USING CORPORA IN CONTRASTIVE AND TRANSLATION STUDIES

(UCCTS)

2010 CONFERENCE

Edge Hill University

27-29 July 2010

Welcome to the UCCTS 2010 conference!

Since the 1980s, the corpus-based approach has revolutionised nearly all branches of language studies. The rapid development of multilingual corpora, including parallel and comparable corpora, is particularly relevant and important to translation and contrastive studies. The corpus-based approach has developed into "a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description, and the practice of translation" (Laviosa 1998), and it has been argued to be "central to the way that Translation Studies as a discipline will remain vital and move forward" (Tymoczko 1998). Likewise, the same is indeed also true of contrastive studies. Multilingual corpus resources such as parallel and comparable corpora have been recognised as a principal reason for the revival of contrastive linguistics that has taken place since the 1990s (Salkie 1999). In addition to significantly advancing translation studies and cross-linguistic contrast, corpus-based contrastive and translation studies have also expanded the field of corpus linguistics.

An international conference like **Using Corpora in Contrastive and Translation Studies (UCCTS)** can clearly benefit translation and contrastive studies as well as corpus linguistics by providing a forum that brings together experts around the world who work in both areas. The 2010 conference is the second in the biennial UCCTS series, launched to provide a forum for exploring the creation and use of corpora in contrastive and translation studies. The UCCTS 2010 conference covers, but is not confined to, the following themes:

- Design and development of comparable and parallel corpora;
- Processing of multilingual corpora;
- Using corpora in translation studies and teaching;
- Using corpora in cross-linguistic contrast;
- Corpus-based comparative research of source native language, translated language and target native language;
- Corpus-based research of interface between contrastive and translation studies;
- Bilingual terminology, lexicology and lexicography.

This year's event is jointly organised by Edge Hill University, the University of Bologna, and Beijing Foreign Studies University. We hope you enjoy your time at the conference!

Dr Richard Xiao
UCCTS Organizing Committee

27-29 July 2010

Programme Committee

Michael Barlow (University of Auckland)

Silvia Bernardini (University of Bologna)

Bart Defrancq (University College Ghent / Ghent University)

Clive Grey (Edge Hill University)

Andrew Hardie (Lancaster University)

Serge Sharoff (Leeds University)

Kefei Wang (Beijing foreign Studies University)

Richard Xiao (Edge Hill University)

Local Organizing Committee

Carol Austin (Conference Centre)

Mike Bradshaw (Department of English & History)

Nicola Kenny (Conference Centre)

Trish Molyneux (Department of English & History)

Richard Xiao (Department of English & History)

Using Corpora in Contrastive and Translation Studies (UCCTS 2010)

Conference Programme

Day 0 (26th July 2010)	
17:30 – 18:30	Registration (Senior Common Room, Main Building)
18:30 – 21:00	Welcome wine reception with buffets and snacks (Senior Common Room, Main Building)
Day 1 (27th July 2010)	
08:30 – 09:10	Registration (Business Foyer)
09:10 – 09:20	Opening by Dr Richard Xiao (Chair of UCCTS Organising Committee) (B001 Lecture Theatre)
09:20 – 09:30	Welcome speech by Dr John Cater (Vice Chancellor of Edge Hill University) (B001 Lecture Theatre)
09:30 – 10:30	Keynote speech A (B001 Lecture Theatre) Chair: Dr Richard Xiao **Title: Parallel corpora and contrastive studies** Professor Hilde Hasselgård (University of Oslo, Norway)

10:30 – 11:00	*Coffee break (Business Foyer)*		
	Session 1-1: Translation Studies Venue: B104 Chair: Professor Mabel Osakwe	***Session 1-2: Contrastive Studies*** Venue: B105 Chair: Professor Marianne Hobæk Haff	***Session 1-3: Corpus & Tool Development*** Venue: B106 Chair: Dr Alex Fang
11:00 – 11:30	Rocío Baños-Piñero (London Metropolitan University) **Understanding the differences between native and translated fictional dialogues in Spanish**	Federico Gaspari (University of Bologna at Forlì) **A corpus-based contrastive study of optional syntactic omission in two varieties of institutional academic English**	Alberto Simões, Sílvia Araújo, Ana Oliveira & Ana Correia (The University of Minho) **Introducing the Per-Fide Project: Parallelizing Portuguese with six different Languages (Español, Russian, Français, Italiano, Deutsch, English)**
11:30 – 12:00	Salma Mansour (Leeds University) **Appraisal emotional adjectives in English/Arabic Translation: A corpus linguistic approach**	Tanja Wissik (University Vienna) **Development and use of comparable specialized corpora of national German varieties**	Stella Tagnin (University of São Paulo) **The COMET Project: Comparable and parallel corpora for the English-Portuguese pair**
12:00 – 12:30	Sanooch Nathalang (National Electronics and Computer Technology	Markéta Malá (Charles University in Prague)	Dechao Li (The Hong Kong Polytechnic University) & Kefei Wang (Beijing

	Centre, Thailand) **Don't use big words with me: An evaluation of English-Thai Statistical-based Machine Translation**	**Copular verbs in English and Czech as seen through a parallel corpus**	Foreign Studies University) **Development and application of bilingual corpora of tourism texts: A new approach**
12:30 – 13:30	*Buffet lunch (Business Foyer)*		
	Session 1-4: Interpreting Studies Venue: B104 Chair: Dr Wallace Chen	***Session 1-5: Lexicon & Terminology*** Venue: B105 Chair: Dr Rafał Górski	***Session 1-6: Teaching & Training*** Venue:B106 Chair: Professor Tengku Sepora Tengku Mahadi
13:30 – 14:00	Rebecca Li (The Hong Kong Polytechnic University) **Corpus-based Interpreting Studies: The state of the art**	Paula Paiva (UNESP/FCLAr - Brazil) **Corpus representativeness in the selection of medical terms be used in translation memory tools**	Guadalupe Ruiz Yepes (University of Hildesheim) **Optimizing the training of translators using corpus search techniques**
14:00 – 14:30	Marlén Izquierdo (University of Cantabria) **Corpus-based identification of English semiperiphrases through contrastive linguistics and translation studies**	Viviana Gaballo (University of Macerata) **A stony ground: A study case in corpus-based terminology**	Miriam Buendía-Castro & Clara Inés López-Rodríguez (University of Granada) **The Web for Corpus and the Web as Corpus in translators' education**

14:30 – 15:00	Marta Kajzer-Wietrzny (Adam Mickiewicz University) **Interpreting universals and interpreter style**	Laura Cantora (Leeds University) **Using Named-Entity recognition systems in the literary domain**	Xiangbing Wang & Lili Ma (National University of Defense Technology) **The influence of specialized parallel corpus on translator competence**
15:00 – 15:30	*Tea break (Business Foyer)*		
	Session 1-7: Translation Studies Venue: B104 Chair: Dr Gernot Hebenstreit		***Session 1-8: Corpus & Tool Development*** Venue: B106 Chair: Dr Wai Lan Tsang
15:30 – 16:00	Hannu Kemppanen, Jukka Mäkisalo & Grigory Gurin (University of Eastern Finland) **The dilemma between corpus statistics and reception of a text: An analysis of foreignizing and domesticating elements in translations**		Alex Chengyu Fang (City University of Hong Kong) & Fenfen Le (Zhongnan University of Economics and Law) **Building a corpus for contrastive studies of British and Chinese Englishes**
16:00 – 16:30	Mabel Osakwe (Delta State University) **Corpora and bilingual translation in Achebe and Soyinka's creative usages**		Rafał Górski (Polish Academy of Sciences) **A project of a BNC-comparable corpus of Polish**
16:30 – 17:00	Emiliana Bonalumi & Diva Camargo (UNESP Brazil) **A study of lexical patterns in a parallel corpus of literary works and their respective translations**		Judith Domingo, Toni Badia (Barcelona Media Innovation Centre) & Carme Colominas (Univesitat Pompeu Fabra) **IAC: A dynamic corpora access interface**

18:00 – 20:00	*Evening meal (Sages Restaurant)*		
Day 2 (28th July 2010)			
09:00 – 10:00	Keynote speech B (B001 Lecture Theatre) Chair: Professor Hilde Hasselgård **Title: A transcultural conceptual framework for corpus-based translation pedagogy** Dr Sara Laviosa (University of Bari "Aldo Moro" / University of Rome "Tor Vergata", Italy)		
	Session 2-1: Translation Studies Venue: B104 Chair: Dr Dechao Li	***Session 2-2: Contrastive Studies*** Venue: B105 Chair: Federico Gaspari	***Session 2-3: Lexicon & Terminology*** Venue: B106 Chair: Prof Paula Paiva
10:00 – 10:30	Sara Castagnoli (University of Bologna / University of Pisa) **Variation and regularities in translation: Insights from multiple translation corpora**	Marie-Aude Lefer (The Catholic University of Louvain) **Genre and domain variation in corpus-based contrastive studies: The case of prefixation in EN and FR**	Stella Tagnin (University of São Paulo) & Elisa Duarte Teixeira (Project COMET) **Let's preserve our identity: Building a Portuguese-English glossary of typical Brazilian cooking**
10:30 – 11:00	Gernot Hebenstreit (University of Graz) **Developments in corpus-based translation studies: A bibliometric approach**	Bart Defrancq (University College Ghent) **Relevance verbs in English, French and Dutch**	Jean-Marie Lessard (Department of Justice Canada) **Legal bilingual and bisystemic dictionary of property in Canada**

11:00 – 11:30	*Coffee break (Business Foyer)*		
	Session 2-4: Translation Studies Venue: B104 Chair: Dr Marie-Aude Lefer	***Session 2-5: Contrastive Studies*** Venue: B105 Chair: Dr Richard Xiao	***Session 2-6: Corpus & Tool Development*** Venue: B106 Chair: Dr Viviana Gaballo
11:30 – 12:00	Wallace Chen (Monterey Institute of International Studies) **Evaluating sight translation: A corpus-based approach**	Rita Calabrese (University of Salerno) **"Living on the edge of two languages": A contrastive analysis of possessive constructions in Smaro Kamboureli's *In the Second Person***	Adriana Mezeg (University of Ljubljana) **Compiling a French-Slovenian parallel corpus**
12:00 – 12:30	Hongwei Huang & Ying Yue (Mechanical Engineering College) **A comparative study of parallel speeches by the Chinese president and American president**	Michaela Martinková (Palacký University) **"I wish you/someone/people would... or mělo by se": A corpus-based study of sentences with I wish and their Czech equivalents**	Silvia Bernardini, Sara Castagnoli (University of Bologna), Adriano Ferraresi (University of Naples "Federico II"/University of Bologna), Federico Gaspari & Eros Zanchetta (University of Bologna) **Introducing Comparapedia: A new resource for Corpus-Based Translation Studies**

12:30 – 13:30	*Buffet lunch (Business Foyer)*		
	Session 2-7: Translation Studies Venue: B104 Chair: Dr Sara Castagnoli	***Session 2-8: Contrastive Studies*** Venue: B105 Chair: Dr Carina Andersson	***Session 2-9: Teaching & Training*** Venue: B106 Chair: Professor Reima Al-Jarf
13:30 – 14:00	Richard Xiao (Edge Hill University) **Can "translation universals" survive in Mandarin: Idioms, word clusters, and reformulation markers in translational Chinese**	Silvia Cacchiani (University of Modena and Reggio Emilia) **Phraseologies in English and Italian historical research articles**	Evaggelia Kalerante (University of Western Macedonia), Simeon Nikolidakis (University of Peloponnese) & Efstathia Georgopoulou (University of Peloponnese) **The teaching of Ancient Greek as a foreign language, for students of immigrant status, at the high school and Lyceum educational levels**
14:00 – 14:30	Viktor Becher (University of Hamburg) **Explicitation and implicitation in translations between English and German**	Renate Reichardt (University of Birmingham) **Does valency theory provide a holistic approach to understanding language?**	Emiliana Bonalumi (FATEC Brazil) **Teaching prepositional verbs through corpora online**
14:30 – 15:00	Miguel Jimenez-Crespo (Rutgers University / The State University of New	Marianne Hobæk Haff (University of Oslo)	Daniel Gallego-Hernández (University of Alicante)

