical variability that can be used for various languages and that seems to be less dependent on text length. This new measure - the A-index - will be presented together with an application to varying language samples in different languages. A discussion of its characteristics and potential should contribute to a fuller appreciation of its use in corpus research.

Xavier Blanco (University Autonoma de Barcelona, Spain) & Pierre-André Buvet (Université de Franche-Comté, France)

Détermination de l'espagnol et du français: études de cas

Nous présenterons des études effectuées dans le cadre du projet DétTAL (Détermination et Traitement Automatique des Langues) dont l'objectif est de décrire exhaustivement la détermination dans différentes langues (notamment l'espagnol et le français) dans la perspective du TAL. Notre travail ne se limite pas au seul recensement morphologique des déterminants, mais aborde également leur dimension syntactico-sémantique.

Des travaux récents sur la détermination justifient les grandes orientations de notre projet. En premier lieu, nous considérons la phrase comme l'unité de base de l'analyse linguistique car les compatibilités entre les déterminants et les noms peuvent dépendre d'autres éléments de la phrase. Ensuite, les déterminants sont étudiés séparément selon qu'ils se rapportent à des noms correspondant à des prédicats ou bien à des arguments élémentaires étant donné que leur fonctionnement dépend des propriétés syntaxiques des substantifs qui les acceptent (les déterminants s'appliquant à des noms qui apparaissent

A partir d'études de cas (les noms d'opérations financières, de crimes et délits et de phénomènes naturels), nous montrerons comment il est possible de formaliser les diverses contraintes sur les déterminants en décrivant la syntaxe des substantifs d'une classe donnée à l'aide d'automates à états finis. Nous indiquerons également en quoi les corpus numérisés constituent un outil performant pour la réalisation de cet objectif. Nous discuterons ensuite de l'élaboration d'un module de correspondances entre les descriptions des déterminants du français et celles des déterminants de l'espagnol afin de systématiser les observations contrastives pour la mise en place d'un système de traduction automatique ou semi-automatique.

Diane Boonen (University of Gent, Belgium), Mieke van Herreweghe (University of Gent, Belgium) & Myriam Vermeerbergen (Vrije Universiteit Brussel, Belgium)

Flemish-Belgian and Dutch Sign Language: syntactic differences due to a different modality

Deaf people in Flanders do not have Dutch as their first language, as their hearing fellow-citizens do, but Flemish-Belgian Sign Language. Because this is a language in a different modality (a gestural-visual modality as opposed to an oral-auditory modality) it has a number of very specific characteristics diverging greatly from those of spoken languages.

For a long time, sign languages were, at best, considered more primitive, more limited than spoken languages. It seemed hard to understand that human capacity for language did not depend on human capacity for speech and to accept that languages can also develop outside the oral-auditory modality. Today it is generally accepted -at least within the linguistic community- that sign languages are capable of the whole range and diversity of expression possible in spoken languages, notwithstanding their use of a different means of expression.

Since the publication of the first modern linguistic analysis of a sign language, William Stokoe's 'Sign Language Structure' (1960), a myriad of books and research papers on sign language usage and structure have seen publication. Part of the research focuses on the similarities between spoken and sign languages. Especially in the early days of sign language studies, linguists often felt the need to prove that sign languages are indeed real, natural human languages by pointing out the parallels with spoken languages. Yet many
studies concentrate on the differences between spoken and sign languages, for the difference in means of expression does of course result in structural differences.

The fact that sign languages are gestural-visual languages opens up possibilities that do not exist -or which exist to a lesser degree- exist for spoken languages. The possibility of using two hands, the head, face and torso at the same time for example, results in a high degree of simultaneity not known to spoken languages that are generally characterized by sequentiality. Another aspect of sign languages that makes them very different from spoken languages is the use of the so-called ”signing space” (the space in front of and on both sides of the signer's head and torso) in morphosyntax and discourse structure. Particularly interesting is the way in which referents are ”localized” in the signing space and the way they can be referred back to by the signer.

In this paper we will discuss some of the structural differences between Flemish-Belgian Sign Language and Dutch. In order to examine these differences we conducted a small-scale corpus study. Deaf signers (teenagers and adults) were asked to look at an animated cartoon in which no language was used (no sign language and no spoken language, not even via subtitles). They were afterwards asked to recount the whole story in Flemish-Belgian Sign Language. The same cartoon was also shown to other deaf people, but this time they were asked to narrate the whole story in written Dutch. The in-depth comparison of these two versions of the same narrative renders a number of interesting syntactic differences.

References:

Lynne Bowker (Dublin City University, Ireland)

A Corpus-Based Contrastive Analysis of the Effect of Multidimensional Classification on English and French Terms

Terms are the linguistic representations of concepts, and, according to Sager (1997:25), term formation is the process of naming the concepts required by a particular special language community for the development of cognitive processes and communication. Sager goes on to add that term formation is a conscious activity which differs from the arbitrariness of general word formation because term formation requires a greater awareness of pre-existing patterns and models, and of the social responsibility for facilitating communication and transmitting knowledge. For these reasons, it is often thought that terms are not prone to variation; however, with the increasing availability of electronic corpora comprised of texts written in Language for Special Purposes (LSP), empirical studies (e.g. Daille et al. 1996, Rogers 1997) are beginning to reveal that
terminological variation is actually a significant phenomenon. Nevertheless, it is often assumed in such cases that the author’s use of variants rather than of a single preferred term is arbitrary and by implication sloppy. But this is an assumption which requires further investigation. Systematic patterns of variation need to be explained, and finding out the causes, the patterns, or the regularities hidden behind such apparent randomness is one of the new tasks facing terminologists.

One of the potential causes for this apparent randomness may be multidimensional classification, a phenomenon that occurs when a concept can be viewed from multiple perspectives and can thus be classified (or referred to) in more than one way. In this paper, we explore the impact that multidimensional classification has on the formation of terms, and particularly on the formation of categoric compounds. I began this study by focusing on the phenomenon as applied to English terms; however, it soon became apparent that multidimensional classification and the resulting terminological variation have serious implications for translation. Therefore, I expanded the study to include a contrastive analysis of this phenomenon in English and French. To this end, I have constructed English and French corpora containing comparable texts in the subject field of optical scanning technology. I have used the software package WordSmith Tools to help me analyze the data.

This paper will be divided into four main sections. In part 1, I consider some of the rules and hypotheses surrounding term formation as found in the terminology literature. In part 2, I introduce and explain the concept of multidimensional classification. In part 3, I examine English terms in context taken from a corpus in the subject field of optical scanning technology and I investigate how multidimensionality has influenced the formation of these terms. Finally, in part 4, I conduct a contrastive analysis with corresponding French terms, compare the effects of multidimensional classification in the two languages, and discuss the implications for translation.

Sylviane Cardey-Greenfield (Université de Franche-Comté, France)

Analyse et conception de lexiques informatisés

Nous présenterons dans un premier temps quelques uns des travaux réalisés dans notre Centre de recherche Lucien Tesnière à Besançon dans le domaine de lexiques bilingues, multilingues tels que :
- un lexique informatisé anglais/français de collocations utilisées en recherche clinique,
- un système informatisé pour la reconnaissance et la traduction assistée ou automatique d'expressions figées anglais/français et anglais/français/italien,
ainsi qu'un système, logiciel, pour l'étiquetage automatique du lexique français en catégories grammaticales ou syntaxiques et pour la levée des ambiguïtés dues à la morphologie (du type "ferme", qui peut être adjectif, nom, verbe, adverbe) : mots simples et composés à deux éléments.
Nous verrons ensuite comment ces travaux ont été menés : entre autres et surtout la part d'analyse purement manuelle et celle qui a pu être réalisée de façon automatique à l'aide de logiciels spécialement conçus par l'équipe pour manipuler le lexique en général.

Sera montrer également pourquoi une partie de la recherche ne peut se faire que de façon manuelle :
- analyse linguistique pure et stratégies de repérage entre autres,
et en quoi l'ordinateur est un outil indispensable :
- tri dans les dictionnaires généraux et extraction de lexiques spécialisés ou spécifiques dans la littérature en général,
- extraction de racines, suffixes, flexions ....,
- validation de la recherche pour des applications telles que la traduction assistée ou automatique nécessitant des résultats sûrs pour des domaines où la sécurité est primordiale (domaines médical, nucléaire, aéronautique...).

Nous pourrons conclure en disant que nos recherches, applications et produits nous ont convaincus que le traitement du lexique nécessitait une part centrale d'analyse linguistique manuelle (conception, compréhension et organisation) mais que sans l'ordinateur il serait difficile de progresser et de contrôler les résultats.

Jonathan Charteris-Black (University of Surrey, Great Britain)

A comparative approach to the translation of English and Malay idioms and phraseological units

The aim of this paper will be to establish a principled basis for the translation of English and Malay phraseological units and idioms. Initially a prototype model for the definition of phraseological units and idioms will be proposed; it is claimed that the non-compositionality of phraseological units originates in a range of semantic, pragmatic, syntactic and lexical features, which may be present to a greater or lesser degree. Pure idioms are the prototypical category because they exhibit these features to the greatest extent; however, the broader notion of phraseological unit also includes other multi-word units, which exhibit these features to a lesser degree.

To facilitate the translator’s task a model is proposed for the cross-linguistic comparison of phraseological units based on an analogy between various types of human relationship and the differing relationships of form to meaning and/or conceptual basis in phraseological units:

1/ ‘Friends’ - the same/similar meaning and the same surface lexis.
2/ ‘Neighbours’- the same/similar meaning and similar surface lexis.
3/ ‘Cousins’ the same/similar meaning but different surface forms
4/ ‘Acquaintances’ - transparent surface forms, with a universal conceptual basis.
5/ ‘False friends’ - different meanings but the same surface forms.
6/ ‘Strangers’ - opaque forms, with a culture specific conceptual basis.

Each category is illustrated with reference to English and Malay, but the model is proposed to be a general one with potential applications to other languages.

There is a consideration of the translation problems and solutions for members of each category. Categories 1 and 4 should be the least problematic for translators because the meanings of phraseological units in one language can be accessed through their surface forms in another. Category 5 is problematic because of the danger of interference. Category 6 is also problematic because of a lack of knowledge of the conceptual basis linking surface forms with their phraseological meanings.

Categories 2 and 3 are also problematic because of the divergence of forms and meanings in the two languages. It is proposed that corpora can help resolve the difficulties where there are different surface forms for the same or similar meanings - and therefore the potential for the use of a phraseological unit in either language. Where the translator is aiming for clarity of meaning, he or she should prefer the surface form (or a literal translation of the surface form) which is found to be most frequent in a representative corpus of each language irrespective of whether this is the surface form of the source or target language. In this way, the relative importance attached to a phraseological unit in a particular language - as evidenced by a comparison of corpora - can influence the translation outcome.

Christine Chodkiewicz, Didier Bourigault & John Humbley (Université Paris 13, France)

Using LEXTER to establish a glossary of Human Rights

This paper aims at presenting the use of a term extractor, LEXTER, in establishing a multilingual glossary of a specialised field by exploiting a limited bilingual corpus of fundamental texts. Stress is put on dealing with multiple translations and extension of a bilingual to a multilingual tool, but mention is also made of extending the scope of the term extractor itself.

The corpus chosen in the pilot project was the French and English versions of the Declaration of Human Rights and the decisions of the European Court of Human Rights, comprising several hundred pages of text. Since LEXTER was only programmed for French texts at the time the project began, terms were first extracted from the French version, and equivalents matched in the parallel English text. After the terms extracted in French and the English equivalents had been validated by a lawyer, some 6000 pairs were found to have more than one equivalent. As synonymy is generally held to be strictly
controlled in special languages, a study was then undertaken to characterise the variation and to determine which alternative translations are licit, which may lead to error and which are themselves errors of the system.

In this paper, a classification is proposed, shedding light on both linguistic and subject-area reasons precluding one-to-one equivalents, and suggestions are made to avoid inclusions of misleading information in the glossary produced. In addition, some information is given on how this experiment was used to increase the capabilities of the software used and how it is extended to other languages.

Hélène Chuquet (Université de Poitiers, France)

*L'imparfait français est-il traduisible en anglais?*

L’imparfait français est une forme qui a déjà été abondamment étudiée, en raison notamment de son apparente polysémie et des multiples effets de sens qu’elle est susceptible de produire en contexte. Que l’imparfait soit décrit prioritairement comme temps du passé ou comme aspect inaccompli, qu’il soit considéré comme appartenant à la fois au plan de l’Histoire et à celui du Discours (E. Benveniste) ou bien défini comme temps narratif de l’arrière-plan dans le récit (H. Weinrich), il se trouvera toujours des exemples pour sembler contredire la description qui en est faite.

L’imparfait ne semble pas à première vue poser des problèmes insurmontables au traducteur vers l’anglais: il est en général relativement aisé de délimiter les contextes d’emploi dans lesquels l’imparfait se traduira par l’une des formes aspecto-temporelles de l’anglais qui peuvent, suivant des degrés de contrainte variables, lui être considérées comme équivalentes — prétérit simple ou progressif, past perfect ou formes modales.

La comparaison de l’emploi de l’imparfait français à ses traductions en anglais soulève néanmoins des questions intéressantes, qui sont (au moins) de deux ordres:
- D’une part la question de la définition même de l’opération linguistique fondamentale que recouvrent respectivement l’imparfait français et le prétérit simple anglais. En effet, l’analyse de ces deux formes, telle qu’elle est effectuée dans la théorie des opérations énonciatives à la suite d’Antoine Culioli, débouche sur un paradoxe pour le linguiste contrastiviste: comment se fait-il qu’une forme décrite comme étant la trace d’un repérage par translation de l’origine énonciative et associée à l’expression d’un aspect inaccompli (l’imparfait) se trouve si souvent traduite par le prétérit simple qui, lui, serait la marque de la rupture par rapport à la situation d’énonciation origine et d’un aspect globalisé (perfectif) ?
- La seconde question est celle de savoir si tous les effets textuels produits par l’imparfait peuvent être restitués en anglais. En effet, selon le cas, on observe dans le passage à l’anglais soit une perte, souvent assortie d’une sorte de “neutralisation” stylistique, soit au
contraire une sur-explicitation de “l’effet-imparfait”, que ce soit sur le plan aspectuel, modal ou lexical.