	Jersey) **The future of translation "universals": What can localization tell us about general features of translation?**	**Counterfactual conditionals in focus - A contrastive analysis of French and Norwegian**	**Acquiring instrumental sub-competence by building do-it-yourself corpora for business translation**
15:00 – 15:30	*Tea break (Business Foyer)*		
	Session 2-10: Translation Studies Venue: B104 Chair: Miss Renate Reichardt	**Session 2-11: Contrastive Studies** Venue: B105 Chair: Dr Bart Defrancq	***Session 2-12: Corpus & Tool Development*** Venue: B106 Chair: Dr Wallace Chen
15:30 – 16:00	Elisabet Murtisari (Monash University) **Relevance-based framework for explicitation: A new alternative**	Carina Andersson (Uppsala University) **Epistemic expressions in contrast: The relevance of polysemy vs. grammatical form and epistemic scale in translation of French *sans doute* / *devoir* into Swedish**	Wai Lan Tsang & Yuk Yueng (University of Hong Kong) **The construction of a Mandarin interlanguage corpus**
16:00 – 16:30	Guangrong Dai (Fujian University of Technology) & Richard Xiao (Edge Hill University) **"SL shining through" in translational**	Azizeh Khanchobani Ahranjani (Islamic Azad University, Salmas Branch, Iran) **The first kind of complex noun phrases in Turkish language and**	Tengku Sepora Tengku Mahadi, Helia Vaezian & Mahmoud Akbari (Universiti Sains Malaysia) **Design and development procedure**

	language: A corpus-based study of Chinese translation of English passives	**their equivalents in English**	**of an English-Malay parallel corpus**
16:30 – 17:30	Keynote speech C (B001 Lecture Theatre) Chair: Dr Sara Laviosa **Title: Translation: Some tough questions and some answers** Professor Raf Salki (University of Brighton, UK)		
19:00 – 21:30	*Conference dinner (Sages Restaurant)*		
Day 3 (29th July 2010)			
	Session 3-1: Translation Studies Venue: 104 Chair: Miss Renate Reichardt	***Session 3-2: Translation Studies*** Venue: B105 Chair: Professor Stella Tagnin	
09:00 – 09:30	Reza Moghaddam Kiya (University of Tehran) & Fahimeh Sahraei Nejhad (Payam-e-Nour University) **A study of implementing the lexical and discoursal modifications in translation: Regarding the translation of *The Kite Runner* from English into Persian**	Yuyin He (Beihang University) **A corpus-based study of the translation of Aerospace China White Paper**	
09:30 – 10:00	Reima Al-Jarf (King Saud University) **Interlingual pronoun errors in English-Arabic**	Haidee Kruger & Bertus van Rooy (North-West University) **Features of non-literary translated language: A pilot study**	

	translation	
10:00 – 10:30	Hammouda Salhi (University of Carthage) **Translating ambiguous lexical items using a parallel corpus: A case study of "good" in the EAPCOUNT**	Ting-hui Wen (National Chiayi University) **Measuring mean sentence length in translated and non-translated Chinese texts: A corpus-based study**
10:30 – 11:00	*Coffee break (Business Foyer)*	
11:00 – 12:00	Keynote speech D (B001 Lecture Theatre) Chair: Professor Raf Salki **Title: Translation corpora and the quest for Translation Universals** Professor Anna Mauranen (University of Helsinki, Finland)	
12:00 – 14:00	*Buffet lunch, networking, farewell (Business Foyer)*	

Note: The paper entitled "**Using corpora to define target-language use in translation**" by Rudy Loock (University of Lille 3 / CNRS UMR Savoirs, Textes, Langage 8163) is presented in absentia due to unexpected emergency.

[Keynote A]

Parallel corpora and contrastive studies

Hilde Hasselgård
University of Oslo, Norway

For a long time corpus studies meant monolingual studies. Only since the late 1990s have multilingual and parallel corpora been available. The first machine-readable parallel corpora were the English-Norwegian Parallel Corpus and its sister project the English-Swedish Parallel Corpus. Just like monolingual corpora have led to new insights and new practices in descriptions of individual languages, parallel corpora have opened up new avenues of contrastive studies. Being machine-readable they can give faster access to more material than was previously possible on the basis of non-electronic parallel texts. They also make it easier to see cross-linguistic patterns of correspondence. In my lecture I will touch on the development and use of multilingual corpora with a focus on work done in Scandinavia. I will also present some case studies to show ways of using such corpora for different purposes and within different fields of language description: lexis, grammar and discourse.

[Keynote B]

A transcultural conceptual framework for corpus-based translation pedagogy

Sara Laviosa
University of Bari / University of Rome, Italy

Corpus-based translation pedagogy is a thriving subfield of Applied Translation Studies, as testified by the rising number of publications in recent years ranging from practical guides on how to use corpora in the LSP classroom (e.g. Bowker and Pearson 2002) to scholarly volumes that examine and illustrate the use of corpora for a variety of teaching and learning purposes (e.g. Granger et al. eds. 2003; Zanettin et al. eds. 2003; Gavioli 2005). These works draw on the theoretical and descriptive branches of Translation Studies as well as neighbouring areas of scholarship such as Corpus Linguistics, Information and Communication Technologies, Computational Linguistics, Machine (Assisted) Translation, Contrastive Linguistics, Terminology, Lexicography, and LSP studies.

This paper takes stock of this important development and examines the main methods currently employed in corpus-based translation pedagogy using the three-level model elaborated by Richards and Rodgers (2003) for the analysis of language teaching methods. This investigation brings to light significant differences between translator training, where corpora are well established, and translator education, where they have just started to make inroads. Arguably, one of the reasons for this gap lies in the lack of interdisciplinary theoretical frameworks conceived specifically for translator education.

On the basis of these considerations, I will explore, in the second part of my paper, the principles underlying an envisioned transcultural conceptual framework, within which corpora can play a significant role in equipping students of language and translation with the competences and capacities they need for the future. These principles are: 'holistic cultural translation', as put forward by Maria Tymoczko (2007) in translation theory, and 'symbolic competence', as elaborated by Claire Kramsch (2006, 2009) in the theory of foreign language education.

[Keynote C]

Translation: Some tough questions and some answers

Raf Salki
University of Brighton, UK

Translation enables us to ask interesting but troublesome questions about language which simply do not arise from a monolingual perspective. Translation corpora enable to us to start answering these questions rigorously. Here I will present some of the questions, and indicate some possible answers.

Question 1:
What do speakers of a language do when their language does not have a particular construction? I call this an *expressive gap* in a language.

Question 2:
Why do speakers of a language use some constructions more than others? I call this phenomenon *expressive (dis-)preference*.

Question 3:
Are there regular, high-level generalisations about the way languages express certain types of meaning? I call these *expressive differences*.

These three questions are controversial, slippery and theoretically suspect. Despite this, I will argue that questions like this are the fundamental ones for contrastive linguistics.

[Keynote D]

Translation corpora and the quest for Translation Universals

Anna Mauranen
University of Helsinki, Finland

Translation universals have been a highly contested area, in part no doubt on account of the associations with the term "universal". Whether we wish to talk about translation universals, translation laws or general translational tendencies, corpus data has turned out to be the most useful data type to search for typical features shared across translations. There are a number of issues that we need to address in designing and compiling a translation corpus – the decision of whether we want a comparable corpus, a contrastive corpus, or whether we reach out for a multilingual corpus. This talk discusses these corpus types in view of the kinds of results they can yield, with examples from the Corpus of Translated Finnish (CTF), and the Finnish-English Contrastive Corpus (FECCS).

Understanding the differences between native and translated fictional dialogues in Spanish

Rocío Baños-Piñero
London Metropolitan University

Several scholars (Baker, 1993, and Toury, 1995, among others) have suggested and proven that the selection of linguistic features in translated and non-translated texts is governed by different norms. In the case of Spanish audiovisual texts, researchers have mainly focused on the differences and similarities between native and dubbed texts with regards to their naturalness (Romero-Fresco, 2009) and their pretended spontaneity (Chaume, 2004; Zabalbeascoa, 2008; Baños-Piñero and Chaume, 2009). Research suggests that original audiovisual texts bear more resemblances to spontaneous conversation than dubbed texts in every language level (Baños-Piñero and Chaume, 2009). As these results do not seem to be restricted to the Spanish language only (see Pavesi, 2008), this is a phenomenon worth exploring. The aim of this paper is to understand the differences between native and translated fictional dialogues using different types of audiovisual corpus. For this purpose, the paper will briefly show the results of a previous study based in the use of a comparable corpus.

The aim was to contrast the linguistic features used by both Spanish scriptwriters and audiovisual translators. This comparable corpus consists of 2 episodes of a Spanish sitcom and 5 episodes of a US sitcom dubbed into Spanish. In order to interpret the results and understand the differences and similarities between these two sitcoms, a secondary corpus will be compiled. This will consist of a parallel corpus (compiled by adding to the previous set of texts the source texts of the US sitcom) and a "draft corpus" -consisting of the preproduction scripts of the native sitcom, as written by Spanish scriptwriters and before being interpreted by the relevant actors. In the case of the dubbed texts, the analysis of this specific corpus will enable us to identify standardisation and levelling-out trends (Baker, 1993). The comparison of pre and post production scripts will show the key role played by actors during the shooting of the TV series, when introducing spontaneous-sounding and nonstandard linguistic features.

A Corpus-based Contrastive Study of Optional Syntactic Omission in Two Varieties of Institutional Academic English

Federico Gaspari
University of Bologna at Forlì, Italy

This paper explores optional syntactic omission in acWaC (Bernardini et al., 2010), a corpus that was built automatically using the BootCaT toolkit (Baroni and Bernardini, 2004) from two sets of academic websites: those of a group of British and Irish universities on the one hand (EN-UNI, made up of approximately 5.4 million tokens, to represent the native/original benchmark), and the English-language sections of the websites of a selection of Italian universities (IT-UNI, 4.2 million tokens) on the other. The latter dataset is bound to include translations of Italian source texts as well as instances of L2 writing in English, and is therefore regarded as a body of mediated language.

Based on the classification presented in Biber et al. (1999:661ff), we considered a range of verbs that take a *that*-complement clause in post-predicate position, such as *hope, suggest* and *ensure*, to investigate the patterns of retention vs. omission of the optional *that* complementizer (ibid.:680ff) in a number of constructions, comparing mediated language and native/original writing. To contrast the incidence of optional omission, we focused in particular on constructions introduced by a verb taking a *that*-clause, followed by the included or omitted optional *that* complementizer (the zero complementizer counts as a case of omission), followed by a personal pronoun, followed by another verb within the same sentence.

The paper discusses the extent to which patterns of *that* retention vs. omission vary depending on the specific controlling verbs in first position, considering whether the observed phenomena can be accounted for, at least partially, in terms of the semantic categories to which the verbs belong and on the basis of their degree of formality. Investigating whether the optional syntactic omission of *that* is lexically triggered helps to put in perspective the overall findings of the study, shedding light on the differences between native/original writing and mediated language in the academic institutional setting.

Introducing the Per-Fide Project: Parallelizing Portuguese with six different Languages (Español, Russian, Français, Italiano, Deutsch, English)

Alberto Simões, Sílvia Araújo, Ana Oliveira, Ana Correia
Universidade do Minho, Portugal

This project involves the creation of a balanced multilingual parallel corpus which will comprise both fiction and non-fiction texts. One of our main concerns is to cover a broad range of domains (literary, religious, journalistic, legislative, scientific-technical) and develop a detailed typology of text-types for these domains. Efforts will be made to include originals in the seven project languages, including Portuguese in its different varieties: European Portuguese, Brazilian Portuguese and African Portuguese. The corpus will contain extensive TEI and XCES-compliant headers and markup for document structure. The first part of the paper will focus on corpus design criteria and the main features of the corpus, particularly those that distinguish this corpus from existing parallel corpora. Secondly, we will discuss the challenges of elaborating a typology of text-types for the religious domain and problems associated with the encoding of the texts belonging to this category. Thirdly, we will highlight the contribution of this project, financed by the Portuguese Foundation for Science and Technology, to the NLP community, literary and translation studies, lexicography, contrastive linguistics, and language teaching. To conclude, we will demonstrate how the Per-Fide Corpus can be used in contrastive and translation studies: (i) a study of pronominal causative constructions in a French-Portuguese-Spanish-German contrastive perspective. The formal correspondences for the French pronominal causative construction *se faire + Vinf* in translations from French to Portuguese, Spanish and German will be presented; (ii) a contrastive analysis of the translation equivalents of idiomatic expressions belonging to a specific semantic field of the religious domain.

Appraisal emotional adjectives in English/Arabic translation: A corpus linguistic approach

Salma Mansour
Leeds University, UK

Evaluation is a concept that has many heterogeneous applications in different disciplines. Even within the field of linguistics, scholars describe the evaluative language as a phenomenon that has various labels; *appraisal*, *stance* and *evaluation*. Although a large body of research has been carried out on English appraisal especially in the late twentieth century, it is surprising that to date, analyzing *appraisal* in Arabic language has not been targeted by any linguistic researchers- as I am aware- despite the fact that a rich of Arabic lexical words is available for describing evaluation.

The main purpose of this paper is to find out different patterns of appraisal emotional adjectives as being the core element in appraisal in English as well as Arabic. In addition, parsing these patterns to reveal lexicogrammatical parameters of appraisal propositions in the two languages. Another aim is to announce the discovery of appraisal theory in the Arabic language.

The paper proposes collocational as well as concordancing analysis of emotional adjectives in English and Arabic using the *BNC* and *British News* corpora on one hand and *Al-Hayat* and *Contemporary Arabic* corpora on the other. This computational technique is adopted mainly to identify words that typically co-occur with the lexical item under investigation.

The study shows that some of the Arabic translations of emotional adjectives found in Arabic-English-Arabic dictionaries are misleading as they do not reflect the full information of the word. Furthermore, the analysis reveals some translations that are suitable for most examples, but do not help certain collocations. The examples illustrated in this paper spell out the main differences between English and Arabic appraisal sentences.

Finally, the results of this study suggest some lexical words as substitutes of the missing/misleading translations came out with the corpus analysis. It is hoped that these results will serve as a guide image to help translators in understanding or choosing appraisal emotional adjectives in English and Arabic.

Development and use of comparable specialized corpora of national German varieties

Tanja Wissik
University Vienna, Austria

German as a pluricentric language is an interesting research topic. In particular, the national variation in specialized communication has been brought into the research focus since previous research dealt with national varieties in general language. By pluricentric language is meant, that German is used as an official language in different (parts of) countries like Germany, Austria, Switzerland, Luxemburg, South Tyrol, and Eastern Belgium. That leads to the development of differences in the standard language (cf. Clyne 1995).

This paper is dealing with the compilation and use of comparable corpora to investigate the national varieties of German used in Austria, Germany and Switzerland in the specialized communication in the area of higher education. The comparable corpora are compiled with a special regard to legal and administrative language used in the university system. This paper will present the experience of developing these three comparable corpora and will discuss issues which arose when setting up the corpus, in particular the practical issues of identification of text types and texts for inclusion and size of the sub-corpora. Furthermore, the paper will illustrate the contrastive linguistic-terminological research, which will be carried out with these comparable corpora.

The COMET Project: Comparable and parallel corpora for the English-Portuguese pair

Stella Tagnin
University of São Paulo, Brazil

The COMET Project, developed at the University of São Paulo, Brazil, consists of three branches: a technical comparable corpus, a parallel corpus and a learner corpus. This paper will focus on the first two. The Project began in 1998 with students attending the Specialization in Translation Diploma Course when they were required to build corpus-based glossaries in various technical areas at their discretion for their final papers. For that purpose they had to compile corpora and then extract the relevant terminology. In subsequent years other groups engaged in similar activities. Five of these corpora were made available on the Web, for research, in 2005 as part of CorTec (Technical Corpus). Users can extract wordlists, n-grams and concordances but have no access to the whole texts due to copyright issues. In 2008, some of these corpora were enlarged and others added. They now cover fourteen technical areas, and their extent varies from circa 200.000 to one million words for each corpus in each language. The parallel corpus branch of the Project, CorTrad (Translation Corpus) was developed as a joint project between COMET and Linguateca and launched on the Web in 2009; it consists of three subcorpora: technical-scientific, journalistic and literary. It has a distinctive feature in that, whenever available, it presents various stages of the translations of the same text, that is, the original text, the first draft translation, the revised translation and the published translation. Its functionalities are text-specific and were developed using Linguateca's DISPARA system. The texts are POS-tagged and marked-up, which allows users to investigate specific categories, syntactic patterns and different parts of a text. English texts are also semantically tagged for "color" and "clothing".

Don't use big words with me: An evaluation of English-Thai Statistical-based Machine Translation

Sanooch Nathalang
National Electronics and Computer Technology Center (NECTEC), Thailand

With the availability of many machine translation systems come the question of how effective they really are. The main purpose of this study is to evaluate the English-Thai statistical-based machine translation (SMT) developed by Human Language Technology Laboratory, National Electronics and Computer Technology Center (NECTEC), Thailand, by a human evaluator. We look in particular for potential areas of difficulty that may cause problems to the SMT system. The corpus from which the data for the current study were extracted consists of 200,000 English-Thai aligned sentences (around 1.3 million words) originally taken from bilingual sources that are mainly educational in nature such as dictionaries and phrase books. We consider the English sentences as the source, and the Thai sentences as the target. In the past few years, there were a few attempts to evaluate the translations generated by our SMT system (e.g. Porkeaw, Supnithi, Wutiwiwatchai, 2007), using the well-known BLEU metric. The results yielded were relatively low BLEU scores of 13-15. This immediately calls for an in-depth analysis of the difficulties that challenge the system. In this presentation, we based our linguistic analysis on the linguistic approach proposed by Baker (1992), paying particular attention to establishing equivalences between English and Thai at word level. Our investigation showed that simple words in English that can find their equivalences in Thai do not pose major problems in translation. However, problematic cases tend to occur where an English word corresponds to more than one word in Thai. Explanations to these problems can be drawn from a number of approaches, ranging from language typology, morphology and syntax, to lexicography. The results of the investigation lead us to conclude that the linguistic differences between the source language and the target language still play a significant role in developing and improving the SMT.

Copular verbs in English and Czech as seen through a parallel corpus

Markéta Malá
Charles University in Prague, Czech Republic

Drawing on the parallel bidirectional corpus of Czech and English being developed as a part of a larger multilingual corpus of Czech and (currently) 22 other languages (InterCorp), the paper explores the area of copular verbs in the two languages.

While English makes use of a broad repertoire of copular verbs, which make it possible to express various types and modifications of the basic copular meaning of ascribing a quality, property or value to the subject, in Czech the class of copular verbs is limited to the equivalents of *be* and *become* only. This raises the question of what means Czech employs to convey the 'modified attribution', and on the other hand, what the constructions used in Czech can suggest of the meaning of the respective copular verbs in English.

Our second goal is a more methodological one: we would like to illustrate also some ways in which multilingual corpora can be employed in contrastive research, in making "meanings visible through translation patterns" (Johansson 2007, p. 28), grouping various forms used to express a certain function in one language with the help of the correspondences in the other, and highlighting differences in the prevalent 'patterns of choice' in the expression of a particular function in the two languages.

Development and application of bilingual corpora of tourism texts: A new approach

Dechao Li[1], Kefei Wang[2]
1. The Hong Kong Polytechnic University, Hone Kong
2. Beijing Foreign Studies University, China

Based on a critical review of the existing monolingual and bilingual corpora of tourism texts both home and abroad, this paper introduces the design rationale and the practical consideration for Bilingual Corpora of Tourism Texts (the Corpora), which are being developed in the Hong Kong Polytechnic University. The practical consideration of the Corpora includes the digitization, the tagging, the alignment and the design of the header for the texts to be included in the Corpora. The paper concludes by pointing out the potentials for the teaching and research of tourism translation based on the Corpora. Some preliminary findings of the researches that are based on the Corpora are also reported.

Corpus-based interpreting studies: The state of the art

Rebecca Li
The Hong Kong Polytechnic University, Hong Kong

As the development of corpus-linguistics goes further, the application of corpus in Translation Studies becomes more and more popular. As a methodology, corpus is an effective way, which coincides with the research requirement of Descriptive Translation Studies, to represent, explain or summarize the phenomena, norms and universals in translation. Fruits in the research are flourishing. However, the application of corpus in interpreting studies is inadequate. The paper proposes to describe the features of Corpus-based Interpreting Studies (CIS), common grounds and differences between CTS (Corpus-based Translation Studies) and CIS, as well as the research function of CIS, through examples. In addition, the paper will summarize the general procedures and steps of the establishment of interpreting studies corpus, with the purpose to pave the way for the development of CIS.

The earliest CIS rose in the world at the end of the last century. Most of the earlier corpora are single language corpora, but Translation Studies is always involved source text and target text. Therefore, there are some limitations for the application of corpora in Translation Studies. After 1993, when Mona Baker proposed the application of comparable corpus in the studies into universals in translations, the application of corpora in translation studies developed rapidly. In the field of interpreting studies, it is deemed that Shlesinger proposed the application of corpora at the earliest time. CIS contains the features of CTS, which can lay the foundation for some same studies in interpreting, e.g. frequencies of words, grammatical constructions, discourse patterns, co-occurrences, lexical density, type-token ratios, etc. In addition, interpreting could be then studied through the use of two types of corpora, in effect extending two increasingly important applications of corpus-based translation studies into this area of research. The paper proposes to describe the features of CIS, common grounds and differences between CTS and CIS, as well as the research functions and aims of CIS. In addition, the paper will summarize the general procedures and steps of the establishment of interpreting studies corpus, with the purpose to pave the way for the development of CIS.

The paper consists five parts. The first part is the introduction. The second part discovers and traces the origins of the relevant theoretical framework, which is divided into two directions: linguistic theories and translation theories. The third part conducts an investigation into the connection and differentiation of CTS and CIS. The fourth part sheds lights on the establishment of the corpora. In this part, first and foremost, the paper discusses the categories of the interpreting studies corpora. There're also two categories of the corpora: parallel corpus and comparable corpus. At the same time, this part compares the different functions of the two corpora in interpreting studies. Second, the part collects and selects four cases, which are of the represents of the resent research of CIS, and includes the cases of studies in China and overseas. This part also generalizes the steps, features and functions of corpus establishment. Last but not least, the fifth part is a summary of the whole paper, proposing questions needed to be covered in the future.

Corpus representativeness in the selection of medical terms to be used in translation memory tools

Paula Paiva
UNESP / FCLAr, Brazil

Scientific journals have become one of the most important sources of information in the medical area as they bring the most recent studies and discoveries carried out in different countries. In Brazil, editors from different medical areas have articles published in two, sometimes, three languages, that is, Portuguese, English and Spanish. This way, study results from Brazilian researchers have become known worldwide. One important aspect from these studies is the common use of new terminology which is connected to recent discoveries. For this reason, a study was carried out in order to observe the use of frequent terms found on journals of anesthesiology, cardiology and orthopedics (Paiva, 2006, 2009).

We have analyzed the most frequent terms used in articles of cardiology and cardiac surgery, based on two parallel and four comparable corpora. The parallel corpora were compiled with thirty articles, originally written in Portuguese as well as their translations in English, which were done by professionals and published on bilingual Brazilian medical journals between 2003 and 2006. The four comparable corpora were compiled with articles from the same medical areas and were originally written in Portuguese and English, however, they were published on international journals by native speakers in both languages. For the selection of keywords to analyze the terms connected to them, we used the program WordSmith Tools 3.0 (Scott, 1999) with the aid of its following tools: WordList, KeyWords and Concordance.

We have already collected the most frequent terms from the corpora of cardiology based on their keyness and representativeness, and compiled a glossary to be used by (future) translators of medical articles. The use of this glossary has been tested on a translation memory tool called Wordfast (Champollion, 1999), which has been the most common program used by professional translators in Brazil (Nogueira & Noguera, 2004). Since the glossary proposed in this study is based on terms which were chosen considering their representativeness in the corpus, it has proved to be a very helpful aid for translators who (want to) work with medical articles, more specifically, in the area of cardiology. The theoretical basis followed was that of Corpus-based Translation Studies (Baker, 1993, 1996; Camargo, 2005); Corpus Linguistics (Berber Sardinha, 2004; Tognini-Bonelli, 2001) and Terminology (Barros, 2004). We hope this study may help scholars and translators showing how research and practice are straightly connected to each other.

Optimizing the training of translators using corpus search techniques

Guadalupe Ruiz Yepes
Universität Hildesheim, Germany

The student body at universities has become very international. Twenty years ago, most of the students at German universities were German. Therefore, in translation programmes, when being taught translation from German into Spanish, they were translating into a foreign language. Nowadays, however, the same translation class will include not only students translating into their mother tongue but also those translating into a foreign language or even from one foreign language into another. What is the most significant consequence of this situation? What does this mean for the translation student? And what does it mean for the translation lecturer? For the students, it means that they will have very different problems from each other when translating the same texts, while for the lecturer it means that he or she will have to address these translation problems and help the students to solve them.

Students usually learn very quickly to deal with pragmatic and cultural translation problems, but they very frequently lapse into repeating linguistic translation errors. Therefore, this paper explains how to deal with linguistic translation problems by using corpus search techniques. With this purpose in mind, the paper focuses on how the achievements obtained by using corpus linguistics methods in the teaching of a foreign language may be applied to the training of translators, as translation students, when translating into a foreign language, are often in a very similar position to that of foreign language learners.

Corpus-based identification of English semiperiphrases through contrastive linguistics and translation studies

Marlén Izquierdo
University of Cantabria, Spain

In the early 80s Lorenzo argued that some phenomena characteristic of a language can only seen by comparing it with (an)other(s) language(s). Corpus-based research on language use has proved this statement right, not only through contrastive linguistics but also through translation. In fact, this paper provides empirical evidence which confirms Lorenzo's thesis by suggesting the existence of an English paradigm of semiperiphrases. Data for this grammatical resource have been driven from a comparable-corpus-based, contrastive analysis of English and Spanish gerund constructions. A complementary descriptive translation study revealed explanatory data for verb patterns of English which have been long studied as individual pieces of language, even though a functional approach finds such an explanation unsatisfactory. These patterns comprise 'stop + v-ing', or 'start + v-ing' amongst others.

The study is framed with a functionalist view of language and the methodology adopted is corpus linguistics, which urges a context-based description, juxtaposition and contrast of the phenomena under analysis. The contrastive analysis unveils the similar functionality of corresponding verb patterns, whereas the insights gained from observing real translations strengthen my hypothesis that these verb patterns are not instances of verb complementation but complex verb phrases which function as a unified whole.

Overall, this piece of research is carried out at the interface between contrastive and translation studies.

A stony ground: A study case in corpus-based terminology

Viviana Gaballo
University of Macerata, Italy

This study originated from the real-world need to provide a lexicographic reference work for the specialized field of stone processing. Very little is available on this specific niche of the lexicon. This contribution will offer lexicographers valuable information as to the identification and designation of materials, activities, and processes related to the quarrying and processing of stones. The study aims at exploring the concepts underlying the mentioned categories, and uses a methodology which merges the input of new technologies with a primarily contextual and functional view of meaning (Tognini-Bonelli, 2001). The research was conducted on the data collected to build a pair of comparable corpora, each containing a variety of texts – from brochures to technical specifications – in one of the source languages investigated: English and Italian. To advance the inquiry, a number of term candidates were identified – based on the frequency and keyword lists generated from the corpora – and analysed in their contexts of use to eventually formulate hypotheses of equivalence in both languages. By studying and contrasting how the selected terms are actually used in the two languages, the cultural norms and linguistic behaviour in the different engineering cultures can be better understood. This work is the result of the growing convergence of different approaches to meaning, all harnessing corpus evidence.