Tout en essayant d’apporter quelques éléments de réponse à la première question posée, je me propose surtout de faire l’étude de quelques cas d’emploi de l’imparfait de l’indicatif français faisant référence au passé à travers un corpus essentiellement (mais non exclusivement) constitué de textes journalistiques français accompagnés de leur traduction en anglais, en vue de dégager certaines régularités dans les phénomènes de non-équivalence entre les potentialités des formes des deux langues. On s’intéressera en particulier aux cas où le recours à l’imparfait permet d’entretenir un certain degré d’ambiguïté sur le mode d’appréhension du procès, et sur les conséquences que peut avoir sur l’interprétation du texte traduit la nécessité linguistique de lever cette ambiguïté.

Références bibliographiques.


Laura Cignonini (Instituto di Linguistica Computazionale, CNR Pisa, Italy),
Stephen Coffey (University of Pisa, Italy) & Rosamund Moon (Cobuild & University of Birmingham, Great Britain)

Idiom variation in Italian and English: two corpus-based studies

This paper will report on two parallel but independent studies of idioms in corpora - one of Italian and one of English - and show how they found strong evidence of similar kinds and degrees of idiom variability in the two languages. In the Italian study, 324 idiomatic expressions were looked for in a corpus of 15 million words; the English study investigated more than 2,800 idioms in an 18-million word corpus.
The introductory part of the paper will set out the bases for the comparability of the two studies, focussing on the kinds of phrasal units under investigation, the size and composition of the two corpora.

The main part of the paper will present our two sets of findings. Here we will systematically list and exemplify the kinds of variation which were found to occur in both languages, whether lexical or syntactic, substantive or minor, regularly occurring or ad hoc. Thereafter we will describe the relatively few variation types which were found in one language only. We also comment on the overall frequency of idiom variation, which was found to be very similar in the two studies.

In our concluding remarks we will underline the fact that these corpus studies suggest a high degree of similarity between the two language systems as regards idiom variation. We will suggest that this relatively unchartered territory in contrastive linguistics could have ramifications in practical applications such as bilingual lexicography, second language pedagogy, translation, and language engineering. We will also stress the importance of carrying out similar studies between other language pairs.

Patrick Corness (Coventry University, Great Britain)

**Practical techniques of parallel corpus alignment for Multiconcord, using Minimark and Word**

This brief presentation will show how a simple automated markup system can be used in conjunction with Word for Windows to prepare an aligned corpus for analysis with the Multiconcord parallel concordancing program, which can then yield useful data for contrastive analysis (CA).

The technique is only partly automated, and it includes a significant manual component, which is believed to be typical of parallel corpus work. Total automation is impossible and manual checking is essential, but ways have to be found to maximise the former while acknowledging the role of the latter, since it is thanks to the power of automated techniques that contrastive analysis is presently coming into its own. To this end, various complex markup schemes and automatic parsers are being developed, but some of the same limitations that apply to machine translation (MT) as opposed to computer-assisted translation (CAT) are found when a high level of automation is attempted in corpus analysis. Much can be done with more straightforward tools and it is the purpose of this presentation to demonstrate some relevant examples based on experience with small corpora.

The present scope of Multiconcord for CA is assessed and plans for enhancement of the functionality to satisfy the needs of researchers are outlined.
Hortènsia Curell i Gotor (Universitat Autònoma de Barcelona, Spain)

The role of a bilingual corpus in the study of the Present Perfect in English and in Catalan

English and Catalan are two languages with totally different sociolinguistic situations. While English is spoken by around 400 million people as a first language and by 350 million as a second language, together with about 100 million people who use it fluently as a foreign language, Catalan only has around 8 million speakers, none of whom is monolingual (Crystal 1997).

The number of corpora available for English in its different varieties is nowadays very large and can be accessed with relative ease. For Catalan, however, corpus creation is very much in its initial stages. Two different institutions are gathering and computerizing data both from spoken and written language at present, but neither of the two corpora is available as yet. Clearly, then, there exists no bilingual corpus, computerized or not, which can be used by contrastivists.

The corpus used in this study was gathered manually by the researcher. It consists of 665 items (each item formed by an English verb phrase and its Catalan counterpart), extracted from four American police novels and four film scripts (their original version and their Catalan sanctioned translation). The verb phrases are presented within the smallest context necessary to their correct interpretation. The context is important for the interpretation of any utterance, but especially so for the present perfect: the use of this form is often conditioned not by syntactic rules but by pragmatic factors.

The aim of this paper is to show that corpus analysis is absolutely necessary for a successful contrastive study. A detailed analysis of data provides insights not found in basically introspective studies. After examining the bilingual corpus in detail, we have been able to support some aspects of existing descriptions, refine others, and even contradict certain statements about the use of the Present Perfect in English and in Catalan. For example, we have found that the present perfect form is not the only one that can express present perfect meaning in both languages. We have also been able to establish determining factors such as adverbials (type and obligatoriness) and the aspectual character of the situation (telic or atelic).
Forward causation in two languages: the case of dus and donc

Previous research on the distribution of the Dutch (forward) causal connectives *dus*, *daarom*, and *daardoor* has shown that the connective *dus* has the broadest functional expression potential, i.e. it can occur in different types of causal contexts (Pander Maat & Sanders, 1995, 1997; Degand, 1996). While *daardoor* can only express non-volitional relations of a content type and *daarom* seems to be restricted mainly to volitional content-relations and deductive epistemic relations, the connective *dus* may be involved in all types of causal relations and causal reasoning patterns: volitional content-relations (1), deductive (2) and abductive (3) epistemic relations, and purely textual relations like restatement or summarizing (4).

In this paper, we want to contrast these various uses of the connective *dus* with the distribution of its French counterpart *donc*. Actually, it appeared from a newspaper corpus analysis that the connectives seem to occur in very similar contexts. Like its Dutch counterpart *dus*, the connective *donc* can appear in a broad range of functional contexts. So, at first sight, the two connectives should nearly always be mutually substitutable, i.e. they should most of the time be translated by each other. However, it appeared from a small bilingual corpus based on information texts delivered by the European Community that this is far from being always the case. In particular, the connective *donc* is not always translated by the connective *dus* in the Dutch text version. As a matter of fact, *donc* appeared to be much more frequent than *dus*. This would then suggest that the French connective fulfils other functional roles that did not appear from its analysis in a monolingual corpus. As a consequence, bilingual corpus analyses are not only useful for contrastive linguistic description, they also enable us to gain deeper insights into the uses and functions of linguistic constructions within one language. Our next task then seems to be to construct vaster and genre-diversified bilingual corpora so as to foster this kind of research.
Olof Eriksson (Université de Växjö, Suède)

La structure phrastique du français en analyse contrastive

Ma communication présentera les principaux résultats d’un livre publié en suédois en 1997 (1) et qui avait pour but de comparer la structure phrastique du français avec celle d’une langue germanique, en l’occurrence le suédois. Elle a partiellement pour base aussi un livre publié en français en 1993 et qui examinait la structure phrastique du français d’un point de vue non contrastif (2).

Dans ma communication, je rendrai compte de la méthode et des matériaux employés : analyse à la fois quantitative et qualitative basée sur un corpus composé de 40 textes littéraires suédois en langue originale et en traduction française (1.000 propositions successives retenues dans chaque texte), recours à des contrôles pour assurer la validité statistique de l’étude (subdivision du corpus en corpus partiels, corpus supplémentaire pour mesurer le rôle de l’interférence dans le processus de traduction, corpus supplémentaire pour garantir la stabilité statistique de chaque texte individuel).

L’étude montre que la phrase française a une « identité structurale » qui diffère fondamentalement de celle du suédois. Je me sers pour caractériser cette différence de termes comme « densité phrastique » (« satstäthet ») et « profondeur phrastique » (« satsdjup »).

Par rapport au français, le suédois est une langue à orientation verbale (« verborientated language »), c’est-à-dire une langue dans laquelle les actions, les procès et les états tendent à s’exprimer par desunités syntaxiques ordonnées autour d’un verbe en forme finie. Cela veut dire que le suédois favorise les membres phrastiques et syntagmatiques qui se rattachent à la sphère fonctionnelle du verbe (syntagme adverbiale, proposition circonstancielle, etc.) aux dépens de ceux qui se rapportent, par subordination (proposition relative, syntagme adjectival, syntagme prépositionnel) ou par prédication (syntagme participial, syntagme infinitival), à la sphère fonctionnelle du nom. Cette dernière tendance caractérise très nettement le français, qu’on peut sans aucun doute qualifier de langue à orientation nominale (« nounorientated language »).

L’étude montre que la répartition quantitative des divers constituants syntaxiques dans les deux langues diffère considérablement en chiffres absolus, mais que c’est en examinant la distribution syntaxique de ces constituants qu’on arrive à saisir les différences les plus pertinentes entre les deux langues. Il s’avère, par exemple, que le français aussi bien que le suédois font un emploi relativement fréquent de la subordonnée complétive mais qu’ils assignent à ce constituant propositionnel des rôles syntaxiques assez différents.

The starting point of this investigation is the hypothesis that German modal particles, utterance medial pragmatic particles, display a relational semantic structure and that their general grammatical function is to connect the current utterance to a pragmatic pretext, i.e. to a proposition "at hand" which is part of the non-verbal context (cf. Diewald/Fischer 1998). The question addressed in this investigation is the following: If this assumption is correct, what happens to this function in languages which do not have modal particles, such as English?

Previous contrastive approaches have analysed the lexical and grammatical correspondences modal particles may have; for instance, auxiliary verbs (e.g. Nehls 1989), intonation (e.g. Schubiger 1965), tag questions (e.g. Fillmore 1981), or, most often, no equivalents at all. However, none of these approaches used a concept of a common function of the word class or a relational semantic structure as a starting point for the analysis; instead, their search for translation equivalents was guided by local decisions on the current pragmatic function of each modal particle occurrence. With this procedure, often no correspondences could be found at all. O'Sullivan/Roesler (1989) therefore argue that in the translation of English texts into German the use of modal particles should be seen as a holistic enterprise depending on the text type. This is also unsatisfactory if it is not clear which properties of the respective text type make the use of modal particles relevant.

In the current investigation, firstly an extract from a German appointment scheduling dialogue recorded in the Verbmobil project will be discussed with respect to the function of each modal particle; secondly it will be shown that the English translations (by professional or semi-professional translators) all allow for the same relational interpretation. This is not only achieved by particular lexical or grammatical means but by the interaction of a large number of factors, several alternatives being possible; see, for instance, the following example of *doch*:

A: Dienstags um 10 ist bei mir jetzt wiederum schlecht weil ich da noch trainieren bin. Ich denke wir sollten das Ganze dann doch auf die nächste Woche verschieben. Geht's bei Ihnen da?
1) Tuesday at 10 doesn't suit me, at that time I'll still be, em, doing sports. I think we'd better postpone the meeting till the week after. Does that suit you?
2) Tuesday at ten is bad on my part since I will be training then. I think we should rather postpone the whole matter until next week. Is that okay with you?
3) Tuesdays at ten is inconvenient for me on the other hand because I am still off training. I think in that case we should postpone the whole thing to the week after. Would that be all right with you?
With the current utterance speaker A suggests postponing the date to the following week. This contrasts with B's earlier proposal to meet in the current week, but it turned out that no date can be found without sacrifices on either side. The implicit pragmatic pretext is here that B would prefer not to postpone the date. A now uses *doch* to relate the current proposal to the previous, *doch* signalling a relationship of contrast. This is rendered by the translators in the following ways: The translator of 1) uses better, which is relational and, like the modal particle *doch*, compares the current to the non-verbalized previous proposal. In 2) we find another relational concept: rather. The translator of 3) renders the relational structure firstly by means of on the other hand (although there has been no on the one hand), and secondly by in that case, thus implying a relevant alternative to which the current proposal contrasts.

The results from this investigation show that the hypothesis that there is a relational semantic basis of modal particles which explains their pragmatic functions is verified by the contrastive analyses; furthermore, only a contrastive analysis on the basis of a hypothesis regarding the general function of modal particles makes it possible to identify the correct translation equivalences. Thus, not only the relational hypothesis can be verified in this contrastive study, but it is also possible to determine how the functions performed by German modal particles can be realized in English.

**References**


**Pierre-Yves Foucou & Natalie Kübler (Université Paris-Nord, France)**

**Comparaison de verbes en français et en anglais technique**

Dans les universités françaises, les formations en informatique intégrant un enseignement obligatoire de l'anglais spécialisé sont devenues la règle. Or, les enseignants d'anglais n'ont pas toujours les compétences techniques nécessaires, et les manuels que l'on trouve sur le marché, ainsi que les dictionnaires anglais-français de l'informatique sont figés, et donc rapidement obsolètes. L'utilisation de corpus alignés anglais-français permet une description contrastive d'une langue de spécialité, basée sur des documents authentiques.
On sait que les francophones ont des difficultés avec le système verbal en anglais, que ce soit au niveau des constructions syntaxiques, des emplois sémantiques ou du système temporel. L'anglais de spécialité tel que l'anglais de l'informatique, ajoute une difficulté supplémentaire par la création de nouveaux verbes (1), par la création de nouveaux emplois de verbes existant déjà, ou simplement, par l'emploi extrêmement fréquent de certains verbes de l'anglais général. De plus, la traduction des verbes anglais en français ne suit pas toujours leur création, ce qui peut poser des problèmes de compréhension, ou dans un autre domaine, de traduction (cf. les verbes to beam, to post et to amp off ci-dessous).