The Web for Corpus and the Web as Corpus in translators' education

Miriam Buendía-Castro, Clara Inés López-Rodríguez
University of Granada, Spain

The Internet has brought with it a new way of organizing and obtaining information. Due to the vast amount of information offered, it constitutes *a fabulous linguists' playground* (Kilgarriff & Grefenstette 2003: 333), from which Translation can also benefit. Since 1997, Corpus Use and Learning to Translate (CULT) has been a fruitful area of research (Beeby, Rodríguez & Sánchez 2009; Bowker 1998, 2000; López 2002, López & Tercedor 2008; Zanettin 1998; Zanettin, Bernardini & Stewart 2003). Both corpora and the Internet have been included amongst the tools of translators, together with lexicographic and terminographic resources, translation memories, etc. Corpora are rich information sources that can provide the translator with both linguistic and conceptual knowledge that is not found in dictionaries. The question that arises within this context is whether the web could be considered as a corpus. Following the distribution made by De Schryver (2002), there are two corpus approaches to the web:

(i) Web for Corpus (WfC): the web is used as source of texts in digital format, for the later implementation of offline corpus, supported by authors such as Sinclair (2005). Despite recognizing the great utility of the Internet for any linguist, Sinclair highlights the fact that the WWW is not a corpus because it has not been designed from a linguistic perspective. In the field of Translation, the notion of DIY (do-it-yourself) corpus (Zanettin 2002: 242) has been used to describe the collection of Internet documents compiled *ad hoc* as a response to a specific text to be translated.

(ii) Web as Corpus (WaC): the web is directly used as a proper corpus (Kilgariff & Grefenstette 2003; Fletcher 2007; Baroni, Marco & Bernardini 2006).

In this study, we compare and evaluate these two approaches in the context of a scientific and technical translation course at university level. We asked a group of students in the BA Translation and Interpreting Degree Program at the University of Granada to carry out a technical specialized translation assignment. Half of the group was requested to do it using the WaC and the rest using conventional methods of WfC. The results obtained show that these two methods should be complementary and students should decide upon their particular needs, more specifically, the translation assignment, novelty of the translation, directionality and specificity of the translation, time allotted, and the level of analysis required.

Interpreting universals and interpreter style

Marta Kajzer-Wietrzny
Adam Mickiewicz University, Poland

Corpus-based translation research has already proved that translations demonstrate a tendency to simplification, explicitation, normalization levelling out/ convergence. These features have been examined on different language pairs, with application of different corpus methods and are therefore, by many, regarded universal.

Although interpreting is also considered a mode of translation, universals in this mode have not been so extensively researched. This paper presents the study design and a corpus-based methodology for an ongoing study, the purpose of which is to establish, whether translations and interpretations of the same source texts into the same target language share potentially universal features or whether it is possible to discern distinct features of translationese and interpretese. Moreover, to examine "universality" of these features the analysis will involve translations and interpretations into one target language (English) from four source languages (French, German, Spanish and Dutch). The analysis is also set to determine, whether potentially universal features are equally reflected in the performance of individual simultaneous interpreters working in the same environment or, whether different realization of these features would rather lead to identification of a particular interpreting style of a simultaneous interpreter.

Using Named-Entity recognition systems in the literary domain

Laura Cantora
Leeds University, UK

Proper names are central to Information Extraction (IE) in general while their interrelations across languages are a major focus of Cross-Language Information Retrieval (CLIR). The ultimate goal of these technologies is to automate the finding of documents most likely to match a specific information request, to extract from them the pertinent facts, and to present the results in a structured form in the requester's language, irrespective of the language(s) of the documents. Central to both IE and CLIR is the automatic recognition and classification of all proper names in a text, a task known as Named-Entity (NE) recognition. From the mid 90s onwards, NE recognition systems have been developed that achieve up to 99% accuracy in the results obtained within different text types within the informative genre. This paper reports on the implications of applying these techniques to the literary domain.

A NE recogniser was used on a trilingual parallel corpus with a total size of 650,000 words comprising British modern novels and their translations into Spanish and Italian. The aim was to automate the extraction of all the names in the SL texts with a view to identifying the translations of those names at a later stage, through the use of a multilingual concordancer.

The retrieval effectiveness of the NE recognition software was not as high as anticipated, with around 50% of all the NEs successfully identified, therefore, additional processing was necessary to extract and classify all the proper names in the texts.

Precision and recall, as well as details of true and false positives, are analysed in this paper with a view to determining the reasons behind such lower accuracy for this text type.

In the end, the decision was taken to maximise recall and let precision drop, but was that the correct way to approach this problem or does it call for the need to develop tools to be used specifically with literary translated texts?

The influence of specialized parallel corpus on translator competence: Taking military translation as an example

Xiangbing Wang, Lili Ma
National University of Defense Technology, China

The influence of corpus on translator competence has become a hot issue in the corpus-based translational studies. This paper, starting from the empirical study of a mini DIY parallel corpus (400,000 words) of military texts, makes a primary probe into the influence of specialized parallel corpus on translator competence and reaches the conclusion that specialized parallel corpus can enhance translator's subject-field understanding significantly and promote translators' competencies on foreign language as well as native language in a tangible way.

A special experiment of corpus-based military translation training was carried out to explore the relationship between a specialized parallel corpus and translator competence. The parameters set for measuring translator competence in the experiment are: translation speed, foreign language competence, native language competence and professional knowledge. The subjects of the experiment, third-year English majors in a military university in China, are divided into two groups. Some texts related to the texts in the corpus in terms of topic are provided for each group to be used as the translation training materials. In the process of translating, one group is provided with the specialized parallel corpus, the other group is the otherwise. Their translations are analyzed in detail from the perspectives of the 4 set parameters. Some enlightening findings are achieved from the analyses. In the end, the software of Statistical Package for Social Sciences (SPSS) is used to analyze the translation mistakes the students have made, which shows that a specialized parallel corpus can reduce translation mistakes significantly.

The dilemma between corpus statistics and reception of a text: An analysis of foreignizing and domesticating elements in translations

Hannu Kemppanen, Jukka Mäkisalo, Grigory Gurin
University of Eastern Finland, Finland

This paper presents results of a study in which we tested a corpus-based method for operationalizing two concepts used in translation studies – the concepts of *foreignization* and *domestication*. The dichotomic categorization of translation strategies introduced by Lawrence Venuti (1995) has been criticized for its fuzziness (Tymoczko 2000, Boyden 2006). This study strives to recognize the statistical features which could be considered as representations of *the foreign* or *the domestic*.

The aim of this study is threefold: 1) to conduct a corpus-based analysis in which translated texts are examined in relation to their foreignizing/domesticating features, 2) to carry out a test where the degree of foreignization/domestication is evaluated by subjects – translation trainers and students, 3) to find a possible correlation between the statistical features of the texts and the results of the evaluation test.

As a starting point of the study is the notion of *keywords* in the sense it is used by Mike Scott (1998): keywords are words "whose frequency is unusually high in comparison with some norm". It is hypothesized that lexical and lexico-syntactic features representing the *keyness* of words (or the lack of it) could be categorized as markers of foreignization or domestication. The material comprises Russian–Finnish translated and non-translated texts representing non-fiction literature on political history as well as a target language newspaper corpus functioning as a word frequency norm of average language use.

The results of the study show that, on the whole, the statistical features of the translated texts do not correlate with the results of the evaluation test. However, several statistical features – such as the number of keywords and the maximum *keyness value* in each keyword list – are in line with each other. The outcome of the study indicates that there is a need for further research into operationalizing the concepts of foreignization and domestication.

Building a corpus for contrastive studies of British and Chinese Englishes

Alex Chengyu Fang[1], Fenfen Le[2]
1. City University of Hong Kong, Hong Kong
2. Zhongnan University of Economics and Law, China

This paper describes a corpus-based research initiative that aims at contrastive studies between Chinese English and British English. For this purpose, a corpus has been constructed, which consists of two comparable sub-corpora of one million word tokens each, including the Corpus of Chinese Media English (CCME) and the Corpus of British Media English (CBME).

This paper will present the motivation of the project, the overall design of the corpus, including the composition, the sampling and the annotation of the resource. As the paper will demonstrate, the two sub-corpora have been built according to a comparable design for text composition. Within the general domain of public media, three broad categories were identified, namely, newspapers, magazines and Internet news. Special care was taken to ensure that the three categories are directly comparable according to their internal subdivisions of texts into news report, business, social life, culture and arts and editorials. This paper will then present a detailed description of the two corpora in terms of the sources of texts and the sampling of the actual texts.

The paper will also describe the grammatical and syntactic annotations applied to the corpus, which include part-of-speech tagging and syntactic parsing. In particular, it will discuss the grammatical and syntactic characteristics of the two annotation schemes that are expected to be particularly useful for contrastive studies. The paper will finally present a summary account of the two component corpora based on the grammatical and syntactic annotations. Intended areas of investigation will be discussed that will aim at an empirical assessment of the hypothetical notion of "the nativisation of Chinese English" regarding its role and use as an emergent variety of world Englishes.

Corpora and bilingual translation in Achebe and Soyinka's creative usages

Mabel Osakwe
Delta State University, Nigeria

This paper reports research outcomes on the translation of the African experience (linguistic and socio-cultural) by two foremost African creative writers: Wole Soyinka and Chinua Achebe. The corpus data analyzed are texts from their creative make-believe, yet real worlds.

Being co-ordinate bilinguals in English and a major Nigerian language, each of these users of English, is a locus of contact; a contact which automatically generates translation into the target language medium of expression of the corpus data. Whereas a study of the translation of a novel such as *Things Fall Apart* (translated into many world languages) may wish to examine formal inter language translation processes, the focus here is on the informal 'idiolectal' usages which throw up idiolectal, diatopic and diatypic linguistic categories.

Linguistic categories cutting across syntax, lexis, phoric references and rhetorics are up for textual analysis. The findings show Achebe's texts exemplifying 'real world' texts, especially in the varieties of language used. Soyinka's samples are restrictive being largely idiolectal expression of an upper zone cline of bilingual. The neologisms, broad vocabulary spectrum and their manner of freedom of occurrence and co-occurrence also provide further corpus data for Research into literary translation study.

A project of a BNC-comparable corpus of Polish

Rafał Górski
Polish Academy of Sciences, Poland

This presentation introduces a project of an English – Polish comparable corpus – or strictly speaking – a Polish corpus directly comparable to BNC. It shall overcome some well known shortcomings of a parallel corpus (small size, influence of the source language on the target language, lack of texts translated from Polish to English, shortage of certain text types). Such a corpus may be useful for English – Polish comparative and translation studies, not instead of but rather complementary to a parallel corpus.

In the talk we shall discuss the degree of comparability at various levels:

a) size – a Polish translation of an English text is always shorter than its source in a certain proportion. Therefore the length of the Polish corpus will be proportionally smaller than the length of BNC.

b) design – each Polish text will be classified according to the classification used for BNC. Because the classification is threefold (genre, channel and topic), every single text should match an English counterpart, according to the three criteria together. It may be difficult to satisfy this requirement, but the genre is a much more important feature than the topic, while those two are more important than the channel. Thus, more care will be given to assure that the corpus is similar in terms of genres than in terms of topic and, even more so, in terms of channel. There are also some features which are relevant for the characteristics of the corpus, but are not overtly stated, e.g. the literary quality of a novel. Regardless of how important these features may be, it is impossible to consider them mainly because they are based on intuitive grounds.

c) tagset – tailoring a tagset dedicated for the corpus is in practice impossible. In fact, the comparability on the level of the tagset is an important requirement as well.
Two other features of an ideal comparable corpus are: the comparability of annotation and a search engine common for the two corpora.

The Polish counterpart of the BNC will be drawn from the resources of the National Corpus of Polish, which is still being compiled. Thus, it shall relay on the structural and linguistic annotation of the latter. We shall reclassify all texts according to the guidelines of classification of BNC.

A study of lexical patterns in a parallel corpus of literary works and their respective translations

Emiliana Bonalumi, Diva Camargo
UNESP, Brazil

This investigation, part of my PhD research, involves a study of lexical patterns in a corpus of literary texts composed by a subcorpus of translated texts of the literary works of Clarice Lispector, translated into English by Giovanni Pontiero (*Near to the Wild Heart*, *Family Ties* and *The Hour of the Star*) and Alexis Levitin (*Where You Were at Night*), and a subcorpus of original texts of the literary works of Clarice Lispector written in Brazilian Portuguese (*Perto do Coração Selvagem*, *Laços de Família*, *A Hora da Estrela* and *Onde Estivestes de Noite*).

This study has as one of the main purposes to observe similarities and differences in the use of lexical patterns (i.e. *at the same time => ao mesmo tempo*; *in fact => na verdade*; *in the middle of the => no meio da*; *at that moment => naquele momento*) in the target texts and the source texts. The research draws on corpus-based translation studies (Baker 1993, 1995, 1996, 2004), as well as studies of lexical patterns (Hunston & Francis 2000). Another objective of the investigation is to examine the extent to which the addition, omission and / or variation of lexical phrases are a feature of translation or a feature of Lispector's style.

In this paper, after an introduction of the methodology adopted, I present my results.

IAC: A Dynamic Corpora Access Interface

Judith Domingo[1], Toni Badia[1], Carme Colominas[2]
1. Barcelona Media Innovation Centre, Spain
2. Univesitat Pompeu Fabra, Spain

Corpora in translation studies are essential not only for research but for training as well. Although, as Colominas & Badia (2008) pointed out, the real use of corpora in translation studies faces practical limitations: interfaces for accessing corpora are often not user-friendly enough to satisfy the real needs of translation students and researchers. Moreover, each interface is built specifically for a corpus; this implies not only great effort but also the fact that interfaces differ from each other in the layout and type of searches that they allow for.

Conscious of these limitations, we have developed IAC (Corpora Access Interface), a non-dependant corpus interface for monolingual and parallel corpora that allows users, without programming grounding, to create searching interfaces between a given corpus and the underlying search tools1. To create a new interface in IAC only two tasks have to be accomplished: first, the corpus has to be formatted according to IAC requirements (tabular format for attributes at word level and xml format for attributes affecting groups of words and metadata) and second, the searching interface has to be designed by means of an intuitive graphical tool (included in IAC) according to the corpus type and the linguistic annotation added. IAC includes also user-controlled access that allows the user to distinguish between private and public corpora. Once the corpus is uploaded and the interface is created, IAC indexes the corpus and a user-friendly searching interface is automatically created allowing for 3 types of searches: simple, expert and frequency based.

As a conclusion, IAC is an extremely flexible and powerful tool that goes beyond current corpora interfaces limitations and provides the creation of user-friendly interfaces for different kinds of corpora.

Variation and regularities in translation: Insights from multiple translation corpora

Sara Castagnoli
University of Bologna / University of Pisa, Italy

In the last decades the search for *laws of translational behaviour*, or alleged *translation universals*, has been at the heart of corpus-based Translation Studies, leaving the question of variation in translation confined to a few studies comparing the style of literary translators. This paper sets out to show that new insights into presumed common features of translated texts such as explicitation and interference can be obtained precisely by analysing the interplay between variation and regularities in translators' behaviour; and that, for this purpose, comparing the performance of several translators translating the same source text (ST) appears to be a more reliable method than relying on traditional corpus resources (i.e. parallel and monolingual comparable corpora). *Multiple translation corpora* (MTCs) are "special" parallel corpora in which several translations into the same target language (TL) are available for each ST; while traditional parallel corpora conceal the variation that would inevitably emerge if translations produced by different translators were available, MTCs make it possible to observe regularities and variations in the way different translators cope with the same ST.

The paper reports on a study of explicitation and interference with respect to connective usage based on a MTC of student translations (English/French > Italian), which showed that the MTC methodology proves revealing in two main respects. First, it will be argued that when most translators translating the same ST are observed to depart from it in similar ways (e.g. by adding connectives in target texts (TTs)), a reason for this may lie in their attempt to approach target language (TL) preferred patterns. Several similar TT renditions may thus indicate shifts towards TL standards, whereas variations from observed regularities may be assumed to constitute less "correct" translations. In addition, the ratio of translators opting for specific shifts can be indicative of the nature of such shifts – i.e. obligatory, optional, (possibly) translation-inherent – based on the comparison with other translation solutions. Second, and somewhat conversely, lack of variation in the translation of specific ST items may not only point to the existence of preferred translations, but it may also betray ST shining-through (i.e. interference). Examples taken from the MTC and illustrating the above methodological suggestions will be provided, along with more detailed information about the corpus, the specific types of analyses it enables, and the statistical measures (standard deviation and binomial test) used to single out "deviant" translations.

Genre and domain variation in corpus-based contrastive studies: The case of prefixation in English and French

Marie-Aude Lefer
The Catholic University of Louvain, Belgium

Genre variation has long been neglected in cross-linguistic analyses. However, more attention has recently been paid to issues related to register, genre and domain in corpus-based contrastive studies (e.g. Teich 2003, Fløttum *et al*. 2006, Hansen-Schirra *et al*. 2007). In my presentation I will assess the role played by genre and domain variation in contrastive studies by examining the use of word-formation devices, and more specifically prefixes, in English and French writing. The study investigates c. 100 prefixes in each language in terms of their realised productivity (i.e. the number of different lemmas formed with them; see Baayen 2008). It is based on the following comparable corpus, made up of three bilingual components of c. 2 million words each:

- Fiction texts (novel excerpts) from the *British National Corpus* and *Frantext*;
- Newspaper leading articles from the *Multilingual Corpus of Editorials* (MULT-ED);
- Research articles in medicine, economics and linguistics from the KIAP corpus (Fløttum *et al*. 2006).

The approach is thus both contrastive (English and French) and doubly comparative (genres and domains). The main objective is to (in)validate general trends brought to light by using a 'mixed-bag' corpus (i.e. the three components described above taken together). To this end, in-depth analyses of genre- and domain-specific features have been carried out. The results show that some semantic categories of prefixes (e.g. negation in English and reiteration in French) are distinctively more productive in one language than in the other in the three genres and domains investigated. This cross-genre and cross-domain similarity suggests that the differences are due to true contrasts between the linguistic systems investigated. On the other hand, the productivity of other categories tends to be genre- or domain-dependent. For example, French number prefixes (such as *uni–* and *bi–*) are more productive than their English counterparts in novels and editorials but not in research articles. The study shows that the variationist approach is a good method for teasing out genuine cross-linguistic contrasts and genre- or domain-specific patterns. It also highlights the critical importance of corpus make-up in contrastive studies.

Let's preserve our identity: Building a Portuguese-English glossary of typical Brazilian cooking ingredients

Stella Tagnin[1], Elisa Duarte Teixeira[2]
1. University of São Paulo, Brazil
2. Project COMET, Brazil

One of Brazil's main sources of income is tourism and tourists must eat. However, more often than not, they have difficulty understanding our menus because of the pitiful translations of some menus. On the other hand, Brazilian cooking is arousing the interest of more and more international chefs who open branches of their European restaurants in our country. They enjoy experimenting with our typical ingredients and creating Brazilianized versions of traditional dishes. Due to this boom, a wealth of cooking schools and colleges has arisen in the last few years, creating an ever-growing demand for bilingual reference material in Gastronomy. To partly meet this demand, we have created a project to build a glossary in the Portuguese-English direction, mainly addressing the most recurrent ingredients in Brazilian Cooking. The glossary is especially aimed at translators, restaurant owners, chefs and Gastronomy students. The corpus-driven *nominata* is being extracted from a comparable corpus consisting of Brazilian recipes. The project will initially address the first one-hundred KeyWords of the Brazilian corpus, along with their collocates. The next step will attempt to identify equivalents in the comparable English corpus. A parallel corpus, consisting of digitalized bilingual English-Portuguese Brazilian Cooking cookbooks has also been compiled to aid in equivalent identification. Each entry will consist of 3 parts: a) an equivalent, provided there exists one, with a usage example; b) a short explanation, which may be inserted in a text as an appositive, for instance; and c) a longer text giving culinary details of the ingredient, the dishes it is most used in, how it is usually prepared etc. As the glossary is to be made available in digital format, it will also include images and links to further information. The paper will provide details on the construction of the corpora and a variety of entry examples.

Developments in corpus-based translation studies: A bibliometric approach

Gernot Hebenstreit
University of Graz, Austria

Over the past years a constantly rising number of publications on translation and interpreting research that are in some way or another based on corpora indicate the growing importance of corpora for translation studies. Some authors even speak of a new paradigm (Corpas Pastor 2008). This paper aims at giving an overview of the development of corpus based methodologies in research on translation and interpreting. Analyzing bibliographic data on relevant publications this study tries to give answers to the following questions: Which branches of translation/interpreting studies have most interest in "introducing" (Olohan 2004) or "incorporating" (Anderman/Rogers 2008) corpora? Of what kind are the research questions that are to be answered based on information provided by corpora? Is there a relation between the adherence to a school of thought and the interest in corpora? What's the relation of quantitative vs. qualitative research? Which methodological questions are being raised? Are there preferred design patterns for building corpora? What role does tagging play? Of what kind are the desiderata voiced in terms of corpus technology?

Relevance verbs in English, French and Dutch

Bart Defrancq
University College Ghent, Belgium

This study deals with a particular group of verbs called 'verbs of relevance' or 'verbs of indifference' in the literature (cf. Karttunen 1977; Lahiri 2002 and Hoeksema 1994; Leuschner 2005 & 2006 respectively). There is no consensus on how the category should be defined, but all authors seem to tacitly admit that verbs of relevance can govern embedded interrogatives. In English, the category covers verbs such as *bother, care, count, matter, mind,…*; in French *compter, s'en foutre, s'en ficher, importer,…*; in Dutch *ertoe doen, schelen, tellen, uitmaken,…*

The purpose of this study is to describe the interplay between pragmatics, semantics and syntax in the way these verbs are used in corpora of English, French and Dutch. It will appear that this interplay is grounded in pragmatic constraints arising from the principle of relevance (Sperber & Wilson 1986).

The basic idea is that, as relevance is presupposed, verbs of relevance are more likely to be used with negative than with positive polarity (except of course if their meaning is already inherently negative, as is the case of some French items). Used with positive polarity, they tend to occur in sentence forms that present them as strongly presupposed, such as clefts and pseudo-clefts. Used with negative polarity, they are more likely to occur in the focal area of the sentence, leading to the use of canonical sentence form and extraposition.

These tendencies can be observed in all three languages, but each language has specific additional features: in English, for instance, corpus data clearly show that the use of a preposition in front of the embedded interrogative correlates with the polarity of the sentence. In Dutch, on the other hand, the most striking aspect of relevance verbs is their high degree of grammaticalization: one of the items, for instance, standardly combines with a fixed modal verb (*kunnen schelen* : 'can differ'). These parallels and contrasts between the items in the three languages involved will be systematically explored in this study.

Legal Bilingual and Bisystemic Dictionary of Property in Canada

Jean-Marie Lessard
Department of Justice, Canada

In Canada, federal law is expressed with equal authority in both official languages, French and English. Adding to this already complex situation of communication, two systems of legal thought and legal rules exist in private law: Civil Law and Common Law. The *Legal Dictionary of Property in Canada* is a bilingual and bisystemic encoding and decoding tool to interpret legislative and judicial texts. In particular, this dictionary aims to facilitate the interpretation of the vocabulary of Property Law within the context of legal and linguistic dualism. Since the coverage of this targeted field of law meets inherent consistency criteria, we can use the term *legal ontology* to describe the layout of the entries.

General structure of the work

It is the first of a four-volume two-tome bilingual and bisystemic legal publication entirely dedicated to Property Law. The current word list slightly surpasses 300 entries. The English and French articles are displayed side by side, with the French component always located on the right hand side of the page in both tomes. Each version has more than 700 pages. The headword pairing, in each language, is provided as a linguistic and legal equivalent. The dictionary's definitions are very short, but for each one, the legal system to which it belongs is indicated. Our goal is to provide readers with a consistent starting point that will facilitate the interpretation of the excerpts from case law included in the article. This exact way of using a corpus is a first in Canada.

Corpus

There are slightly more than 2,500 aligned bilingual documents within the corpus that were selected from more than 18,000 judgments from an overall text database relating to Property Law. The selected documents were taken primarily from the Supreme Court of Canada, the Federal Court of Appeal and provincial courts of appeal. These judgments were rendered over a period of approximately 30 years.

Evaluating sight translation: A corpus-based approach

Wallace Chen
Monterey Institute of International Studies, USA

This paper outlines a corpus-based approach to evaluating sight translation, specifically the evaluation of learners' sight translation assignments involving specialized texts. The data used in the research include a Reference Corpus (RC) and a Learners' Corpus (LC).

As an experiment, the RC is composed of over 25,000 words of naturally-occurring, non-translated and topic-specific English texts in the areas of rainforest conservation, laptop computing and telecommunications. Being authentic and professionally written, the texts contained in the RC can serve as a benchmark against which translator trainers can compare learners' sight translation outputs in various linguistic levels, including collocation, terminology, idiomatic expression, verification of intuition, translation equivalent, target language patterns and new expressions. The LC, on the other hand, contains over sixty transcripts of learners' sight translation outputs from two Chinese texts used in two examinations.

By consulting the LC, trainers are able to systematically identify learners' error patterns, individual styles and issues, length of delivery, and the possible connections between lengthy delivery and language use. It is further suggested in this paper that the corpus-based approach to translation evaluation offers an empirical tool to complement traditional approaches that are based on intuition, personal experience, subjective judgment and restricted knowledge of subject matters. It is also argued that, by incorporating these electronic corpora in the learning process, students will be better equipped with the necessary tools to become independent learners and will have greater awareness of language use in specialized translations.

'Living on the edge of two languages': A contrastive analysis of possessive constructions in Smaro Kamboureli's *In the Second Person*

Rita Calabrese
University of Salerno, Italy

Smaro Kamboureli's poetic diary records/reports on the reconstruction of a woman's identity deconstructed by living as a Greek immigrant in the 1980s Canadian society. The inner dualism implicit in her bilingualism causes the splitting of the self as mentioned in the book title and makes her live «on the edge of two languages, on the edge of two selves named and constructed by language» (p. 34).

In this paper I analyse the possessive constructions occurring in the text as a structured category unified under cognitive principles (Langacker 2000; Taylor 2000; Fónagy 2004). Following the procedure adopted in a recent study on the semantic relations encoded by N N and N Prep N instances from a parallel corpus of English and Romance languages (Girju 2008), I carried out a similar study by matching corpus-based evidence and the linguistic diagnostics (cross-linguistic syntactic and semantic mappings) adopted in previous research.