(1) to boot = amorcer
to beam = envoyer un message par courrier électronique
to amp off = faire tourner en arrière-plan

(2) to post = envoyer un message à un forum de discussion
to run = tourner, faire tourner, lancer

(3) to access = accéder

Disposant de corpus alignés anglais-français, nous examinons les verbes anglais et les comparons avec les verbes français.

Nous décrivons tout d'abord les structures syntaxiques des verbes pour ensuite relever les ressemblances et les différences entre les deux langues. Les différences nous permettront d'établir une liste des difficultés, donc des erreurs potentielles des francophones en anglais. Disposant d'un corpus de textes rédigés par des francophones en anglais, nous pouvons vérifier quelles erreurs se produisent réellement. Dans l'exemple (4), le verbe accéder en français a une structure prépositionnelle à deux arguments, le verbe anglais access a une structure à deux arguments sans préposition. La comparaison nous permet de prédire l'adjonction par le francophone d'une préposition to en anglais, ce qui est vérifié dans notre corpus d'erreurs.

(4) SQL allows users to access data in relational database management systems
    SQL permet aux utilisateurs d'accéder aux données des SGBD relationnelles
    *SQL allows users to access to the data in relational database management systems

Les verbes ayant plusieurs emplois peuvent aussi poser des problèmes aux francophones. Le verbe to run par exemple correspond au verbe tourner, au causatif faire tourner ou à lancer :

(5) dosemu needs to run as root
dosemu a besoin de tourner avec l'identité "root"

(6) Running dosemu on an xterm or in X Windows (…)
Lancer dosemu dans un xterm ou X Windows (…)
Les critères syntaxiques ne suffisent pas ici à faire la distinction. On pourrait penser que la présence de la préposition on en (6) explique la traduction lancer du français. Ce n'est pas le cas puisque l'on trouve aussi :

(7) Linux runs on any platform
    Linux tourne sur toute les plates-formes

Nous utilisons l'approche plus sémantique des classes d'objets, élaborée au LLI par G. Gross pour affiner la description. Cette approche nous permet de différencier les emplois sémantiques des verbes en fonction de leurs arguments. La différence d'emploi en anglais, et, par conséquent, l'utilisation de deux verbes différents en français pour run s'explique par le type d'arguments admis en position sujet; en (6), le sujet doit obligatoirement être un humain, implicite ou non, alors qu'en (7) la position sujet peut être remplie par une classe d'objet que nous pouvons qualifier de "exécutable" (i.e. qui englobe les programmes, les systèmes d'exploitation, les commandes).

Les outils informatiques utilisés pour la consultation des corpus consistent en un concordanceur utilisant des expressions régulières et des catégories syntaxiques, ainsi que des outils d'étiquetage de textes et d'analyse statistique. Les résultats du concordanceur peuvent être traités et manipulés pour produire des exercices portant sur les problèmes relevés par la comparaison.

Jacques François (Université de Caen, France)

La transitivité indirecte dans la traduction d'allemand en français :
L'incidence de caractères ontologiques de la prédication verbale

Selon G. Lakoff (1977) les rôles sémantiques Agent et Patient sont des "gestalts" linguistiques, c'est-à-dire des catégories présentant un noyau prototypique et des variantes plus ou moins abâtardies. Cette conception a été développée dans la théorie multifactorielle de la transitivité de P. Hopper & S. Thompson, dans celle des macro-rôles de R. van Valin, et celle des rôles prototypiques de D. Dowty, qui prévoit une sélection variable du sujet et de l'objet lorsque certaines des propriétés prototypiques de l'agent font défaut.

Dans l'esprit de ces travaux, je cherche à tester l'hypothèse de l'existence de facteurs ontologiques de transitivité indirecte (transitivité "bâtarde") susceptibles de fonder cognitivement une observation contrastive : La sous-classe des verbes allemands de valence 2 à régime au datif dont le sujet est éventuellement ou obligatoirement prédicatif (c'est-à-dire exprimé par une complétive, une infinitive ou une nominalisation d'événement ou d'action, [N/Prop:x V N_datif:y] ex. "es fällt mir ein, dass ich ihren Brief nicht beantwortet habe" est préférentiellement traduite en français par des verbes (ou équivalents fonctionnels) à régime transitif indirect [N/Prop:x V à N:y].
Gertrud Gréciano (Université des Sciences Humaines de Strasbourg, France)

**Quel corpus pour la phraséologie comparée?**

Pour la phraséologie, le regard interlangagier est tentation, plaisir et nécessité. D’abord il s’est fixé sur du matériel statique pour vite glisser vers des données dynamiques. On sait aujourd’hui que le système ne se construit que selon les modalités de l’emploi. Evolution du corpus lexicographique avec la saisie aussi électronique des propriétés du signe phraséologique, de son signifiant et signifié, vers des corpora discursifs avec la saisie surtout manuelle du comportement phraséologique dans les textes et actes en situations parallèles. Vérifications empiriques. Réponses d’un bi-texte nouveau aux attentes actuelles de la phraséologie comparée.

André Hantson, (Facultés Universitaires Notre-Dame de la Paix Namur, Belgium)

**The English -ing Form. A Contrastive and Corpus-Based Approach**

The first part of the paper will discuss the syntax of verbal ing-form constructions and contrast them with related constructions / equivalents in a number of other languages. Traditionally, verbal ing-forms are analysed as being either participial (e.g. while reading that paper I realised that ...) or gerundial in nature. As regards the latter, it is essential to
make a further subdivision into nominal gerunds (e. g. a more careful reading of that paper would show that...) and verbal gerunds, i.e. gerund clauses (e.g. I don’t mind reading that paper more carefully). It is true that the need for the distinction between (verbal) gerunds and present participles was called into question a few decades ago by Kruisinga and Erades and more recently by Quirk et al (1985). But in this paper the view will be defended that the distinction needs to be maintained for reasons relating to both the coherence of the syntactical system of English (the most important argument being the possibility of using a possessive determiner with a verbal gerund but not with a present participle) and contrastive linguistics. As for the latter, English ing-clauses will be compared with Latin gerunds, gerundives and present participles, Spanish gerundives, French gerundives and present participles, Russian gerundives and present participles, Modern Greek gerundives, German (and Dutch) constructions of the type ‘Dieses so nachdrücklich immer dasselbe wiederholen geht einem auf die Nerven’ ‘That so emphatically always the same thing to repeat gets one on one’s nerves’, a construction that invites comparison with an English instance such as She was fed up with it all: the getting him off to his office, the keeping in clean shirts... Also the Norwegian translations of the gerunds in the following will be discussed: (a) Being tied physically gives you lots of scope mentally, (b) Tell me now more about your husband wanting to murder you. The translations, which are taken from a small corpus collected from Norwegian translations of English novels, are the following (a’) Det aa vaere bundet saann i rent ytre mening kan gi en indre frihet ‘It (or ‘the’?) to be tied physically gives you ...; (b’) ‘Fortell naa mere om dette at Deres mann vil myrde Dem!’ ‘Tell now more about that that your husband wants to murder you’. They will be briefly compared with similar constructions in the other Scandinavian languages and Russian.

The second part of the paper will focus on the findings of an investigation of a one million word corpus consisting of 750,000 words of written English (taken from novels, newspapers, academic texts, plays) and 250,000 words of spoken English (taken from radio discussions and interviews). Apart from highlighting statistically significant findings, mention will be made of constructions that either need to be studied in considerably larger corpora or require comparison with data from French, Scandinavian, Spanish, etc. corpora. Moreover, attention will be drawn to constructions that may have been becoming either increasingly or decreasingly frequent over the last few decades. In particular it will be shown that it would be crucially important to check whether there has been any considerable decrease in the frequency of possessives as contrasted with personal pronouns (in sentences of the type I don’t mind him / his doing that), which in my corpus- which is representative of British English as used in the fifties and sixties - were about equally frequent, whereas NPs in the genitive were far less frequent than NPs in the common case. Indeed, for there to be such a dramatic decrease would mean that the basis of the gerund-participle dichotomy was being gradually eroded and a merger might take place in the not too distant future.
Liesbet Heyvaert & Kristin Davidse (Katholieke Universiteit Leuven, Belgium):

On the so-called 'middle' construction in English and Dutch

Constructions such as This peach eats easily, Sheila seduces easily have, in the English tradition ever since Sweet (1892), been analyzed predominantly (Jespersen 1927, Halliday 1967, Fawcett 1980) as having a 'passive ' meaning. In this type of analysis, the following elements tend to recur:
- the Subject is a Patient;
- an Agent, distinct from the Patient is inherently implied but cannot be explicitly expressed by this construction;
- the construction expresses the 'facility’ with which the process can be carried out, or the ‘aim’ for which it is carried out;
- this ‘facility’ or ‘destiny’ can be expressed by Circumstances (easily, well, etc.), modal auxiliaries (will, can, etc.) or simply by the polarity of the finite verb (These shoe laces won’t tie. - Yes, they tie!)
- specific characteristics of the Patient/Subject are held responsible for the facility or destiny associated with the process.

While subscribing to the correctness of all these descriptive claims, we will argue that the passive analysis itself faces a number of problems.

One of the most important problems is the presence of the active finite verb. The authors referred to above have all stated that the construction has a passive meaning in spite of its active verb, but they have never proposed a positive motivation for it. They have remained equally vague about the alleged passive meaning of the construction: if anything, as suggested by their paraphrases, they seem to be thinking of something like the ordinary passive of ‘an action being effectively carried over from Agent to Patient’ (e.g.‘this peach can be eaten easily by me’). Directly connected to this is the explicit or implicit claim that the construction requires a transitive verb.

However, as we will argue, examples such as That road drives well, That race track runs easily and That pitch bowls well, show that it can also be construed with intransitive verbs. On this point an interesting light can be thrown by a comparison between English and Dutch. Not only has descriptive work on Dutch grammar explicitly recognized that ‘circumstantial middles’ such as Die stoel zit gemakkelijk illustrate an intransitive subtype of the middle, further evidence for the existence of middles with an intransitive verb is provided by examples such as Dat asfalt rijdt goed, Die berg klimt goed, in which, despite the existence of the overtly marked transitive verbs berijden and beklimmen, the intransitive verb is used. Clearly, the extension of the category is broader than assumed in the traditional English ‘passive’ analysis: what typifies this construction is a non-agentive Subject, which may be a Patient or a Circumstance, on the one hand, and a non-passive finite verb form, of either a transitive or intransitive verb, on the other. Once this has been acknowledged, it becomes clear that examples such as That rock music dances well,
Pencils do not smudge easily and Narrow tyres manoeuvre more easily should, against the position held by adherants of the ‘passive’ analysis, be considered as instances of the same construction.

In the approach which we will propose to this construction type, we will argue that two constructional layers should carefully be distinguished: the utterance’s actual syntactic categories (Subject-Finite/Predicate) and its process-participant roles. It is the Subject-Finite/Predicate pattern that construes the Subject (in virtue of its specific characteristics) as ‘responsible’ for the facility or destiny of the process. In this context, we will elaborate on our interpretation of the Subject as the ‘instantiator’ of the utterance. In terms of process-participants roles, we will propose that one should at least consider the possibility that the active lexical verb is primarily elaborated by the inherent - transitive or intransitive - Actor, rather than by the Patient or Circumstance, which would give the construction a more ‘active’ flavour than has usually been assumed. This ‘active’ interpretation of the construction is foregrounded more clearly in examples
- which contain transitive, rather than ergative, verbs, e.g. Pan cleans easily;
- which are less ‘generic’ than the stereotypical ones usually considered, e.g. When he scraped a fingernail over the blotchy stains, they sloughed off easily;
- which contain ‘mental’ or ‘verbal’ process verbs, e.g. That story tells well, That doesn’t listen so bad.

This illustrates again the need to investigate a wide, representative, range of corpus examples, rather than the ‘garden selection’ of concocted examples traditionally considered in the literature.

References


Stig Johansson (Oslo University, Norway)

On the Role of Corpora in Cross-linguistic Research

The availability of bilingual and multilingual corpora in machine-readable form have opened up new possibilities for language comparison. It may even be justified to talk about ‘contrastive linguistics in a new key’. While traditional contrastive studies often focused on a comparison of systems in the abstract, we can now more easily compare languages in use, with reference to parallel corpora in two or more languages. But how do we design such corpora? I will present the model chosen for the English-Norwegian
Parallel Corpus, which incorporates both original texts and translations (English to Norwegian and Norwegian to English) and comparable original texts in the two languages. In other words, it is not necessary to choose between a ‘translation corpus’ and a ‘comparable corpus’. Both can be combined within the same overall framework, and each can then be used as a means of controlling and supplementing the other, as I will show with reference to examples from our corpus.

I will also touch on the recent expansion of our project to include other languages, making it possible to compare across a number of languages using English original texts as a starting point. The project will focus particularly on a comparison of English, Norwegian, and German, with special reference to (a) word order and information structure and (b) expressions of modality.

Philip King (University of Birmingham, Great Britain)

*Level playing fields and drops in the ocean: the translation of metaphors and idioms in the proceedings of the European Parliament*

Metaphors and idioms are both characterised by the fact that the meaning of the whole is not the sum of the meanings of the parts; nor in many cases can it be inferred simply from summing the parts. An added dimension to this problem for translators is what it is possible to replicate and how in the TLs. The corpus on which this paper is based consists of transcripts of European parliamentary debates in eight of the EU languages. Thus for every SL instance of an idiom or metaphor there are seven translations. This paper attempts to classify the techniques used by translators on expressions ranging from those like "crocodile tears" which translate painlessly into every language exemplified, to "level playing field", which shows signs of extensive reworking in every translation.