In order to perform empirical investigations of the semantics of possessive constructions encoded by nominal phrases (namely N Prep N, N's N) and compounds (NN) in English, and to test the interpretation of such instances in Italian, I collected the data by digitalizing a bilingual edition of the diary published in 2007. The English version was syntactically parsed using VISL applications/linguistic tools which can provide both syntactic and semantic information on a given constituent structure. Then each N N and N P N instance was manually mapped to the corresponding translations to verify the corpus distribution of the semantic relations per each syntactic construction as well as the role of English and Italian prepositions in the semantic interpretation of possessive constructions.

Compiling a French-Slovenian parallel corpus

Adriana Mezeg
University of Ljubljana, Slovenia

Since the development of the first corpora and general awareness of their advantages, they became indispensable in virtually all the areas dealing with the study of language: grammar, lexicology, lexicography, translation studies, pedagogical didactics, etc. Through large national projects, usually financed by public and private institutions and carried out by experts in linguistics and natural language processing, many countries, at least European, managed to develop large reference corpora for their respective national languages. From a national perspective, parallel corpora are generally not that vital, therefore their compilation is usually undertaken by individuals, mostly linguists or translation scholars, who find it urgent to provide modern corpora-based contrastive descriptions of languages in the form of grammars and language teaching materials, to compile modern bilingual general and specialised dictionaries or modernize the already existing obsolete ones, etc. Such motives lie behind the compilation of the first French-Slovenian parallel corpus.

Making a corpus of texts in language A and their translations into a language B is a long and complex process for several reasons: (non-)availability of large quantities of (electronic) texts for less translated language pairs, securing permission from copyright-holders of texts for both languages in question, alignment and annotation. These issues will be discussed with regard to the development of a 2.5-million word French-Slovenian corpus of contemporary literary and journalistic texts, which is, in spite of its smallness, an invaluable source of data for contrastive studies and exploration of translation phenomena for the language pair in question.

A quantified comparative study of parallel speeches by the Chinese president and American president at a press conference

Hongwei Huang, Ying Yue
Mechanical Engineering College, China

This paper analyzes the similarities and differences between source native language (English) and translated language (English translated from Chinese) as used in the speeches delivered by President Hu of China and President Obama of the U.S. at the press conference held in Beijing on 17th of November, 2009. The speeches are considered parallel and comparable because they address similar issues and were delivered on the same occasion, for the same purposes and to the same audience. Therefore, one serves as a good example of native source language and the other as English used by non-native speakers for international communication.

Although the non-native variety of English is already widely accepted for international communication, and it is by no means surprising that non-native English deviates in differing degrees from the native one, the authors of this paper still hold the view that the closer the non-native variety is to the native variety, the better for mutual understanding, and that only by quantifying the differences can non-native speakers be made aware of the specific areas for their improvement. Therefore, this study first focuses on a quantitative analysis of the two speeches at the levels of words, word collocations, clauses and sentences; then compares the expressions the two presidents use to refer to themselves, to each other, and expressions they use to refer to their own country and each other's country and other countries; third, compares the usages of words for expressing views, standpoints and attitudes; and finally compares the types of clauses they employed in their respective speeches.

From such analysis, a deeper insight is gained into the speeches, and suggestions are offered to improve translation of similar speeches in the future.

I wish you/someone/people would... or *mělo by se*: A corpus-based study of sentences with *I wish* and their Czech equivalents

Michaela Martinková
Palacký University, Czech Republic

Finite clauses following *wish* are often analyzed as its content clause complements. In my presentation I will focus on the modal function of the *I wish* construction, using its Czech equivalents in texts translated from and into English (all taken from the Intercorp project).

While literal translations can be found (*přeju si* 'I wish', *přál bych si* 'I would wish'), and the dependent clause follows also other Czech verbs in the conditional (*chtěl bych* 'I would want', *byl bych rád, kdyby* 'I would be glad if...', *rád bych* 'I would gladly...'), these are not the dominant ones; there are also simple exclamative sentences with *kdyby* (*I wish I knew* 'to kdybych věděl'), and a significant frequency of the 'wish particle' *kéž* and other particles (*ať*) or particle-like expressions. The expression *škoda* 'pity, damage' , for example, comes up not only as an equivalent of *I wish* complemented by a clause with a verb in the past perfect (Dušková 1994:607) or with *could* followed by a past infinitive, but also of *I wish* complements containing a verb in the past simple. In these translations then, as well as in those where *I wish* is translated as *mrzí mě, že, je mi líto, že* ('I am sorry that') the Czech sentence has a reversed polarity, which supports theses often stated in linguistic literature about negative entailments or presuppositions of the *I wish* complements (Huddleston and Pullum 2002:1009). Variance can be found in translations of *I wish* complemented by a clause with *would* (cf. Searle 1975:65), where Czech may have a verb in the imperative. Interestingly, the same cannot be said about translations into English, where *I wish* occurs only very rarely, and the reason is perhaps not only a lower number of Czech to English translations in Intercorp.

Introducing *Comparapedia*: A new resource for corpus-based translation studies

Silvia Bernardini[1], Sara Castagnoli[1], Adriano Ferraresi[1, 2]
Federico Gaspari[1], Eros Zanchetta[1]
1. University of Bologna
2. University of Naples "Federico II"

Special purpose comparable corpora are among the most valuable resources for translators. Typically however they are not publicly available (differently from reference corpora), such that they have to be constructed for specific tasks as the need arises (Varantola 2003). This solution is clearly not ideal, since the resulting corpora are likely to be very small (if constructed manually) and rather low-quality - if an automatic procedure is used, e.g. the BootCaT method (Baroni and Bernardini 2004) - or else absorb more time and effort than the average translator is willing to spend on the task.

The present paper describes an attempt at tapping the potential of Wikipedia to build and make available to translation professionals and CBTS scholars *Comparapedia*, a large bilingual corpus formed of special purpose comparable sub-corpora of English and Italian. Adapting methods developed for Web-as-corpus construction (Baroni et al. 2009), all the bilingual data (i.e. the linked entries) are downloaded, cleaned, lemmatised, part-of-speech tagged, and indexed using the Corpus Workbench (Christ 1994). Based on the human-generated categories included in the entries, a topic categorisation is derived that is used to group entries into topic-specific corpora.

Compared with traditional corpora, *Comparapedia* is a completely new resource, which opens up novel possibilities for language professionals and translation scholars. Since comparability is established both at the micro-level (matching bilingual entries) and the macro-level (matching topic categories), the user can move between the two levels, searching the comparable corpus for hypothesis generation, then browsing the single bilingual text pair(s) for hypothesis confirmation. From the translation scholar's viewpoint, *Comparapedia* could be viewed as a hybrid *comparallel* corpus, since some of the entries may have been partly or entirely translated from their matching entry. This raises methodological issues concerning the adequacy of our traditional corpus typologies, as well as prompting theoretical questions that oblige us to rethink the status of both translation and corpus linguistics in the Web era.

Can "translation universals" survive in Mandarin? Idioms, word clusters, and reformulation markers in translational Chinese

Richard Xiao
Edge Hill University, UK

This paper explores three linguistic features which have so far been rarely studied in translation research, namely idioms, word clusters and reformulation markers, in translational Chinese as represented in a one-million-word balanced corpus of translated Chinese texts in comparison with native Mandarin represented in a comparable corpus of non-translated Chinese texts, in an attempt to verify whether some English-based, genre-specific features of translated texts can be generalized as translation universals in the light of evidence from Chinese, a language which is "genetically" distinct from English. The implications of our findings for translation universal hypotheses are also discussed. It is our hope that the study of translational Chinese will help to address limitations of imbalance in the current state of translation universal research, which has so far been largely based on translational English and confined to its closely related languages.

Phraseologies in English and Italian historical research articles

Silvia Cacchiani
University of Modena and Reggio Emilia, Italy

Research articles (RAs) have long been a major concern in research in English for Academic Purposes (for one, Swales 1990). Recent developments into corpus compilation and the development of query tools have increasingly enabled researchers to shift the focus on other genres and on cross-linguistic variation. Whereas EAP studies and register studies alike have chiefly looked at language variation across genres and disciplines (e.g. Hyland and Bondi (eds.) 2006), it is the purpose of this paper to concentrate on cross-linguistic and cross-cultural variation in English and Italian RAs of history.

Specifically, using the *WordSmith Tools* (1996) suite of programmes and *ConcGram 1.0* we shall query for metadiscoursive devices the *HEM-History* (c. 1.000.000 tokens), a collection of RAs in History built and currently held at the University of Modena and Reggio Emilia), and the *HEM-History, Italian subcorpus*, currently in its final stages of construction. We chiefly bank on Hyland's (2004, 2008) comprehensive account of metadiscourse and complement it with suggestions and insights from work on the cohesive role of bundles (Biber 2006), coherence relations (Knott and Dale 1994, Knott and Sanders 1998) and the and on Siepmann's (2005) taxonomy of second-level discourse markers, which also takes into account studies on the pragmatics of discourse markers (Fraser 1988), and work in rhetorical structure theory (Mann 1999).

The main emphasis will be on the phraseologies of: i. reformulators and resumers (English: *In a word, (And) (m/More) specifically/Specifically, also called, in another way*; Italian: *in altre parole; (e) (più) in particolare; si tratta di; in (estrema) sintesi)*; and ii. the related and partially overlapping category of summarizers and concluders (English: *finally, altogether, To conclude,* In conclusion, *So, X provides us with a grounds for concluding that*; Italian: *Concludendo; Per concludere, Veniamo ora alle conclusioni*).

The teaching of Ancient Greek as a foreign language, for students of immigrant status, at the high school and Lyceum educational levels

Evaggelia Kalerante[1], Simeon Nikolidakis[2], Efstathia Georgopoulou[1]
1. University of Western Macedonia
2. University of Peloponnese

In our article methods and manners are examined for the teaching of Ancient Greek as a foreign language for immigrant students in Greek schools. We are primarily referring to Albanian immigrant students, who comprise the majority population of immigrants, and have established themselves in Greece, whether they have studied in the school system of Albania, or if they have entirely done their studies in the Greek Educational system. Both of these groups have a difficulty in learning Modern Greek and getting accustomed to the Greek culture while simultaneously learning their mother tongue, Albanian, their language of origin. As well, the teaching of ancient Greek is for them yet another foreign language taught in Greek schools.

Our proposal, for the methodical education of the Ancient Greek language to immigrants is co-related with the cultivation of the language as the essence of civilization itself. From this, essential informative extracts, positions, and perspectives for man, society and social life in connection with space and time can be attained. In this manner we are referring to the teaching of a language which promotes the content of grammar and syntax as a mechanism to be used towards the better understanding of the language and towards the analysis of the texts' significance and their connection to the philosophy, history, and the development of science.

With this model as a base, a particular importance is placed on:

a) The choice of texts
b) The thematic approach
c) The bi-lateral scientific connection

Upon this base we can foresee the teaching in a group formation by which foreign students find worthy bibliographical sources found in libraries or via the internet. At the same time, at higher education levels, it is foreseen that communication and the exchange of prospects with corresponding departments in foreign countries will enhance pilot programs for the teaching of the ancient Greek language.

Explicitation and implicitation in translations between English and German: Evidence for the Asymmetry Hypothesis

Viktor Becher
University of Hamburg, Germany

Explicitation is the relative increase in explicitness in the process of translation (cf. Vinay and Darbelnet 1995), *explicitness* being definable as the verbalization of information that the addressee might be able to infer if it were not verbalized. Conversely, *implicitation* is the relative decrease in explicitness from source to target text. This paper presents first results from a study whose aim is to test Klaudy and Károly's (2004) Asymmetry Hypothesis, which claims that in any given language pair the number of explicitations will be higher than the number implicitations, regardless of the translation direction.

The hypothesis is investigated using a corpus of English-German and German-English translations of business texts (ca. 60 texts, 50,000 words). Explicitating as well as implicitating shifts were identified manually and categorized according to the level of the linguistic system on which they occur (e.g. verbal explicitations vs. nominal explicitations).

The results obtained so far (from the analysis of the first quarter of the data) may be summarized as follows:

1. The number of explicitations is indeed higher than the number of implicitations, regardless of the translation direction. There are 267 explicitations vs. 129 implicitations in the English-German translations and 203 explicitations vs. 146 implicitations in the German-English translations.
2. The type of explicitation performed most frequently is highly directiondependent. For example, English-German translators explicitate predominantly on the level of the noun phrase, whereas German-English translators prefer explicitations on verb level. This can be linked to previously observed contrasts between the two languages (cf. e.g. Fabricius-Hansen 2007).

The results do not only offer evidence in support of the Asymmetry Hypothesis, but also confirm and extend previous observations regarding grammatical and stylistic contrasts between English and German (see e.g. Hawkins 1986), which are highly relevant to translators working with these two languages.

Does valency theory provide a holistic approach to understanding language?

Renate Reichardt
University of Birmingham, UK

Over the last 30 years the distinction between lexis and grammar as separate areas of study has moved towards a theory which emphasises the strong interaction between the two (Singleton 2000:17). At the same time, there has been an increased emphasis on the study of phraseology and phrase patterns, at the expense of sentence grammar, in the English language classroom (Granger 2009).

However, if we accept that

- learners will always relate a new language to previous knowledge of language, mainly their native language (Lightbown and Spada 1999:45; Nunan 1999:40),
- word meaning depends to a great degree on the surrounding words (Firth 1957:11; Sinclair 1991:110) and
- language use is a creative process which requires negotiation amongst its users (Teubert 2004:98).

then, as a result, it appears to be important to provide learners with a more holistic approach to address the lexis-syntax continuum.

Valency theory, the property of a word to combine with or demand a certain number of elements in forming larger units (Emons 1974:34), offers an approach to investigate the interface of local grammar and lexis, which works for monolingual, as well as bilingual, analysis.

This paper looks at the verb CONSIDER and applies a bilingual (English and German) corpus study to investigate whether valency patterns relate to word meaning, i.e. translation equivalents. Using a corpus linguistic approach for grammatical analysis inevitably highlights the difficulties that are often faced by students and scholars alike when working with authentic texts where the analysis is often more varied and difficult than textbooks on general grammar usually imply (Hoey 2005:46).

It will emerge that although the valency patterns for CONSIDER to some extent show preference for translation equivalents, there is, on the other hand, also a great degree of freedom in the translations (Kenny 2005:162). This is a valuable finding for the language classroom, where language is often presented as ‘fact’.

Teaching prepositional verbs through corpora online

Emiliana Bonalumi
FATEC, Brazil

Corpus Linguistics offers to the teacher and to the foreign language learner the opportunity to research and extract data of the language in use, thorough corpora online. This paper has as one of the main purposes work with several prepositional verbs, through corpora online, allowing teachers and foreign language learners to observe differences of meaning found in the same verb with different prepositions, in comparison to their source languages. This works aims at developing research in the classroom, encouraging the active position of the teacher and/or student and, searching evidences or concrete examples of the frequencies of the linguistic features of a prepositional verb. Three inductive principles were used: (a) identification; (b) classification and; (c) generalization. Through these three inductive principles, teachers and / or learners will be able to identify the linguistic feature, through corpora online, classify it, and after a wide investigation, make generalizations about the observed data. This study draws on Data Driven Learning principles (Johns, 1991) and Corpus Linguistic studies (Berber Sardinha, 2000; 2004). After a brief introduction, teachers and learners will be familiarized to corpora online, aiming that by the end of the paper, they will be able to make their own research, identifying the prepositional verbs, classifying them and making generalizations about the observed data.

The future of translation "universals": What can localization tell us about general features of translation?

Miguel Jimenez-Crespo
Rutgers University / The State University of New Jersey, USA

Since the emergence of Corpus-Based Translation Studies, research into potential regularities in translational behavior or "general features of translations" has been at its core. However, translation scholars have also criticized this type of descriptive research from methodological and theoretical grounds (i.e. Tymozco 2005, 1998; Snell-Hornby 2006; House 2008). These have been mostly due to the tendency to overgeneralize the results obtained in limited translation subsets (Chesterman 2004). Additionally, it has been argued that general claims about translational behavior need should apply not only to current and past types (Venuti 2005), but also to those that might appear in the future. This paper reflects on these issues through an overview and discussion of previous corpus-based studies by the author into of a new translation modality, web localization. It intends to discuss how this and other new translation modalities, such crowdsourcing web localization (O'Hagan 2009), can help realign theoretically and methodologically how these research constructs are conceptualized.

Methodologically, the data is provided by the Spanish Web Comparable corpus compiled in 2006. It contains a body of 20,000 original webpages from original Spanish corporate websites and 20,000 localized webpages from the largest US corporations addressed at users in Spain. The overview of results will summarize the findings from previous studies on explicitation, conventionalization and sanitation carried out on the Spanish Web
Comparable Corpus (--- 2008a, 2009a, 2009b, 2009c), as well as an additional feature of translations, the tendency to "clone" source text structures (Larose 1998; --- 2009c). This tendency has not been conceptualized as a general tendency in previous studies, partly due to the fact that it cannot be directly studied using current methodologies and lexical analysis tools mostly focused on word-based metrics. The methodological and theoretical implications for research into "T-Translation" tendencies will be discussed, together with a reflection of how different aspects of web localization can help reshape basic tenets such as the notion of stable translated texts or the one to one relationship between a single translator and a translation product.

Counterfactual conditionals in focus: A contrastive analysis of French and Norwegian

Marianne Hobæk Haff
University of Oslo, Norway

This paper explores differences and similarities between French and Norwegian as regards the uses of tense forms in counterfactual conditionals. My analysis accounts for constructions with a protasis introduced by ***si*** in French, and by ***hvis*** or ***dersom*** in Norwegian. I have chosen this issue for at least two reasons. In the first place, because important reference grammars of French (including some recent ones) partly give an inadequate description of the issue, following grammatical tradition (e.g. *Foundations of French syntax* 1996, *Le bon usage* 2008, *Grammaire méthodique du français* 2009). In the same way, a recent reference grammar of Standard Norwegian (Bokmål), *Norsk referansegrammatikk* 1997, simplifies the presentation of the issue. I want to show that actual usages both in French and Norwegian are more complex than they claim. Secondly, to the best of my knowledge, a comparison of French and Norwegian conditional counterfactuals has not been undertaken before.

Two of the main patterns in French will be discussed: 1) *si* + *imparfait* + the apodosis VP in *conditionnel présent* (= ***Si*** + **IMP** + **COND PRES**) and 2) *si* + *plus-que-parfait* + the apodosis VP in *conditionnel passé* (= ***Si*** + **PQP** + **COND PASSÉ**). (Thus other possible combinations of tense forms will not be discussed). According to the grammars of French mentioned above, counterfactual present is exclusively expressed by the first of the patterns (1), and the second pattern is used only to render counterfactual past (2):

(1) Si j'étais riche, j'achèterais une grande maison.
(2) Si j'avais été riche, j'aurais acheté une grande maison.

They do not mention, however, that the second pattern too (**si + PQP + COND PASSÉ**) can express the counterfactual present, but not always with exactly the same meaning. Nor do they suggest the existence of a counterfactual used about the future (discussed in Robert Martin, 1971, 1983, 1991). It is interesting that Norwegian has two corresponding patterns to express counterfactual present, but used differently, as I will show. To interpret these constructions it is necessary to take into account the context, for instance the opposition telic vs. atelic verbs, or the presence of eventual adverbs in the clause. My analysis is based on examples from the following two text corpora: the Oslo Multilingual Corpus and a text corpus of *Le Monde*.

Acquiring instrumental sub-competence by building do-it-yourself corpora for business translation

Daniel Gallego-Hernández
University of Alicante, Spain

The aim of this paper is to share our experience in teaching how to build DIY corpora in business translation courses. Business and finance texts have a significant presence on the web and there is free software for Windows that can assist the translator in the different stages of the process of building DIY corpora from web resources. The model we propose in our courses takes into account these two realities and develops some of the sub-competencies that translation competence consists of, especially the instrumental one, which concerns the use of information and communication technologies and documentation resources.

The model has four basic stages. The first stage involves source-text analysis (genre and topic recognition). It requires the developing of extra-linguistic sub-competence related to business and finance texts, and instrumental sub-competence related to the sources containing parallel target-language texts. The second stage concerns browsing the Internet in order to find these texts and retrieving these resources by the use of search engines. In this stage, instrumental sub-competence is activated by the use of search-engine query languages and the evaluation of resources according to the needs of the translator. The third stage entails massively downloading these parallel texts. Instrumental sub-competence is also activated since now the translator can use download managers or offline browsers. Finally, the fourth stage is related to the conversion of parallel texts to TXT format so that the translator can exploit them with corpus query tools.

Each stage allows the use of different resources or different tools which can be used for corpus building even if the software the translator uses has not been specifically developed for this purpose. Translators must be aware of what they are reading or downloading. The goal of this model is basically to develop translators' instrumental sub-competence and to introduce for the first time those who are not familiar with command-line interfaces to the basics of corpora building as a translation resource.

Relevance-Based Framework for Explicitation: A New Alternative

Elisabet Titik Murtisari
Monash University, Australia

This paper will discuss an alternative way to approach the phenomenon of *explicitation* in corpus-based research by using Relevance Theory. The concept of explicitation itself, which is generally understood as 'the spelling out of information which is otherwise implicit in the source text', has been of special interest in translation studies because of its elusiveness. Different methods have been applied in the study, e.g. by the use of the discourse based concept of explicitness and the traditional encoded/inferred distinction. The studies, however, are somewhat difficult to compare since every study seems to have its own concept of explicitation.

In this paper I'd like to demonstrate how Relevance Theory may be able to shed more light on the elusive nature of explicitation and may also bring all the different approaches together in its future research. I will begin by explaining the current approaches in explicitation, and then compare them with the Relevance-based framework. Examples of analysis are taken from the comparison between John Steinbeck's novel *The Grapes of Wrath* and *Of Mice and Men* and their translation.

Epistemic expressions in contrast: The relevance of polysemy vs. grammatical form and epistemic scale in translation of French *sans doute / devoir* into Swedish

Carina Andersson
Uppsala University, Sweden

The aim of the presentation is to evaluate the importance of on one hand polysemy vs. grammatical form and on the other hand position on the epistemic scale and construction type for the appearance of a certain translation equivalent in the translation of French and Swedish expressions of epistemic modality.

Data is retrieved from C-ParaFraS, a Swedish-French bidirectional parallel corpus consisting of fiction and non fiction texts of approximately 2 millions words, mainly from the period 1970-2009.

French and Swedish have partly different means of expressing epistemic modality. This is typically demonstrated in the fact that French has a modal auxiliary verb *devoir* ('must') corresponding to a set of Swedish adverbial expressions (*antagligen, förmodligen, nog, väl*, etc.) and a smaller set of modal auxiliaries (*måste* 'must'*, borde* 'ought to'*, bör* 'should' and the older form *måtte*). The auxiliary *devoir* is polysemic and also expresses non epistemic meaning. French also has monosemic adverb *sans doute*.

The equivalent pattern shows that *devoir* in 2/3 o the cases corresponds to a Swedish verb, in 1/3 to an adverbial expression. *Sans doute* on the other hand corresponds only to a very small extent to a verb.

Sans doute indicate a relatively high degree of certainty. Its relative position on the epistemic scale can be demonstrated by test of the type *sans doute, même certainement*. The Swedish equivalent form does not always hold an equivalent position on the epistemic scale, but express the same, a higher or a lower degree of certainty. This appears to be the case when looked at both the source text equivalents and the translation equivalents. The hypothesis is that the construction type, for instance the concessive construction, correlates to the type of translation equivalent. Another hypothesis that will be tested is the idea that the translation paradigms differs according to text type (fiction or non fiction).

The construction of a Mandarin interlanguage corpus

Wai Lan Tsang, Yuk Yueng
University of Hong Kong, Hong Kong

Since the boom in corpus linguistics in the 1980s, different types of English corpora have been compiled (e.g. The British National Corpus (BNC) [general corpora] and The HKUST Computer Science Corpus [specialised corpora]). Mandarin Chinese, which is one of the key international languages with a growing number of learners of different nationalities, is trying to catch up with its development of corpora. In addition to the effort shown in various overseas universities (e.g. The Lancaster Corpus of Mandarin Chinese and The UCLA Written Chinese Corpus), many institutions in Mainland China, which is the largest provider of Mandarin language courses, have been contributing to Chinese corpus compilation.

The project reported in this paper aims at contributing to the development of interlanguage corpora (or learner corpora), one type of specialised corpora. It is a current initiative of constructing a Mandarin interlanguage corpus with both written and spoken output from Mandarin Chinese learners. These learners, with different L1s, are attending a two-year Certificate course on Mandarin Chinese at a tertiary institution in Hong Kong. Both their written and spoken production in the form of coursework and examination, amounting to 650,000 words and 10 hours of speech, is included in the corpus. The rationale, methodologies (i.e. collection, transcription and annotation) and features of the corpus are introduced. Potential theoretical and pedagogical contribution of the corpus is also discussed.

"SL shining through" in translational language: A corpus-based study of Chinese translation of English passives

Guangrong Dai[1], Richard Xiao[2]
1. Fujian University of Technology, China
2. Edge Hill University, UK

The translational language as a 'third code' has been found to be different from both source and target languages. Recent studies have proposed a number of translation universal (TU) hypotheses which include, for example, simplification, explicitation and normalization. This paper investigates the "source language shining through" put forward by Teich (2003). The hypothesis is that "In a translation into a given target language (TL), the translation may be oriented more towards the source language (SL), i.e. the SL shines through" (Teich 2003: 207), which has attracted little attention in translation studies. This study presents a detailed case study of English passive constructions and their Chinese translations based on comparable corpora and parallel corpora. This research explores a new aspect of TUs and offers another perspective for translation studies.

The first kind of complex noun phrases in Turkish language and their equivalents in English

Azizeh Khanchobani Ahranjani
Islamic Azad University, Salmas Branch, Iran

Words come together to form complex phrases in English language as well as in Turkish language. Although grouping of these phrases in different parts in English language is not the same as in Turkish language, by investigating the researches of English linguists and grammarians one can find out their equivalent forms. In Turkish language complex noun phrases consist of two parts: dependants, and independent ones. Taking into account these parts, they are subdivided in three groups. A comparative study of the first group in Turkish and English languages reveals both similarities and differences.

Design and development procedure of an English-Malay parallel corpus

Tengku Sepora Tengku Mahadi, Helia Vaezian, Mahmoud Akbari
Universiti Sains Malaysia, Malaysia

The current article aims to introduce an ongoing corpus compilation project at the School of Languages, Literacies and Translation of the Universiti Sains Malaysia. The research project includes constructing a 500,000 word English-Malay parallel corpus of legal texts, developing an English-Malay translation memory of legal texts from the corpus, and finally building a corpus-based glossary of legal terminology.