Olivier Kraif (Université de Nice Sophia Antipolis, France)

*Réflexions autour des concepts de correspondance lexicale et d'alignement textuel*

**INTRODUCTION**

Le bi-texte, constitué de deux textes parallèles dont l’un est traduction de l’autre, et dont les segments correspondant sont appariés par une fonction d’alignement, est un objet empirique intéressant dont les applications sont nombreuses: aide à la traduction, terminologie différentielle ou lexicologie, enseignement des langues. Dans le but d’évaluer plus rigoureusement les systèmes capables d’aligner deux textes automatiquement, nous éclairons ici les concepts d’alignement textuel et de correspondances lexicales.
I Le concept d’alignement
Aligner, c’est apparier des portions de texte en relation d’équivalence traductionnelle. Le concept d’alignement, tel qu’il est couramment présenté dans la littérature, repose sur une hypothèse: la compositionnalité de la traduction [Isabelle]. En outre deux conditions sont requises pour le parallélisme des textes: la quasi-bijectivité (appariements bi-univoques des segments) et la quasi-monotonie (séquence des appariements respectée). Suivant la taille des segments impliqués, on parle d’alignement au niveau des paragraphes, d’alignement au niveau des phrases, voire même d’alignement au niveau des mots.

II Le concept de correspondance
Ainsi, au niveau sub-phrastique, on a coutume d’apparier des unités lexicales en relation de traduction, et l’on parle des correspondances comme d’un cas particulier de l’alignement, au niveau des mots. Cette conception se heurte à deux problèmes étroitement imbriqués:
♦ Problèmes de segmentation: la réalité des traductions étant bien loin du mot à mot, on se voit contraint de mettre en correspondance des unités hétérogènes le long d’un continuum: mots, mots composés, syntagmes, phrase entière. Selon quels critères objectifs décide-t-on de décomposer une expression ou de la considérer comme un tout? En outre, ces appariements sont dépourvus de portée générale, car ils ne dépendent pas de constantes des systèmes lexicaux, mais de l’isomorphie structurelle de la traduction, i.e. des choix particuliers du traducteur, de sa manière d’interpréter et de restituer une réalité extra-linguistique (qui sert toujours de plan de médiation dans l’activité traduisante), de son idiosyncrasie.
♦ Problèmes de divergence sémantique: il est fréquent que deux phrases en relation de traduction présentent des contenus sémantiques différents. La compositionnalité traductionnelle atteint souvent ses limites en deçà de la phrase. Or ces divergences forment un continuum: à partir de quels moments considère-t-on l’écart sémantique trop important pour refuser l’appariement?

III Correspondance lexicale vs alignement maximal
Ces problèmes trouvent un début de solution dans la distinction les correspondances lexicales et l’alignement maximal.
Les correspondances lexicales concernent des unités lexicales (lexèmes, phrasèmes et semi-phrasèmes au sens de Mel’cuk), reconnues comme telles dans chacune des langues séparément, se recouvrant au niveau de leurs désignations extra-linguistiques (concept, référent) dans le contexte de la traduction. Tandis que l’alignement maximal est un type d’alignement, tel que défini précédemment, recherchant les segments les plus petits possibles reconnaissables comme équivalents traductionnels, dans leur contexte: il se situe à la limite de la compositionnalité traductionnelle.

CONCLUSION
La distinction ainsi établie, en brisant l’illusion naïve que la traduction puisse se fonder sur des correspondances mot-à-mot, permet deux ouvertures:
- la détermination de critères plus précis pour établir des corpus de référence utilisables dans l’évaluation de l’alignement ou de l’extraction de correspondance lexicale.
- la juste interprétation de ce que constitue un bi-texte: une collection de segments textuels reliés entre eux non par une quelconque propriété linguistique, mais par l’intermédiaire d’un plan extra-linguistique où se joue l’équivalence traductionnelle, dans
le mixte de contraintes contextuelles et pragmatiques dont est tributaire l’acte de traduire, en tant qu’acte interprétatif.

Belén Labrador de la Cruz, (Universidad de León, Spain)

A Corpus-driven Contrastive Project (English-Spanish): A Quantitative Study on Quantification

As in many research centres and universities all over the world, some scholars at the University of León are adhering to the principles and methodology of corpus linguistics. A corpus-driven contrastive project was begun in 1994 and four years on, it continues to develop both in the number of researchers involved and the amount of research produced. We have been granted funds by the Government both to individual members and to the whole project and several doctoral theses are currently being written on contrastive studies of particular communicative functions compared in two languages - English and Spanish.

Two monolingual corpora are being used, due to the lack of a bilingual English-Spanish corpus specifically-designed for this purpose. These are the Bank of English of CoBuild (see website http://titania.cobuild.collins.co.uk/) and CREA (Spanish Corpus of Current Reference), created by the Royal Academy of the Spanish Language. They are both large and representative corpora, composed of a variety of genres and include written as well as oral texts.

The final aim of the project is to serve practical purposes in the fields of Second Language Teaching and Translation Training. Teaching materials dealing with English-Spanish Contrastive Studies in Translation and English degrees are badly needed in Spain at present. Contrastive grammars and lexicons based on real use, which focus on differences in the usage of words with similar meanings are in high demand and would undoubtedly be very helpful.

Some of the semantic areas currently being studied by the members of the project are: quantification, degree, characterization, modality, discourse markers, aspect, tense and voice. This paper will summarize part of the research carried out on quantification and will show some of the results obtained with the comparison of English and Spanish universal quantifiers. It is a two-to-three relationship as only two words, *todo* and *cada* translate the meaning of all, every and each. Whereas inflectional variations are relevant in this case in Spanish (morphemes of gender and number: *todo*, *toda*, *todos* and *todas*), derivation occurs in English (*everybody*, *everything*, *everywhere*). I will describe and classify their most frequent collocational patterns and give an account of the analysis and results of the study. One of the most relevant conclusions is that even though some functional words across languages seem to be approximately equivalent in meaning and, in most cases, translatable, their linguistic behaviour is not parallel, which is the main
source of difficulty and leads to the production of unnatural and atypical utterances both in the second and in the target languages.

Lan Chun, (The Hong Kong Polytechnic University, Hong Kong)

A Corpus-based Contrastive Study of Spatial Metaphors in English and Chinese

This research takes a corpus-based approach to an investigation of spatial metaphors in English and Chinese. The main objectives of the research are, to find out in real life English and Chinese, (1) along which metaphorical extensions two English space concepts, namely UP and DOWN, and two Chinese space concepts, namely SHANG (UP) and XIA (DOWN), develop; and (2) what the differences and similarities are between the two English space concepts and their Chinese counterparts.

The English corpus chosen is the 5-million word Word Bank of the Collins Cobuild English Language Dictionary (1996, CD ROM version), from which 5728 instances of up and 4781 instances of down are sorted out through an English concordance searching program. The Chinese corpus is a self built one which contains 1,800,000 characters of written texts taken from newspapers and magazines published between 1st April and 30th June 1998 in Mainland China, Taiwan, Hong Kong and among overseas Chinese communities. Altogether 8079 instances of shang and 4387 instances of xia are sorted out through a Chinese concordance searching program.

About 10% of all the instances of up, down, shang and xia are randomly selected, put into a database and analyzed one by one along the parameters sense, trajector, landmark, path, vertical axis and metaphorical extension. The following findings are obtained:

In both languages, upward trajectories are linked with things considered to be desirable in the culture and downward trajectories are linked with things considered to be undesirable in the culture.

The two English concepts are mainly used to structure changes in four non-spatial target domains, which can be put in order of prevalence as STATES of people/ objects/ events, QUANTITY, SOCIAL HIERARCHY of people/ organizations and TIME.

The two Chinese concepts are mainly used to structure five target domains, which can be put in order of prevalence as ABSTRACT LOCATION, STATES, QUANTITY, TIME, and SOCIAL HIERARCHY.

Both the English concepts and the Chinese concepts can also be elaborated from expressing vertical movements to expressing horizontal or movements.

A high level of correspondence is observed between the findings for the two English concepts and the findings for the two Chinese concepts. This offers evidence for the possible existence of a universal spatial metaphorical system. However, differences also exist both at the generic-level and especially at the specific-level metaphorical extensions.
The above findings are then compared with the results from a dictionary-based study of the same four concepts carried out earlier by the author. A high level of correspondence is observed between the results yielded by the dictionary data and that yielded by the corpus data, although the former seems to produce more specific-level metaphorical extensions than the latter.

François Maniez (Université Lumière Lyon II, France)

Utilisation des corpus informatisés et de mesures de fréquence lexicale pour la résolution de certains problèmes de traduction et de délimitation d'un lexique spécialisé

I - Applications du traitement automatique de données textuelles en langue étrangère
L'auteur retrace brièvement l'histoire des divers projets de corpus informatisés des trente dernières années, leurs applications lexicographiques, ainsi que l'utilisation d'outils de traduction intégrant l'utilisation de corpus bilingues alignés. Les problèmes de création d'un corpus exploitable par un micro-ordinateur de type PC sont également évoqués (taille et composition). Les limites actuelles des logiciels de traduction automatique sont soulignées.

II - Problèmes de traduction
Trois types de problèmes sont passés en revue :
- polysémie de certains items lexicaux
  A partir d'un exemple précis (la traduction en français de la lexic composée sedimentation rate), l'auteur évoque l'utilisation de banques de données intégrant l'indexation des lexic composées et des collocations les plus fréquentes. La prise en compte de la fréquence lexicale des divers sens d'un même terme polysémique est également étudiée.
- ambiguïté syntaxique
  La prise en compte des fréquences comparées de la collocation based on dans deux corpus distincts (langue générale et langue médicale) montre comment une erreur de découpage syntaxique peut être évitée en traduction automatique si l'on intègre au programme une "sensibilité à l'environnement lexical".
- collocations disjointes
  A partir d'un autre exemple (l'absence de détection de la collocation mixed results dans la phrase "Results of trials of selective gut decontamination have been mixed", l'auteur se penche sur les facteurs qui peuvent contribuer à une interprétation correcte de cette phrase par un humain bilingue, et sur la manière dont cette connaissance peut être répliquée par l'utilisation d'outils informatiques.

III - Délimitation du "vocabulaire spécialisé"
A partir de l'étude de deux corpus informatisés de taille comparable et portant sur le même sujet, l'un en langue générale et l'autre en langue médicale, l'auteur pose le problème de la délimitation du lexique de la langue de spécialité. L'efficacité de diverses méthodes de partition est discutée, ainsi que l'utilité de la délimitation d'un niveau intermédiaire, celui d'une langue "semi-spécialisée", dont le lexique est l'intersection de la langue générale et de la langue de spécialité. Les éléments de ce lexique ont un comportement spécifique de cet environnement lexical (réalisation de certaines extensions à l'exclusion des autres pour les termes polysémiques, combinaison en collocations spécifiques). L'auteur tente de démontrer que les items lexicaux de cette langue semi-spécialisée gagneraient à être intégrés aux dictionnaires bilingues des divers domaines de spécialité ou à faire l'objet d'un traitement distinct, dans la mesure où bon nombre de traductions spécialisées sont réalisées par des spécialistes d'un domaine scientifique ou technique dont la connaissance lexicale de la langue de départ n'est pas équivalente à celle d'un natif.

Mirela Moldoveanu (Université du Bas Danube, Roumanie)

Phraséologies marquées [+ affectif]. Domaine franco-roumain

Le but de notre contribution est de réaliser un corpus franco-roumain des idiomaties qui servent à l’expression de l’affectivité dans les deux langues.

Dans la littérature de spécialité en français et en roumain, il y a beaucoup d’études qui portent sur l’expression des sentiments, l’un des domaines les plus vastes qui s’offrent à la recherche linguistique, lieu de rencontre des approches lexicale, sémantique, syntaxique, pragmatique. Mais les études contrastives sont très peu nombreuses et portent surtout sur les aspects syntaxique et pragmatique ; de plus, un corpus franco-roumain dont le centre d’intérêt soit les lexies marquées [+affectif] n’existe pas encore.

Notre point de départ est un corpus constitué en principal par des textes littéraires, mais aussi par des extraits de presse.

Notre intervention comporte deux grandes parties. Dans la première, nous essayons de dresser un inventaire si complet que possible des phraséologies qui expriment ou décrivent des états affectifs. Le principe qui préside à la systématisation de cet inventaire est la taxinomie des émotions que nous empruntons aux recherches menées par les psychologues.

Dans la deuxième partie, nous visons à analyser les équivalences et les différences qui se manifestent dans les deux langues. Il y a deux aspects principaux que nous prenons en considération à ce niveau :
1. Les éléments constitutifs des unités du corpus. Nous entreprenons une étude contrastive de la nature de l’élément central des idiotismes (verbe ou nom) et de leurs particularités syntaxiques (la manière dont les unités inventoriées répondent aux tests de la commutation, de la suppression des constituants, de la pronominalisation, de l’impérativisation, etc). Cette démarche est incontournable si l’on veut établir un tableau des variantes de traduction qui enregistre autant l’existence de ces structures en langue, que leur fonctionnement au niveau discursif.


Nous considérons que notre contribution pourrait être un point de départ pour un inventaire complet des lexies marquées [+affectif] en français et en roumain, susceptible de devenir un outil de travail indispensable pour la traduction des textes littéraires.

Dirk Noël, (University of Gent, Belgium)

**English passive matrix clauses as evidentials: empirical contrastive support**

English verbs of the believe type, which alternate finite with infinitival clausal complements, more often take the latter complement as passives than as actives, and sentences like (1)-(2) are on the whole far more common than sentences like (3)-(4).

(1) Magnesium is thought to affect the heart muscle. (BNC A2P 346)
(2) Car dealers are found to offer better deals to white people than black. (BNC ABK 54)
(3) If the decision is illegal it can be quashed; otherwise the court cannot (with one exception) intervene, even if it thinks the decision to be wrong in some respect. (BNC EBM 94)
(4) We have found causation to consist, at bottom, in connections between particulars. (BNC EVX 747)

There are good information/thematic structural reasons for this (see Noël 1998), but the present popularity of the passive pattern may also be a concomitant of a grammaticalization (auxiliarization?) process which is making the passive pattern available as an option in systems of evidentiality (cf. Chafe & Nichols, eds., 1986).

Interestingly, the Dutch and French prototypical equivalents of believe-type verbs do either not alternate finite complements with infinitives (Dutch) or do so only under strict syntactic conditions (French) (see Noël & Defrancq 1996). Building on observations by
Halliday (1985) on ‘interpersonal grammatical metaphor’, and on work by Thompson & Mulac (1991) on ‘epistemic parentheticals’ and by Verhagen (1996) on Dutch matrix clauses, and continuing my own work on the distribution of infinitival and finite complements after English verbs of the believe type (Noël 1997, 1998), this paper will use contrastive data from the INTER SECT, TRIPTIC and Canadian Hansard corpora to attest the evidential status of the passive matrices of infinitival complements in English.

References

Noël, D. & B. Defrancq (1996) To be or not to be after believe/croire-like verbs. Contragram 8: 4-6.

Michel Paillard (Université de Poitiers, France)

Des figures de style aux unités lexicales : approche contrastive de quelques métonymies et hypallages en anglais et en français


La métonymie désignant un animé humain ou une institution par le biais d'un objet concret qui lui est associé (les gros bras, le petit écran) est particulièrement productive en anglais (jet set, shopping-bag lady, strap-hanger ; bank holiday, redbrick university, meals on wheels). De l'abstrait au concret, en revanche, l'anglais résiste à la désignation par un prédicat nominalisé de l'agent, l'objet, le lieu ou le temps du procès. (société de consommation : consumer society ; à la réception : at the reception desk ; l'allongement de la scolarité : the raising of school-leaving age). Ceci est à relier à la problématique générale des noms abstraits et de la concrétisation (cf. Defrancq & Willems 1996) et à la réticence de l'anglais à mettre en relation des notions ayant des propriétés hétérogènes (cf. Guillemin-Flescher 1981, Celle 1997).

L'hypallage consiste en un décalage entre portée syntaxique et portée sémantique d'une qualification. Ici, c'est le français qui résiste à la distorsion, parfois sur le plan textuel (after ten rainy miles : après avoir parcouru dix miles sous la pluie) et, plus encore, dans le cas de l'hypallage lexicalisé (lucid interval, married years, sick bag). De tels groupes
nominaux lexicalisés présentent une ellipse syntaxique (peu courante en français dans ce schéma N+Adj: *panne sèche*) caractéristique de la composition en général. Larreya & Méry (1992) montrent qu’en tant que déplacement de portée, l’hypallage est un processus omniprésent en anglais jusque dans les transformations syntaxiques de "montée" (*it was found to be missing*).

Références bibliographiques


Elena Paskaleva (Bulgarian Academy of Sciences, Bulgaria)

**Linguistic Boundaries in Data-driven Processing of Bilingual Texts**

Within the frameworks of two Copernicus’94 projects a large volume of bilingual corpora – English-Bulgarian, French-Bulgarian and French-English were collected, aligned and processed on the data-driven platform. Two main operations were fulfilled by specially designed software systems:

- **MARK ALISTER** – a system for alignment and markup of parallel texts with a procedure of searching translation equivalents – the searching engine points out the segment where the equivalent is located.

- The extracting module of **MARK ALISTER** fulfils the real extraction of translation equivalents. It is based on the parallel calculus of the co-occurrences of complex terms components in the aligned texts.

An evaluation of the ‘linguistic power’ of both procedures is made that takes into account:
The immanent possibilities of the method used: the alignment procedure is based on Gale-Church algorithm with its genuine constraints; the extracting procedure being statistically based requires a minimal number of cooccurrences and therefore a large volume of processed material.

The portion of linguistic knowledge participating in the procedures: in the sentence alignment only some linguistic refinement of the sentence boundary setting can improve the results; in the extracting procedure where the lemma indexing of both texts is an obligatory module, the full tagging and disambiguation of both texts will reduce the noise in the set of extracted translations.

The type of the processed texts – investigation of the error rate is made for different types of texts – fiction, scientific, legal.

The language pairs – better results are obtained in the extraction of translation equivalents in the English-French than in the French-Bulgarian experiment.

The well-known correlation (inversely proportional) between:
the depth of linguistic knowledge (involved in and obtained by the operation) and
the universality of the procedure lowers the general benefits of the data-driven approach – its language independence and large scope.

An attempt to break this stable balance between the “large” and the “shallow” is made in the design of the system interface – ergonomized and user oriented. The editing and updating facilities, the variety of processed text formats and the rich concordancing support in different stages of the procedure make MARK ALISTER (with or without its extracting module) not only an interesting tool for linguistic and translation research but also a basic environment for parallel texts processing.

References


Hans Paulussen, (Facultés Notre-Dame de la Paix Namur, Belgium)

TRIPTIC: a trilingual database for the contrastive analysis of prepositions

In foreign languages teaching and translation training, the learning of prepositions and particles is often restricted to mere rote learning, because these so-called function words are generally considered devoid of any meaning. Over the last few years, however,
numerous studies in cognitive linguistics have revealed that these grammatical words show a strong polysemous structure, the meanings of which are interrelated. Since these studies are mainly based on introspection and the use of questionnaires, the results are biased towards potential usage, rather than real usage. Moreover, in most cases the studies are limited to intralingual analysis, so that it remains unclear in which way two or more languages diverge or converge in prepositional meanings.

In order to validate the cognitive approach on an empirical and contrastive basis, which assumes the use of large scale authentic linguistic material in different languages, a trilingual corpus in English, Dutch and French has been compiled at the University of Namur. The Namur corpus, containing 2,000,000 words equally divided over fiction and non fiction texts, is parallel aligned at paragraph level, so that for each nth paragraph in one language the same nth paragraph in the other two languages can be selected automatically. This source corpus consists of a collection of text files and therefore can easily be explored by using simple text manipulation tools, such as GREP and AWK. Nevertheless, such explorations are limited to general quantitative queries. In order to facilitate a detailed semantic and contrastive analysis, a separate trilingual database TRIPTIC —which stands for TRilingual Parallel Text Information Corpus— has been developed which allows fast selections of records on different search keys. Although the TRIPTIC database has been designed specifically for the contrastive analysis of prepositions and particles in English, Dutch and French, the architecture is such that the same principles can be used for similar types of contrastive research projects.

At present, the TRIPTIC database contains a sample set of the Namur corpus, which was used for the analysis of the preposition/particle up and on in English and their equivalents sur and op in French and Dutch. Within the limits of the source corpus, the analysis shows a number of important features, on the interlinguistic and intralinguistic level, which would not have been possible if the research had only been based on introspection and small sized data.

For example, the corpus data clearly show that although on and up can both be preposition or particle, the first is more often used as a preposition, whereas up is a typical particle. In the case of Dutch, op is almost equally used as preposition and particle. Since French, similar to the other Romance languages, does not use the particle system common to English and Dutch (and the Germanic languages in general), it is not surprising that this use is a recurrent stumbling block for French learners. The corpus also reveals that on/up and op in English and Dutch are more frequently used than sur in French. On the basis of the data stored in the TRIPTIC database, it has been possible to analyse the meanings effectively used in the different languages, and to show in which way a preposition in one language is rendered in the other two languages.

In this talk, I will give an overview of the Namur corpus and show the link with the TRIPTIC database. This is followed by a number of examples from both the intralinguistic and the interlinguistic analysis. These illustrations will make clear that parallel corpora are valuable empirical data in contrastive linguistics research.
Tiina Puurtinen, (University of Joensuu, Finland)

*Using electronic corpora in translation studies: translation universals and translationese*

The paper presents an ongoing research project “Translation Universals and Translated Finnish: a study on corpus data” at the Savonlinna School of Translation Studies.

The project carries out basic research on the nature of translated text: the aim is to compare, with the help of computer software, the language of translated and non-translated Finnish texts and to find potential differences between these two. The corpus is used to test hypotheses proposed by several translation scholars about “universals of translation”, linguistic features which are assumed to be common to all translations regardless of source and target languages. These hypothesised universals include simplification, explicitation and normalisation. Indicators of simplification, for instance, might be a lower type-token ratio and a lower lexical density in translations than in non-translated target language texts (Laviosa 1996). In addition to testing universals, the project also investigates features of translationese, i.e. linguistic differences between translated and non-translated language on the basis of Finnish.

The paper will report on some preliminary findings made on the basis of the current state of the corpus, and discuss some advantages and limitations that the approach confers to the study of translation universals and specific features of translationese.

The Savonlinna Corpus consists of Finnish translations and originally Finnish texts in electronic form, i.e. it is a “comparable corpus” (Baker 1995), and the source texts of the translations are not included in the corpus. The collected texts represent three genres: adult fiction, children’s fiction, and academic literature. The source languages of the translations are English, German and Russian.

Raphaël Salkie (University of Brighton, Great Britain):

*How can we unpack a translation corpus?*

Translation corpora are often regarded as problematic: questions of quality, representativity, and “translationese” are raised. It is more useful, however, to see a translation corpus as an annotated monolingual corpus, similar to a lemmatised or tagged corpus. The annotation is unusual in that it is itself a corpus, but in principle this is no different from practice in the Text Encoding Initiative, where information about a text is stored in a Document Type Definition (DTD), which can be a separate file from the document itself.
Questions about the quality of a translation corpus can now be reformulated as questions about the quality of the annotation. More significantly, we can ask:

1. What kind of annotation is present in a translation corpus?

Eugene Nida has argued that translators engage in three activities:

- Analysis of the source text
- Transfer of the analysed text into the target language
- Restructuring of the target text

A translation corpus will therefore contain the results of these operations.

2. How can we extract the information represented in the annotation?

The basic methodology is to compare the information in a monolingual corpus with the information in a translation corpus. To find information in a monolingual corpus we rely on the words in the corpus and their collocations. To find information in a translation corpus we rely on translation equivalence. This paper will look at three examples of the methodology:

- Word senses in lexicography
- Word sense disambiguation (WSD) in statistical corpus analysis
- Discourse connectives in English and French

Jean Senellart (Université Paris 7 et Université de Marne-la-Vallée, France) & Cédric Fairon (Université Paris 7, France et Université Catholique de Louvain, Belgium)

**Classes d'expression bilingues gérées par des transducteurs finis (FST), dates et titres de personnalité (anglais-français)**

Le point de départ de cette étude est la description complète que l'on possède dans chacune des deux langues de classes d'expressions :
- les adverbes de dates (M. Gross) productifs ou figés (ex. : *tard dans la soirée du 17 janvier 1997, à la saint-Glinglin*)
- les titres de personnalité (ex.: *le ministre français délégué à la culture, the Chancellor of the Exchequer*)

Cette description est réalisée à l'aide d'automates finis et construite à partir de la consultation de vastes corpus informatisés. Elle est directement utilisable dans le logiciel de traitement de corpus INTEX (M. Silberztein). Chacune de ces descriptions représente dans ce système plusieurs centaines de graphes interdépendants.
Nous avons étudié la mise en correspondance de ces deux descriptions à l'aide de transducteurs obtenus à partir des bibliothèques d'automates des deux langues respectives. Cette mise en correspondance passe par la production d'un code intermédiaire formalisé et commun aux deux langues.

Grâce à cette représentation, il est possible de mettre en correspondance des expressions bilingues ou monolingues telles que :

- about every month in the coming few years à peu près tous les mois dans les ans à venir
- at the beginning of each january tous les ans, au début de l'année
- the Chancellor of the Exchequer Mr. X Finance Minister Mr. X. of Great Britain

Ce système ne règle pas la question du choix de la traduction. Plusieurs traductions sont proposées en sortie et il revient à l'utilisateur de faire le choix.

Notre étude, basée sur la description linguistique d'expressions complexes (bien au-delà des mots simples et même des mots composés), permet d'obtenir une traduction de très bonne qualité. Et la méthodologie utilisée est extensible à tous les phénomènes linguistiques complexes admettant une description par automates finis. Le LADL a entrepris depuis plusieurs années une telle description, essentiellement pour le français (d'autres langues sont également en cours de description par d'autres laboratoires : anglais, italien portugais, espagnol, etc.). Comme nous le montrons dans notre étude, la description complète d’un phénomène dans une langue est une étape préalable à celle de sa traduction.

Elke Teich, (University of the Saarland, Germany)

Contrastive-linguistic analysis of translations: an integral model of grammatical, semantic, textual and contextual factors

The increasing availability of multilingual corpora, including those consisting of texts and their translations, makes it possible to investigate translation not only locally on a sentence basis, but in actual texts. However, one of the crucial questions before any analysis, no matter whether automatic, semi-automatic or manual, can start, is which categories to use for the analysis.

As is well known, translation requires knowledge not only about the contrasts and commonalities between language systems at the grammatical level, but crucially about differences and similarities when these grammatical means are deployed in texts and according to situational contexts. A model for the contrastive-linguistic description of
translations must therefore take into account not only grammatical differences and commonalities between language systems, but also contrasts in the instantiation of linguistic systems in text and context.

In the paper, I present a model of contrastive-linguistic description that integrates the system and the instantiation view. The model is based on Systemic Functional Linguistics (Halliday, 78; Halliday, 85) and has been developed in the context of automatic multilingual text generation (Bateman, 95; Teich, 95) and functional language typology (cf. Caffarel and Martin, in preparation). I will argue that because of the inherent property of Systemic Functional Linguistics of integrating grammatical, semantic, textual and contextual factors, the model is particularly suitable for the contrastive-linguistic description of translations. The data I use are taken from a small corpus of English and German popular-scientific texts—the English source language texts are taken from Scientific American and the German translations appeared in Spektrum der Wissenschaft.