This paper explains the steps followed by the project team to design the corpus including the decisions over the size of the corpus, the number and the length of the texts, and the time span of the documents to be included in the corpus. The procedures followed to construct the corpus are also explained in detail. The paper finally elaborates on the process of developing the other products of the research project, namely the Translation Memory and the Glossary.

A study of implementing the lexical and discoursal modifications in translation: Regarding the translation of *The Kite Runner* from English into Persian

Reza Moghaddam Kiya[1], Fahimeh Sahraei Nejhad[2]
1. University of Tehran, Iran
2. Payam-e-Nour University, Iran

The present paper, while acknowledging the need to accommodate certain modifications in translation as a process to establish equilibrium between the source and target text, is aimed at focusing on the ways in which these modifications, both in lexical and discourse levels, are implemented in real situations. In general, modifications imply any alteration or manipulation in the form of the source text in order to convey its meaning properly when decoding it into the target language.

Lexical modifications, which are typically made in the cultural words, are implemented by the application of techniques such as borrowing, loan translation, explanatory-descriptive translation, coining, or functional or cultural equivalence. Discoursal modifications can also be implemented by methods such as adding, omitting, transposition, or by changing a word's part of speech or changing the whole discourse. These modifications are made for adjusting the grammatical structures, establishing linguistic or cultural equilibrium, for aesthetic purposes, or adjusting linguistic metaphors in both languages

In this paper, the image schemata which are the basis for the formation of the linguistic metaphors are categorized into two general classes namely *experiential* and *conventional*. Experiential schemata are gestalt, abstract and recurring patterns with bodily bases in our conceptual system which enable one to understand, reason and make sense of not only the human experiences and the physical world around him, but also the non-physical and abstract phenomena by their abstract extensions and metaphorical mappings. There can be as many experiential schemata as human physical experiences. However, the most prominent ones according to Mark Johnson (1987) are *Containment Schema, Path Schema, Force Schema, and* so on.

Conventional Image Schemata which are culture-specific are those abstract and unified information structures in our mind based on our social conventions or agreements. These are used as shared background knowledge among the language community and are crucially effective in the formation and understanding of a discourse where some of the facts in the discourse is taken for granted it is taken for granted by the participant and as such, are omitted from the discourse.

Based on randomly-selected passages from the translated novel, *Badbadakbaz*, and comparing it with its original English manuscript, *The Kite Runner*, it was concluded that all kinds of modifications in translation were implemented in line with the image schemata that existed in the minds of the translation readership or addressees. Furthermore, all lexical modifications are specifically made in line with the conventional schemata, while discoursal ones were made in line with both conventional and experiential schemata.

A corpus-based study on the translation of Aerospace China White Paper

Yuyin He
Beihang University, China

Wolfgang Teubert, one of renowned post-Firthian corpus linguists, has argued that corpus research is "a key element of almost all language study." Corpus-based study has been widely employed in translation study. Mona Baker has built Translational English Corpus. Professor Wang Kefei has built Chinese-English Parallel Corpus with insightful research results regarding features of translated Chinese.

This investigation, based on both the Chinese versions and English versions of Aerospace China White Paper 2000 and 2006, aims to study language features between Chinese and English and syntactic translation features from Chinese into English. Research tools involve Antconc, ICTCLAS, GoTagger and WinAlign of CAT software SDL Trados 2009.

First, Antconc is used to show token/type ratios of English versions of White Paper 2000 and White Paper 2006. Then, both the Chinese versions and English versions of Aerospace China White Paper 2000 and 2006 and their POS tagging and segmentation versions either by ICTCLAS or by GoTagger, altogether eight versions are put to use in Antconc for their word lists to display language features between Chinese and English and the changes of content foci between 2000 and 2006. What's more, ICTCLAS is utilized to study keywords and their weighting in the Chinese versions of 2000 and 2006. Last but not the least, Trados WinAlign function is adopted to achieve syntactic alignment between Chinese and English. Syntactic translation features are studied concerning long Chinese sentence segmentation and Chinese subjectless sentences.

Some limitations should be taken into consideration. A comparable corpus of aerospace white paper from English native countries should have been created for a better understanding of the translated English in this investigation. Different rhetoric structure in white paper composition makes it hard to find an appropriate counterpart for the Chinese one.

Interlingual pronoun errors in English-Arabic translation

Reima Al-Jarf
King Saud University, Saudi Arabia

Unlike English, Standard Arabic has two forms of subject pronouns: (i) **Independent**: *?na (I), nahnu* (we), *?anta (you masculine singular), ?anti (you, feminine singuar), ?antumaa (you, dual), ?antum (you, masculine pural), ?antunna (you, feminine plural)*; *huwa (he), hiya (she), humaa (they, dual), hum (they, masculine plural), hunna (they, feminine plural*); and (ii) a **pronominal suffix** that is an integral part of the verb such as *katab-tu (I wrote), katab-naa (we wrote)*; *katab-ta (you wrote, masculine singular), katab-ti (you wrote, feminine singular), katab (he wrote, uninflected form), katab-tumaa (you wrote, dual), katab-tum (you wrote, masculine plural), katab-tunna (you wrote, feminine plural); katabaa (they wrote, masculine dual), kataba-taa (they wrote, feminine dual), katab-uw (they wrote, masculine plural), katab-na (they wrote, feminine plural).* Independent subject pronouns are commonly used in nominal sentences, not verbal sentences. Use of independent pronouns in verbal statements depends on syntactic, pragmatic, discoursal and semantic factors available in a particular context.

The present study investigates translation students' awareness of the syntactic, pragmatic and discoursal restrictions that determine the use of Arabic subject pronouns when translating connected discourse from English into Arabic. An error corpus of faulty uses of Arabic independent subject pronouns was collected from the translation projects of senior students majoring in translation. Syntactic, pragmatic and discoursal criteria were used to judge the deviations. It was found that students translate imitatively rather than discriminately (L2 to L1 transfer). Since English sentences begin with a subject pronoun such as *I, he, they,* the students used an independent pronoun followed by a verb + pronominal suffix in declarative, affirmative statement, without realizing that the subject is contained in the verb, and use of *?na* or *huwa* is redundant. Implications for increasing students' awareness of pragmatic, discoursal and syntactic constraints in translating English pronouns into Arabic will be provided.

Features of non-literary translated language: A pilot study

Haidee Kruger, Bertus van Rooy
North-West University, South Africa

Corpus-based research into the features of translated language has most often been based on comparable corpora of translated writing and original writing. Much of this research has utilised the Translational English Corpus (TEC), or subcorpora drawn from it, and subcorpora from the British National Corpus (BNC). These (sub)corpora largely consist of literary or other imaginative texts. The question that arises is whether the features of translated language that have been identified (particularly features related to explicitation, simplification and normalisation) may not, in some way, have been influenced by the particular generic make-up of the corpora.

In the terms of Biber (1988), fictional writing is characterised by a blend of features of informational and involved production. However, compared to more informational writing, such as academic prose, official documents and press reportage, fiction contains a greater incidence of some features of involved production, such as *that* deletion, the use of contractions, and a lower type/token ratio. These are some of the features that have been investigated in studies of explicitating tendencies in translated literary language (see Olohan, 2004), which have suggested that the greater prevalence of more explicit forms (where a choice exists between the explicit and the more economical form) in translated fiction means that translated fiction is closer to informational production than involved production.

The research question investigated by this project is whether the same tendencies towards explicitation, normalisation and simplification are evident in translated texts that are not literary or imaginative in nature: in other words, in text types that, by their informational and non-involved nature, are already characterised by high levels of explicitness and fairly standard usage, unlike literary texts. As such, it links to recent work by, for example, Williams (2005).

In order to investigate the research question, this pilot project compiles a small corpus of approximately 400 000 words of translated non-literary, largely informational, English texts generated in the South African context, specifically service translations mostly done from Afrikaans to English (but also a number of other languages). It replicates and adapts some of the methodologies in existing studies to investigate whether explicitation, normalisation and simplification are features of service translation as much as of literary translation, and whether these features are also prevalent in translation taking place outside the Anglo-European context. The corpus of translated texts is compared with a reference corpus drawn from the ICE-SA corpus, compiled to be as similar as possible to the corpus of translated texts. Using WordSmith Tools, the two corpora are investigated for reduction features typical of more informal registers, but still found in lower proportions in written texts, such as *that* deletion, the use of contracted forms and the omission of optional sentence elements, as well as indications of informational density, such as average sentence length and type/token ratio, and findings are compared with findings from existing studies.

Translating ambiguous lexical items using a parallel corpus: A case study of "good" in the EAPCOUNT

Hammouda Salhi
University of Carthage, Tunisia

The paper derives from a feeling of much apprehension and bewilderment about the way lexical ambiguity has been dealt with by translation researchers over a considerable period of time. While linguistic research puts an emphasis on the centrality of this phenomenon in language, where most lexical items are claimed to be ambiguous to some extent (Pustejovsky, 1995), translation literature seems to focus on the restricted, exceptional and accidental side of the problem. Therefore, in order to be more efficient in a translation classroom, traditional approaches should, arguably, be revised to handle lexical ambiguity as the norm rather than the exception in language.

The paper attempts to present an empirical and systematic corpus-based method for trainee translators allowing them to discuss the often undermined and neglected problem of *complementary polysemy* (CP) (Pustejovsky, 1991 and 1995) of some lexical items in SL texts and help them find appropriate equivalent, or near-equivalent, terms in the TL. As an example, I will examine the highly polysemous adjective *good* and its Arabic equivalents in the EAPCOUNT, the English-Arabic Parallel Corpus of United Nations Texts, a parallel corpus of about six million word tokens. Results show that almost with every usage of the adjective, there is a different novel meaning and, therefore, a different equivalent term and that the aims of resolving the ambiguity of this adjective and of establishing equivalence at both word and collocation levels heavily depend on the head noun that *good* modifies. The results strongly suggest that a corpus-based approach is highly appropriate in the translation classroom when dealing with the problem posed by lexical ambiguity.

Measuring mean sentence length in translated and non-translated Chinese texts: A corpus-based study

Ting-hui Wen
National Chiayi University, Taiwan

Simplification first emerged as a topic in applied linguistics and as a technique in second language acquisition for teachers, learners and writers/editors (Ferguson 1971; Blum and Levenston 1978; Davies and Widdowson 1974; Mountford 1976; Honeyfield 1977; Blum-Kulka and Levenston 1983). Later, Blum-Kulka and Levenston (ibid) were the first to point out that simplification was also a strategy adopted by translators when they encountered lexical voids, where there is a gap between source language and target language. In 1996, Mona Baker proposed four translation universals: explicitation, simplification, normalisation and levelling out. Sara Laviosa-Braithwaite (1996) and Laviosa (1997, 1998, 2002) then based on Baker's hypothesis and Blum-Kulka and Levenston's assertion, created the Translational English Corpus (TEC), adopted measures from corpus linguistics and investigated simplification in translated texts. Unlike Blum-Kulka, Baker and Laviosa proposed simplification as a universal or a recurrent translation feature, independent from the influence of the source language.

Simplification in translation can be manifested in the following three levels: translated texts tend to display a shorter average sentence length, draw on a more restricted vocabulary and contain a lower information load, than non-translated texts in the same language.

The manifestations may be quantified through corpus-based methods of comparative analysis, measuring: 1) mean sentence length; 2) lexical variety with type/token ratio, percentage of high frequency words and percentage of list heads; and 3) information load with lexical density.

A corpus of modern Chinese mystery fiction (CCCM) has been compiled especially for the purpose of the current project, with two subcorpora of translated and non-translated mystery fiction.

The present research aims to investigate, using corpus-based methods, the phenomenon of simplification in translated, compared to non-translated, Chinese texts. This paper focuses on measuring sentence length, analyzing the results of mean sentence length and its additional measures: mean sentence length in terms of characters, mean sentence sub-unit length in terms of words and mean sentence sub-unit length in terms of characters. Mean sentence length is proposed as a measure of simplification: if a text has shorter mean sentence length, it is assumed to be simpler for readers to comprehend. Since translated texts are hypothesized to be simpler than non-translated text, they would presumably exhibit shorter mean sentence length.

The measure of mean sentence length and its additional measures render consistent results showing that the translated texts exhibit shorter sentence length than the non-translated texts.

Using corpora to define target-language use in translation

Rudy Loock
University of Lille 3 / CNRS UMR Savoirs, Textes, Langage 8163, France

When translating a text from the source language to the target language, translators have to convey the same semantic/informational content while abiding by the morphological, syntactic, and semantic constraints of the target language. In order to provide translations that are as natural as possible, translators also have to be aware of language use, i.e. the way native speakers of the target language actually use that language. This will enable them to make a choice between two grammatically-acceptable translations, selecting that one that will sound the most natural for the speakers and readers of the target language.

From a lexical point of view, this explains why translators translating from English to French should translate *swine flu* with *grippe A* or *grippe H1N1* rather than resort to a literal translation, *grippe porcine*, which does exist, but is only rarely used by the French media. Syntactically, the language-use constraint explains why many SVO English sentences will be translated with a cleft sentence in French: *Anna died before dawn*, which can be translated literally (SVO sentence), can also be translated with a cleft structure (*c'est avant l'aube qu'Anna est morte*).

The aim of this presentation is to explain how the use of corpora can help define language use. This is a difficult task requiring thorough knowledge of the target language. Through the use of press articles and novels written in English and French, and the translation of these same texts into English or French, as well as the compilation of statistics, we will suggest a methodology for defining language use. This methodology must be followed alongside the morphological and syntactic rules that translators have to respect.