References


Elena Tognini Bonelli (University of Lecce and The Tuscan Word Centre, Italy)

“Functionally Complete Units of Meaning” in English and Italian: a case study from a Corpus-driven Perspective

In this paper the author describes a corpus-driven approach aimed at identifying correspondences and divergences between two languages. It is shown how, given a set of comparable corpora, a series of methodological steps in generalisation and degeneralisation allow to formalise the contextual parameters of a word at the lexicogrammatical, the semantic and the pragmatic levels; the aim is to identify systematically sets of “functionally complete units of meaning” between the two languages. A unit is seen as functionally complete when it has reached its pragmatic realisation in the surrounding text.

It is maintained that these units, inevitably multi-word units, provide a better base than traditional one-to-one translation equivalents for the creation of a translation data-base. The fact that they are shown to have both semantic and pragmatic realisation makes them ideal candidates for a contrastive approach that takes into account the structure of
discourse and the evaluative stance that a speaker/writer takes with respect to the subject s/he is treating. The method adopted is proposed as a stepping stone towards partial automation of the translation process.

The case study presented deals with some phrases built around the set of words *case-caso*. In particular *in case of*, *in the case of* and *in case* in English and, on the other hand, *in caso* and *nel caso* in Italian. These sets show interesting divergences, of the type that are certainly not identified by current bilingual - or even monolingual - dictionaries but that need to be made explicitly and systematically when one is working on two different languages.

Åke Viberg, (Lund University, Sweden)

**Swedish verbs of motion from a crosslinguistic perspective**

This paper is one in a series of papers dealing with various aspects of Swedish verbal semantic fields from a crosslinguistic perspective with the aim of establishing the lexical profile of Swedish. The typological profile of a language is a crosslinguistically valid characterisation of its structure highlighting in particular what is universal and what is language-specific. The analysis is based on data from parallel corpora and on typological data collected with a questionnaire and on second language learner data.

Certain characteristics of the Swedish system of motion verbs reflect universal tendencies, such as the dominance (in terms of frequency and structural/semantic possibilities) of the nuclear verbs meaning 'go' and 'come' (Viberg 1996a). Other characteristics can be related to typological patterns of variation. With respect to the expression of path (or direction), Swedish is a satellite-framed language (Talmy 1985, Slobin 1996). Languages also differ with respect to the contrast between self-propelled and non-selfpropelled motion. In English, a simple sentence such as Bill went to the railway station is neutral with respect to the means of transportation. In Swedish, there is an obligatory differentiation between the verbs *gå* 'go, walk' which can only represent motion by foot when the subject is human and the verb *åka* 'go/ride in a vehicle'.

Verbs of motion have many extended meanings of various kinds. Some examples will also be given of characteristic patterns of polysemy of Swedish motion verbs.

**References**

**General:**
Earlier studies on motion verbs in the project Crosslinguistic lexicology:


Cornelia Wermuth (Katholieke Vlaamse Hogeschool Antwerpen, Belgium), Werner Ceusters (Language and Computing, Belgium) & Johan Decruyenaere (Universitair Ziekenhuis Gent, Belgium)

Contrastive analysis of the semantic and linguistic structures of surgical procedure expressions based on multilingual data

Goals

As a result of the ongoing specialisation in the field of medicine, different medical sublanguages can be identified. Based on the results of previous analyses of the semantic and linguistic properties of (medical) sublanguages carried out in the field of (computational) linguistics and medical informatics, this paper focuses on the contrastive analysis of written natural language expressions from the subdomain of cardiac surgery as they can be found in the rubrics of classification systems. The multilingual corpus has been analyzed both at the semantic and formal level focusing on the following aspects:

- the analysis of the semantic structures which implies the identification of a set of corpus-relevant concept-types and semantic relationships
- the description of the language specific encoding of the conceptual structures in grammatical form which must illustrate, among others, to what extent a detailed multilingual corpus analysis can support the identification of implicational (i.e. not structural encoded) meaning
- the potential possibilities of computerised corpus analysis illustrated by means of the medical language processing tool “CASSANDRA”

Material

This paper summarizes the results of an empirical study which has been carried out on the basis of a restricted set of multilingual terminological data consisting of 142 cardiovascular expressions in English, Dutch, German and French from the ICD-9-CM classification (9th version, vol. 3, codes 35-39). These so-called “rubrics” are used as natural language definitions of associated numeric codes (e.g.: “35.0: Closed heart valvotomy”).
Method
In the first section of this paper, we focus on the semantic and syntactic level. The corpus specific concept types are identified on the basis of the CEN1828:1995 (1) standard resp. its augmented version (2). In a next step, a set of basic semantic relations is determined. Following the work of Dik, Fillmore, Frawley and others, we assume that these relations specify how arguments relate to the predicative head of an expression. With respect to the assignment of semantic roles (e.g. “Theme”, “Patient”, “Goal” etc.) we proceed from the assumption that they are mediated by grammatical functions at the abstract syntactic level (e.g. the lexical elements which represent the “Direct Object” receive the “Theme” resp. “Patient” role). For each expression, the predicative element is identified together with the corresponding argument structure (i.e. the number of arguments and their semantic roles). This analysis results in a corpus specific set of argument structures. We propose for each of the identified semantic roles a (re)definition which is based on the empirical data as they occur in the corpus. The analysis at the morpho-syntactic level focuses on the description of the linguistic encoding of the identified argument-predicate-structures in order to identify existing language specific divergences. Examples are provided which illustrate how a detailed multilingual corpus analysis is useful both in clearing up structural ambiguities and in identifying implicational meaning in surgical procedures expressions.

In the second section of this paper, the “CASSANDRA” tagging technique is presented. We illustrate its potential for annotating multilingual corpora automatically using concrete examples from the analyzed corpus.

Karin Aijmer, Anna-Lena Fredriksson, Annelie Adel & Mia Boström Aronsson (University of Göteborg, Sweden)

**Contrastive work in progress at the English Department, Göteborg University**

The aim of the presentations is to give a report of work on the basis of the English-Swedish Parallel Corpus in two different areas on the basis of pilot studies.

The passive will be analysed contrastively in several different text types (Anna-Lena Fredriksson). The passive is however a fuzzy category as is apparent from Quirk et al.’s discussion of the passive in English. The contrastive study of the passive in Swedish and English is of special interest since Swedish has several possibilities to express the passive. We can also expect interesting differences since Swedish has an indefinite generic pronoun (‘man’) competing with the passive construction.

Annelie Ådel will discuss metalinguistic comments in the English-Swedish Parallel Corpus with focus on differences in use in different text types. Metalinguistic or self-reflexive comments are a pervasive feature of language. However, they can be expected to be expressed by different grammatical and lexical means when one compares two languages.

**Mirella Conenna (Université de Bari, Italie)**

**Traitement automatique comparé des proverbes français et italiens**

Nous présentons une description systématique d’une classe syntaxique de proverbes français et italiens. Nous avons construit deux dictionnaires électroniques qui font partie de la série DELA, les Dictionnaires électroniques du LADL (Laboratoire d’Automatique Documentaire et Linguistique de l’Université Paris 7).

Nous avons regroupé les formes équivalentes dans les deux langues dans une bibliothèque d’automates transducteurs à états finis. Il s'agit d'un outil efficace pour traiter simultanément les proverbes de deux (ou plusieurs) langues. Les applications sont nombreuses, notamment dans le domaine traductologique ainsi qu'en analyse automatique des textes.

On connaît les problèmes traditionnellement posés par la traduction des proverbes. Un classement homogène effectué selon des critères strictement linguistiques et sur support informatique contribue à améliorer les études parémiologiques.
Elizabeth Dawes, (Université de Winnipeg, Canada)

Les modèles phraséologiques dans le dictionnaire

La recherche récente en phraséologie a révélé que la variabilité est tout aussi caractéristique de la locution que la stabilité. Dans les dix dernières années, plusieurs nouveaux dictionnaires phraséologiques ont été publiés. Malheureusement, aucun ne présente un traitement systématique des variantes phraséologiques.

Dans le but de constituer un dictionnaire de locutions françaises et de leurs variantes phraséologiques, j’ai établi un corpus informatisé d’environ 3 000 locutions relevées dans une grande variété de journaux et magazines français et québécois. Presque 2 000 des expressions relevées ont été retenues pour mon dictionnaire, à savoir toutes celles ayant des variantes phraséologiques, y compris 1) les locutions normatives ayant des variantes normatives (ex. tomber en panne - être en panne - rester en panne); 2) les locutions normatives ayant des variantes québécoises (ex. passer l’éponge - passer l’efface [efface “gomme”]); et 3) les locutions québécoises ayant des variantes québécoises (ex. faire la babouine - faire la baboune “faire la moue, bouder” [babouine, forme archaïque de babine; baboune, création québécoise signifiant “visage, bas du visage, joue, lèvres”, le suffixe -oune existant dans quelque 325 mots québécois]).

L’analyse systématique de mon corpus m’a permis d’établir une typologie de la variabilité phraséologique comprenant une soixantaine de types de variantes qui se laissent classer en deux grandes catégories. Les variantes non fonctionnelles concurrencent la forme de base de la locution sans en modifier ni le sens phraséologique ni l’emploi de la locution et se manifestent aux niveaux graphique (ex. chercher midi à quatorze heures - chercher midi à 14h), morphologique (ex. de tout poil - de tous poils), syntaxique (ex. séparer le bon grain de l’ivraie - séparer l’ivraie du bon grain) et lexical (ex. donner le ton - donner le la). Les variantes fonctionnelles modifient le sens ou l’emploi de la locution et se manifestent aux niveaux morphologique (ex. dernier cri [adjectif] - dernier cri [nom]), syntaxique (ex. taper sur les doigts de qqn - se faire taper sur les doigts), lexical (ex. regarder qqch d’un bon oeil - voir qqch d’un bon oeil) et sémantique (ex. à l’aveuglette “à tâtons” et “au hasard”).

En comparant les variantes présentées dans les dictionnaires avec celles que j’ai relevées dans les textes, j’ai pu constater que les dictionnaires ont une forte tendance à négliger les variantes fonctionnelles morphologiques, syntaxiques et lexicales ainsi que les variantes non fonctionnelles graphiques qui dépendent toutes de règles linguistiques (ex. ponctuation, formation du lexique, aspect du verbe, voix, etc.). Quoique le traitement des locutions comme des unités isolées ait favorisé l’assimilation de la phraséologie au lexique, les locutions entrent aussi dans le domaine de la grammaire grâce à leur statut de syntagmes.
Je montrerai donc comment l’analyse du corpus, qui m’a permis de dégager systématiquement ces types de variantes négligées dans la tradition lexicographique, me permettra de proposer un traitement lexicographique des locutions qui tienne compte des modèles de variabilité phraséologique. Ces modèles, tout comme les conjugaisons que certains dictionnaires tels que le *Petit Robert* signalent pour chaque verbe, seront utiles aux usagers qui cherchent à manipuler les locutions en fonction de leurs propres besoins.

Caroline de Schaetzen (Centre de Terminologie de Bruxelles, Institut Marie-Haps, Belgique)

**Lexicographie sur CD-ROM. Développements récents**

1. **DES DICTIONNAIRES PARTOUT**

Les dictionnaires s’intègrent à d’autres outils. Un exemple est le QUICKTIONARY de WizCom, lecteur optique portable couplé à un dictionnaire électronique et traduisant les mots scannés par l’utilisateur. Un autre exemple est le *Kramer’s*, inclus à la version néerlandaise de l’encyclopédie ENCARTA. Ou encore, nous avons programmé au CTB un CD-ROM enseignant les formants grecs et latins, c’est-à-dire les mots, racines, préfixes, suffixes, infixes grecs et latins contribuant à former des mots dans les langues occidentales vivantes. Nous y avons inclus un dictionnaire de racine, pour faciliter la création lexicale par fixation. Ses entrées sont accessibles de n’importe quel endroit du cours ou des exercices. Comme le tutoriel comporte des termes de lexicologie qui pourraient être inconnus des futurs spécialistes auxquels ce CD-ROM est également destiné (le premier public-cible est constitué d’étudiants en traduction et en philologie classique), nous avons également incorporé un lexique de lexicologie, accessible en mode hypertextuel depuis le tutoriel. Un clic sur le mot *préfixe* fait sauter l’apprenant à l’entrée définitoire correspondante.

2. **UN METIER RENOUVELE**

Les équipes confectionnant les dictionnaires et encyclopédies multimédias sont multidisciplinaires: graphistes, scénaristes, vidéastes, lexicographes. Cette tendance s’observe aussi pour les tutoriels de langues.

On assiste à l’émergence de services neufs : la mise sur CD-ROM de dictionnaires, y compris leur pressage, leur présentation et leur diffusion, par exemple (société AndComplex).

Les logiciels d’ingénierie lexicale se spécialisent. On peut acheter des gestionnaires de fichiers destinés à la confection ou la conversion de dictionnaires pour CD-ROM (MULTITERM DICTIONARY) ou des matrices (coquilles à remplir) d’exercices multimédia de vocabulaire, pour les professeurs de langues spécialisée ou générale (Wida). Des banques de terminologie se dédient à certains profils d’utilisateurs: TIARA est une banque de termes destinée aux seuls interprètes du Parlement européen.
Les auteurs de dictionnaires édités qui souhaiteraient une seconde diffusion, sur CD-ROM, de leur ouvrage, ignorent souvent que leur éditeur ne peut s’opposer à cette rediffusion si le contrat signé avec lui est muet sur ce chapitre: s’il ne les a pas cédés explicitement à l’éditeur par contrat, l’auteur est titulaire de tous les droits dérivés sur son ouvrage.

L’informatique met la charge de la typographie, de la composition et de la mise en page sur les auteurs au lieu des éditeurs. Tous les manuscrits des dictionnaires (comme ceux des autres livres) doivent être remis sur disquette ou transmis par INTERNET. De plus, l’éditeur conditionne souvent un contrat d’édition à la remise du livre en version prête à cliquer.

Les dictionnaires sur CD-ROM présentent l’originalité de vendre le mode de consultation du dictionnaire en même temps que l’ouvrage. C’est un peu comme si les livres devaient être pris de manière différente sur les rayonnages, puis ouverts et lus selon un mode variant de l’un à l’autre...

3. SAVOIRS

Les étudiants des métiers linguistiques ne devraient-ils pas acquérir des connaissances typographiques nouvelles, empruntées notamment à la Gestalt? Savent-ils par exemple que la mémoire humaine ne tolère que sept niveaux de saut hypertextuel d’une entrée de dictionnaire à d’autres, que l’œil lit d’abord ce qui se trouve dans la partie inférieure droite d’un écran? INTERNET a enseigné aux jeunes à enchâsser correctement les opérateurs booléens requises dans les consultation de dictionnaires sur CD. Par contre, les futurs auteurs de dictionnaires n’évitent pas encore les renvois circulaires ou dépourvus de cohérence; or si ces voies sans issue sont acceptables dans les dictionnaires sur papier, elles ne le sont plus sur ordinateur, en raison de la rapidité de ses sauts.

Adela Dragan, (Université “Dunarea de Jos” Galati, Roumanie)

**Valeurs modalisatrices des interjections en français et en roumain**

L’une des situations d’emploi des interjections est celle où l’on exprime une attitude. Le locuteur exprime une attitude (subjective, qui passe par les critères d’appréciation personnels) vis-à-vis d’un référent. Ce référent peut être un Objet, compris comme action, phénomène de la réalité mais aussi comme la qualité et la quantité de l’Objet en question (par Objet, nous entendons tout élément découpageable de la réalité, auquel on fait référence).

Si l’interjection est toujours à considérer comme une réaction (manifestation linguistique de nature réactive), nous avons en vue pour cette démarche les interjections dont l’emploi tient au comportement langagier (emploi conventionnel), et non pas les interjections “sensitives”.

46
Le statut d’“holphrase” de l’interjection la place sous le signe de l’exclamation comme modalité de la phrase, traduisant des nuances de la pensée et des sentiments (désir, souhait, protestation). Plus ou moins exactes, les paraphrases qu’on peut donner des interjections nous aident à identifier les valeurs modalisatrices des celles-ci. A côté des paraphrases on a les énoncés qui suivent une interjection et qui ont généralement, le rôle d’expliciter et de soutenir sémantiquement l’interjection. De cette façon le verbe de la paraphrase ou de l’énoncé qui suit peut marquer, de manière explicite cette fois, l’attitude exprimée. Nous considérons que les valeurs modalisatrices des interjections sont proches des valeurs qu’on attribue aux adverbes de phrase comme: probablement, certainement, heureusement, et qu’elles peuvent entraîner des tours comme: il est heureux que, il est triste que, etc.

Nous avons identifié des ressemblances et des différences à tous les niveaux: sémantique et pragmatique, lexical et syntaxique, tout comme au niveau du registre de langue et de la fréquence.

Nous avons utilisé un corpus représentatif provenant de traductions d’ouvrages roumains et français et nous avons isolé les éléments qui conduisent à des transpositions moins réussies dans la langue cible.

**Ana Frankenberg-Garcia (Insituto Superior de Linguas e Administraçao, Lisbon, Portugal)**

**Using bilingual corpora to produce second language teaching material**

Working with Portuguese learners of English has brought to my attention different types of errors of English which have to do with my students´ prior knowledge of Portuguese. Identifying those errors and interpreting them in the light of contrastive linguistics can help indicate ways in which bilingual corpora can be harnessed to prepare English as a second language teaching materials for native speakers of Portuguese. This poster highlights a few practical examples of the kind of information that a teacher of English to native speakers of Portuguese would like to extract from Portuguese-English bilingual corpora.

**Tove Jacobsen (University of Bergen, Norway)**

**Automating the bilingual dictionary-making process: a case-study**

A Norwegian-French dictionary, which also exists in a machine-readable version, was published in 1991. It was assumed that relevant information could be retrieved
automatically from the Norwegian-French data to form a database for a new French-
Norwegian Dictionary.

In this paper I will describe briefly the sorting, deleting and collapsing operations the
Norwegian-French dictionary was submitted to before a satisfactory result was reached,
and then discuss some of the lexicographic issues involved.

In conclusion, I will look at other possibilities for reusing the material.

Annette Myre Joergensen (University of Bergen, Norway)

Equifunctionality of coordinated verbs in Norwegian and Spanish

The coordination of functionally equivalent members is a criterion for coordination,
defined by S. Dik in his work Theory of Functional Grammar II (1997, 189). Coordination is also a universal phenomenon, so this should be a universal property. In this paper I will part from the fact that coordinated members are to be functionally equivalent:

a. Min far liker filmer, men ikke bøker
a'. A mi padre le gustan las películas, per los libros no.

b. Folk kommer og går uten stans
b'. Les gents vont et vient sans cesse

This is not always the case when it comes to coordinated verbs. There is a kind of coordination in Scandinavian languages called "pseudocoordination" where the coordinated verbs are not considered to be equivalent. Some of these coordinations can be translated into Spanish:

c. De hadde snakket lenge om det, plutselig tar han og sier...
c'. Habían hablado un rato largo del asunto, de repente él coge y dice...

d. Når vi diskuterer saker, hopper han ofte til og sier noe dumt
d'. Cuando discutimos asuntos, el salta y dice tonterías muchas veces

In this paper I will show how the application of a semantic analysis on the deep structure of these utterances mentioned above, exposes the semantic relation between the coordinated members. My conclusion is, on the one hand that these two coordinated verbs/predicates are not completely equivalent members, in spite of being coordinated, the interesting thing being that they can be translated to Spanish, and on the other that the Spanish language uses both pseudocoordinations and verbal construction to express a complex action, while Norwegian only has pseudocoordination.
Alexander Künzli, (Université de Genève, Suisse)

*Translators' use of reference material: A cross-sectional study*

This study investigates to what extent professional translators and translation students vary in their use of reference material. Three professional translators and three translation students of the Ecole de Traduction et d'Interprétation of Geneva University translated a short text from an English magazine into German (L2 into L1) while thinking-aloud. The experiments took place at the translators' usual working place. They had thus access to all types of reference material normally available to them when translating: (a) printed and electronic dictionaries, (b) e-mail, fax and telephone to contact human informants, and (c) a small corpus of comparable texts, i.e., texts of the same text type that had been written independently in German and thus had the same communicative function as the original text to be translated. The focus of the analysis of the think-aloud protocols was on the identification of the verbalizations that revealed how participants made use of reference material. The written translations were evaluated from a communicative point of view, i.e., with respect to the speculated effect that errors have on the communication between the author of the source text and the potential target-text readers. Data analysis is in progress. The preliminary results show that the study has implications for research and pedagogy. Thus, reliable information on translators' use of reference material is best obtained by on-line measures. Secondly, better understanding of how certain types of translation errors are related to inadequate use of reference material will provide students (and professionals) with more efficient strategies for knowledge acquisition.

Alessandra Lombardi, (Università Cattolica del Sacro Cuore, Italie)

*A Corpus-Based Analysis of Collocational Patterns in Domain-Specific Texts.*

The recent interest in corpus linguistics has been labeled as a “major paradigm shift” (Woodward 1996) in the field of linguistic research and teaching. Overestimating the impact of a corpus-based approach, however, risks stirring up futile controversies. Corpus work should be gradually integrated into language training at university and assessed through single case studies targeted on specific learner-groups within the scope of already existing curricula.

The present paper describes a project developed at the Catholic University of Brescia and specifically designed for students of German (as first or second foreign language) specializing in Tourist Management. In order to translate or write specialized texts students need to learn how to express themselves in an appropriate manner according to text type and subject domain. Knowing how and when to use collocational associations properly is essential for developing communicative competence in a specialized register.
The students joining the project have been trained to single out the collocational patterns recurring in the domain-specific texts selected for the corpus. Exploiting the collocational knowledge gained through the corpus analysis, each student had to create a web site in German promoting a local tourist resort (Lake of Garda, Valtellina, Valcamonica, etc.). The selected corpus includes texts (brochures, leaflets, etc.) written by target language experts promoting tourist resorts in German speaking countries.

The project has three aims:
- to increase the pragmatic competence of the students in this domain-specific register
- to enhance the stylistic quality of the translations and/or descriptive texts promoting local tourist resources
- to gradually set up an electronic data-base organized per functional “segments” in line with current research in the field of tourist marketing.

The paper will provide (a) a discussion of the project (b) the data-analysis (c) a first account of the results and (d) the conclusions I have drawn with respect to the usefulness of looking at collocations in domain-specific texts.

Carmen Perez Tuda (University of Leon, Spain)

A Corpus-driven Contrastive Project English/Spanish. A Qualitative Study of Degree

Due to the recent boom both in contrastive studies and in corpus linguistics, my home university – the Spanish University of León – embarked on a corpus-driven contrastive project in 1994. This long-term project aims to cover the most relevant communicative functions and their expression in both Spanish and English in order to provide empirical data on which to base manuals for future translators. Moreover, the results of this project may well prove a priceless aid in the field of foreign language teaching.

Up to now a team of PhD students in translation, supervised by the doctor Rosa Rabadán, one of the key theoreticians in translation nationwide, has started to do contrastive research into several semantic/pragmatic areas (such as degree, characterization, quantification, modality, tense, voice, aspect and discourse markers). The number of lecturers and PhD students involved in this project is constantly increasing.

The lack of a bilingual corpus English/Spanish has pushed this research group to choose two monolingual ones: the Cobuild/Bank of English and the Spanish CREA. As a PhD member of the above team, I started working on the expression of degree in adjectives and adverbs two years ago. I have carried out research visits to the Cobuild Centre of Advanced Research in English at the University of Birmingham (UK) in order to learn the possibilities of exploitation of this English corpus. We have also contacted the RAE, the owner of the corpus CREA.
The expression of degree is a fascinating field in itself, but still more if it is looked into from a contrastive viewpoint. Spanish people have a reputation for being more emotional and expressive than the British, and this feature is shown in the variety of syntactic options each language has in order to convey different points on the scale of intensity. The gradability of adjectives and adverbs in both English and Spanish will be the topic of this paper. We will deal with the different options each language offers to convey each of the semantic/syntactic aspects of degree (lexematic, iterative, approximative, progressive, comparative, consecutive, pseudo-negative gradability and so on). But we will focus on what really constitutes a linguistic frontier, i.e. the different contexts and frequency of occurrence, extra-linguistic connotations, collocational restrictions and vitality of each of these aspects.

We are in favour of an eclectic approach to contrastive studies: descriptive empirical linguistics and explanatory linguistics are complementary. Corpus-driven data have become the testing ground for our hypotheses, but corpora are only implemented to process the data. The mission of these computerised tools is merely ‘to show’; it is the human analyst’s job ‘to interpret the results’. It results in a fruitful ‘trial and testing’ relationship.

**Noelia Ramon Garcia (University of Leon, Spain)**

**A Corpus-driven Contrastive Project English/Spanish. Characterization in English and Spanish**

The University of León has been engaged for several years now in a wide-reaching project on corpus-driven Contrastive Analysis involving two languages: English and Spanish. Government grants have been awarded to the project as a whole, as well as to individual researchers working on different semantic fields. Several lecturers are involved and a number of doctoral dissertations are currently being written on distinct communicative areas such as characterization, degree, quantification, modality, verbal tense, etc.

Due to the fact that there are no manuals on English/Spanish Contrastive Studies focused on corpus-driven data available at the moment, the future applications of the results are extremely promising. All the previously mentioned research projects can be directly applied to the improvement of Translator Training, and to the development of Foreign Language Teaching (FLT) in Spain.

There are two main corpora that are being used in this project: the COBUILD Bank of English and the Spanish corpus called CREA. Both are corpora of the general language in use. They are well known, their use is widespread, and they provide information that is essential for the development of our project.
This paper is centred on characterization as a distinct semantic field in both languages. English and Spanish use different grammatical resources to describe nouns as well as actions. Although adjectives and adverbs are the two main syntactic resources used with this purpose, there are a series of constructions specific to each language. In general, characterization in English follows a much more detailed approach, at the same time accumulating adjectives and adverbs in positions where this could not be possible in Spanish. We will try to identify the ways more currently used for this type of description in Spanish, so as to determine the procedure translators should adopt in such cases.

Gudrun Rawoens (University of Gent, Belgium)

**Bilingual lexicography. Dutch-Danish dictionary project**

In 1997 the Department of Nordic languages started up a Dutch-Danish dictionary project. The aim is to create a bilingual dictionary of about 40,000 words. The project is an initiative taken by the CLVV who also coordinates other bilingual dictionary projects. The Dutch-Danish dictionary is quite remarkable because it makes use of a very new editor called OMBI which was created for bilingual dictionary making. RBN (Referentiebest and Nederlands - corpus based) forms our material for the source language. In a poster we intend to give an overview of the few existing Dutch-Danish dictionaries and point out the need for this new version, as well as describe our way of working with the OMBI-editor (network-advantages-problems) and the RBN.

Morris Salkoff (France)

**A formal comparative French-English grammar**

I have recently completed a comparative French-English grammar in which the comparisons between French structures and their English equivalents are formulated as translation rules. Each such rule consists of a French schema (of a particular grammatical structure) and its associated translation into an equivalent English schema. The grammar contains all the rules giving the English equivalents under translation of the principal grammatical structures of French: the verb phrase, the noun phrase and the adjuncts (modifiers).

In addition to its intrinsic linguistic interest, this comparative grammar has two important applications. The translation rules it contains can provide a firm foundation for the teaching of the techniques of translation. Furthermore, such a grammar is a necessary preliminary to any program of machine translation, which needs a set of formal rules, like
those given here, for translating into the target language the syntactic structures encountered in the source language.

The importance of this comparative grammar is seen from the new results obtained here, which can have an immediate application to research in MT. These are the translational sub-classes and the passe-partout (all-purpose) translations. The first are sub-classes of French words and their English translations, which are related only as translation equivalents, but are not otherwise distinguished in their respective grammars. The need for such sub-classes emerges only when certain difficulties are encountered in finding suitable English equivalents for French schemata. The second, the passe-partout translations, are approximate translations which are useful in two cases. In the first case, there are several related French schemata whose translations are different for various hard-to-define subclasses of the categories they contain. Such schemata could be separated only on the basis of semantic considerations that would be difficult to formalize. The passe-partout translation provides a ‘cover’ translation that is approximately correct for all the schemata, and does not require establishing hard-to-define semantic sub-classes. In the second case, some one-to-many translations of a French word can be approximated by a single passe-partout translation which suggests approximately the same range of meaning as the original one-to-many translations.

Such results could not have been found from the sole consideration of grammatical phenomena in French grammar or English grammar independently of each other. As the reader will observe from the data collected here, a comparative study is necessary both for pedagogical purposes, and for research in machine translation.

Chikako Shigemori-Bucar (J.W. Goethe-Universität, Germany and Ljubljana University, Slovenia)

Meaning and use of lexical and derived forms in voice---Japanese and Slovene

This paper is a report on research that aims to establish an effective teaching method of Japanese basic verbs to Slovene speaking learners at beginners’ level. The starting point of the research is the fact that there are many transitive/intransitive verb pairs in Japanese which are said to be difficult to learn for second language learners. The main problem to be tackled is how to develop a methodology in which the distribution and frequency in practice would be taken into account to support and/or refine morpho-syntactic contrastive analysis of such language combination (Japanese and Slovene) for which practically no bilingual databank is available.

Starting with Japanese and Slovene basic lexical verbs (cca. 250 vbs. each), their semantic and syntactic characteristics in simple sentences with one, two and three arguments, their derived forms as well as periphrastic means for passive, indirect passive and causative expressions are compared. The analysis shows certain characteristics of the
Japanese and Slovene verbs when contrasted in form and meaning. It is presumed that quite a large part of the Japanese causative construction (especially the derived causative of intransitive vbs) corresponds to the Slovene basic transitive construction, and that a large part of the Japanese basic intransitive construction corresponds to the Slovene derived passive construction without the mention of agent.

In order to support the assumptions, and to find out to what degree they hold true, corpus analyses must be conducted. Since data in Japanese and Slovene is very limited, four different approaches are suggested:
1) error analysis,
2) short essay writing based on visual materials,
3) translations of well-read texts of 3rd language origin,
4) corpus analysis of larger but different computerised texts in each language. The merits and demerits as well as limitations of each approach are discussed.

Elke St John (University of Sheffield, Great Britain)

An Investigation into the Frequency of Pronouns in the German Basic Law and its Translation into English, with the aim of presenting conclusions concerning Translators' Choices and Translation Techniques

Translation can be understood in two ways: as a ‘process’ and as a ‘product’. This paper concentrates on translations as products, comparing a corpus of the German Basic Law (original text) with the corpus of its (published) English translation. It is a case study approach to a subset of English and German structures.
This study concentrates on cross-linguistic comparison, namely the use of pronouns in the German Basic Law and its translation. Traditional linguistic studies do not manage to induce a systematic translation strategy. In this case a solution can be offered by a quantitative analysis of the features of these target structures found in a parallel corpus of translation.

The choice of the German Basic Law is based on five presuppositions:
• The great majority of linguistic features of translation occur in all text types
• Most of the relevant factors in legal translation can be studied and described with linguistic methods
• The choice of legal language proved to be fruitful at least as far as the frequency of pronouns was concerned.
• The translation of texts dealing with specialised subject matters into a target language (here English) being a foreign language to the translator is quite a common task for professional translators. The contrastive analysis of this study may therefore turn out to be of some practical value.
• As specialised texts are in a sense more restricted in their syntax and lexis (compared with fictional texts anyway), they are particularly useful for the purpose of contrastive
discourse analysis and the development of appropriate and efficient translation strategies.

The general question this paper addresses is whether there are differences in the use of pronouns between an original law in German and its translation in English. Since pronouns appear scattered around the length of the text, it is not uncommon for these patterns to change in translation, given that translators do not always have these patterns immediately available to them as they prepare the translation of a text. As a result it is predicted that what was originally a pattern of the source language texts may become a different pattern altogether in the translation. I want to investigate whether there are shifts at all and if there are shifts whether they are just due to stylistic preferences or whether there are also other reasons behind it. It needs to be stressed that this study is only concerned with text type-specific translation strategies. German and English are two genetically and typologically closely related languages and it will be of interest to find out which strategies translators employed and whether concessions were made in any respect in this specific text type. The comparison of word frequencies of pronouns will also give some insights into literal translation.

The approach involving a close comparison of source and target language texts meets the objective of this study, although some insight into the act of translation would be interesting, particularly where translators’ choices seem arbitrary or strange and thus difficult to explain and analyse. The source language text together with the target language text are scrutinised to find out whether regularities and/or preferences in usage can be discovered. Presupposing that a constitutive shift is inevitable in the target texts as a result of the linguistic differences between English and German (for example, word-formation and pronoun agreement) I will compare the translators’ choices with the source language text and describe them in terms of tendencies or strategies. Where the motivation of translators’ choices is discussed, the explanations I give are probable rather than absolute. Although no claim to an ideal translation is made, a number of target language text realisations are criticised in the examination process and alternatives are mentioned.

Lioubov Iouzina-Tarvainen (Helsinki University, Finland)

*Rhythmno-syntactic cliches in Russian and English verse*

The Russian novel in rhymes by Alexander Pushkin *Eugene Onegin* (over 5,500 lines of iambic tetrameters organized into an intricate 14-line stanza) has been translated into English at least twelve times during the last century. Of interest is the fact that only one translation, that by V. Nabokov, was literal, while the rest versions were attempts to render Pushkin in verse, literary.

It seems tempting to check an intuitive guess, based on the idea of C. Rabin "... the more and the longer translations are made from language A to language B, the easier it becomes to translate from A to B ...", that every new rhymed version is closer to the original
The distance between the ST and its TTs seems to be reduced even more after publishing a thorough literal version, which might confirm the idea that literal and literary translations are not mutually exclusive but mutually additive modes. To this end, contrastive linguistic analysis is planned to be carried out at the levels of the whole text, a stanza, a sentence, a line, a phrase, and a single word. The general approach is not evaluative but based on the degree of deviation from the original, the Nabokov version serving as a gloss and taking no other part in the analysis.

The level of comparison suggested for the present symposium is that of a phrase, or formula. As is specified by Marina Tarlinskaja of the USA: "Grammatical patterns containing repeating words used under the same metrical conditions in a verse line, that is, 'formulas', are a structural element in at least two substantially different literary traditions: Russian and English."

Michael Gasparov of Russia stresses the idea that verse 'formulaity' is of interest because it is a meeting point of the two levels of the structure of a poetic text: lexico-stylistic and metro-rhythmic. Such a formula is, first of all, based on a more or less frequent word (lexical factor), such a word occupies a marked position in the line, normally at its end (rhyme factor), and its adjoining words correspond to the meter of the line (rhythm factor).

M. Gasparov singles out ten rhythmo-syntactic cliches, of which the seventh type is most abundant (300 lines in Eugene Onegin), which is further subdivided by him into ten subtypes, the most often used being AbB (governing noun + adjective + governed noun), which M. Tarlinskaja scans as w S / w S w s w / S (e.g. Lesov tainstvennuju sen’), where S stands for a stressed ictus, s for an unstressed ictus, w for an unstressed non-ictus, W for a stressed non-ictus, / signifies a word boundary. Gasparov lists 112 cases of this particular construction and accounts for the predominance of this double-attribute construction in the rhythmo-syntactic repertoire of the Russian I4.

In an attempt to list all the formulas used by A. Pushkin, the Russian original was typeset as a text, with every line scanned as a rhythmic variation of iambic tetrameter in two ways: with and without word boundaries. As a result, the lists of formulas presented by M. Gasparov have been modified. In my report I intend to present the results of a comparison which might reveal the corresponding structures in the translations under investigation.
**Knut Hofland (University of Bergen, Norway)**

**Aligner and browser software for parallel texts developed in the ENPC project**

In the English-Norwegian Parallel Project (ENPC) software has been developed to align parallel texts (original and translation) and later to index and search the aligned texts. The texts are marked up according to the TEI recommendations. The alignment may be manually corrected before indexing.

The aligner makes use of a bilingual dictionary ("anchor list") and is described in Hofland (1996). The aligner has been used for other language pairs like English-Swedish, French-Norwegian, German-Norwegian, French-English, German-English.

The browser allows the user to search for words (or part of words) in one of the languages, optionally together/not together with words in the other language. The search result is given as pairs of sentences. A special version of the browser has been developed to allow the user to search with a Web browser. A demo site containing some public English, German, French and Spanish tests is set up at http://khnt.hit.uib.no/webtce.htm. The browser is described in Ebeling (1998).

**References**

ENPC Web site: http://www.hf.uio.no/iba/prosjekt/

**Philip King (University of Birmingham, Great Britain)**

**Multilingual Parallel Concordancing**

Multiconcord is a multilingual parallel concordancer working on simply marked up text on which the paragraphs have been pre-aligned, categorising, sorting, saving and testing of parallel concordance sentences or paragraphs. Designed originally to work with Danish, English, French, German, Greek and Italian texts, it has recently been officially extended to cope with Spanish, Portuguese, Swedish and Finnish, and has shown itself capable of handling a variety of other languages.

**Elena Paskaleva (Bulgarian Academy of Sciences, Bulgaria)**
**MARK ALISTER - a system for aligning, searching and extracting translation equivalents from parallel texts**

The main function of MARK ALISTER (a PC tool made in DELPHI interface) is the alignment (based on the Gale-Church method) of parallel texts in different input formats. In the preprocessing stage of the alignment an SGML marking of sentences and paragraphs is carried out. The system has various editing options which make it possible to ergonomize the processus of alignment and to neutralize the natural limitation of the algorithm used. It is also possible to search for translation equivalents in pattern-matching style - setting the correspondence between the text element and the segment in the parallel text containing its translation.

The additional extracting module of the system based on Ted Dunning's method of coincidence means it is possible to extract automatically the translation equivalents of selected phrases (in the general case - terms in non-fiction texts). In this way a bilingual dictionary of terms is compiled where the editing facilities and the concordancing support make it easy to establish a custom dictionary of terms.

The results of both procedures can be saved in different formats and used for the aims of different translational investigations.

The main ideology of the system is to compensate for the linguistic insufficiency of a data-driven approach by the linguist-friendly interface.

**Fleur Schut (TRADOS, Belgium)**

**TRADOS translation tools**

TRADOS is a leading specialist in the field of Computer Assisted Translation tools (AKA CAT tools). With 14 years of experience, TRADOS possesses in-depth knowledge of user needs and requirements, the current market situation and linguistic knowledge of specialized languages.

TRADOS has offices in Germany, Belgium, France, Great Britain, Sweden, Switzerland, Ireland, Japan, China and the U.S. The flagship products of TRADOS are the Translator's Workbench and MultiTerm '95 Plus Products.

The Translator's Workbench is a CAT tool based on Translation Memory technology. It automatically stores every translated sentence in a so-called "translation memory" during translation and immediately presents it to the translator whenever he has to translate a similar or identical sentence again. This means that translators never have to translate the same sentence twice, while the consistency - and thus the quality - of the translations will
improve. The Translator's Workbench is the first product on the market which makes use of neural networks for fuzzy matching in translation memories and recognition of terms and collocations.

A terminology component is integrated in the Workbench, so that all translators can be provided with the same relevant terminology for a specific translation project. The component is called MultiTerm '95 Plus! and can be purchased separately as well. MultiTerm '95 is primarily geared towards those translators who want to store and manage their terminological knowledge in flexible, customizable databases. The program accommodates all types of requirements, whether simply to store word equivalents or to create complex dictionary articles with graphics. MultiTerm '95 is fully network compatible. Any number of users can access the database on the network, so all users see the most current information available.

The Translator's Workbench and MultiTerm '95 are both fully integrated in current word processors like Word 6, 7 or 97 and WordPerfect 6.1. Even translation material from previous translation projects - generally containing extremely valuable information - can be re-used by the Translator's Workbench, with the help of the visual text alignment tool WinAlign. WinAlign is a text alignment program that allows the user to align bilingual corpora (source text + translation) for automatic creation of translation memories.


Jean Senellart (Université Paris 7, LADL et Université de Marne-la-Vallée, France) et Cédric Fairon (Université Paris 7, LADL, France et Université Catholique de Louvain, Belgium)

Logiciel INTEX pour Windows 95 & NT

Le logiciel INTEX développé par Max Silberztein est un logiciel de traitement automatique de corpus qui intègre de nombreux travaux de description réalisés au Laboratoire d’automatique documentaire et linguistique (Paris 7). Son fonctionnement est basé sur l’usage de dictionnaires électroniques à large couverture (pour le français, le dictionnaire des formes simples non fléchies compte 120.000 entrées et le dictionnaire des formes fléchies 890.000 entrées), de grammaires locales (Finite State Transducers) et de règles transformationnelles du Lexique-grammaire.

Les outils du système INTEX peuvent être utilisés pour la recherche de patterns lexicaux ou syntaxiques, pour la levée d’ambiguités ou encore le marquage d’unités lexicales simples ou complexes.
Le système INTEX est déjà adapté pour plusieurs langues européennes: anglais, espagnol, français, italien, portugais. Et il est naturellement adaptable à d’autres langues: les utilisateurs peuvent ajouter leur propre alphabet, dictionnaires, et grammaires locales.

Développé à l’origine sur NextStep, Intex fonctionne aujourd’hui dans un environnement Windows. Des versions d’évaluation (entièremenent fonctionnelles pendant une durée déterminée) pourront être données aux personnes intéressées.