The Institute for Language and Speech Processing / "Athena" Research Centre is hosting a colloquium on

Learner Corpora for less commonly taught languages:

Design, processing and prospects for Second Language Acquisition and Education

Learner corpora have been established as a valuable instrument for detecting patterns of language development, serving, thus, as a basis for empirical research and educational interventions. The primary goal of the colloquium is to examine and reflect on concepts, principles and perspectives of Learner Corpus research and applications. The colloquium focuses on less commonly taught languages and seeks to explore best practices and lessons drawn from research in widely spoken languages, such as English. Recent advances will be presented by invited speakers, leading figures in the field of Learner Corpus research.

9.00 – 9.30	Registration
9.30 – 10:00	Opening Developing a learner corpus for language-minority children *Dr Maria Tzevelekou, Research Director, ILSP/"Athena" R.C.*
10:00 – 11.00	Towards more and better learner corpora for educational applications *Prof. Sylviane Granger, CECL, UCL*
11.00 – 11.30	Coffee break
11.30 – 12.30	On systematically characterizing learner language: a computational and corpus linguistic perspective *Prof. Detmar Meurers, Department of Linguistics, University of Tübingen*
12.30 – 13.30	The ASK corpus - a learner corpus of Norwegian as a second language; design, annotation and search interface *Prof. Kari Tenfjord, Department of Linguistics, University of Bergen*
13.30 – 14.30	Lunch
14.30 – 15.30	Building a Learner Corpus of Czech; Lessons Learned *Dr Jirka Hana, Researcher, Institute of Formal and Applied Linguistics, Charles University, Prague*

"Kostis Palamas" Building, University of Athens
Akadimias 48 & Sina

Tuesday 8 April 2014

Sylviane Granger
Director of the Centre for English Corpus Linguistics at the Université catholique de Louvain and President of the Board of the Learner Corpus Association

Towards more and better learner corpora for educational applications

Learner corpora are electronic collections of written or spoken data produced by second or foreign language learners. At first limited to one language – English – and a highly restricted range of text types, the scope of learner corpus research has progressively expanded and diversified. In the first part of my presentation I will provide an overview of the main learner corpora collected at the Centre for English Corpus Linguistics of the University of Louvain. These include the *International Corpus of Learner English* (Granger et al 2009), the *Louvain International Database of Spoken English Interlanguage* (Gilquin et al 2010), the *Longitudinal Database of Learner English,* the *Varieties of English for Specific Purposes dAtabase, the Corpus of Native and Non-native EFL Classroom Teacher Talk* and *the French Interlanguage Database.* I will highlight the benefit that can be derived from complementing learner corpus collection with other types of corpus data, in particular comparable corpora representing the target norm, bilingual corpora and textbook corpora. In the second part of my presentation I will focus on three of the main stages in learner corpus research (Granger 2012): corpus design, corpus collection and corpus annotation. For each of them I will highlight some of the lessons learnt from past research and suggest some priorities for the road ahead. As regards applications I will advocate a learner-centred, needs-driven and localized approach and briefly describe a web-based writing aid based on this approach (Granger & Paquot 2010).

References

Granger, S. (2012). How to use foreign and second language learner corpora? In Mackey, A. & Gass, S.G. (eds). *A Guide to Research Methods in Second Language Acquisition*, 7-29. Basil Blackwell.

Gilquin, G., De Cock, S. & Granger, S. (2010). *Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM.* Louvain-la-Neuve, Presses universitaires de Louvain.

Granger, S., Dagneaux, E., Meunier, F., Paquot, M. (2009). *The International Corpus of Learner English. Handbook and CD-ROM. Version 2.* Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. & Paquot, M. (2010). Customising a general EAP dictionary to meet learner needs. In Granger, S. & Paquot, M. (eds.) eLexicography in the 21st century: New challenges, new applications. Proceedings of ELEX2009. Louvain-la-Neuve, Presses universitaires de Louvain, 87-96.

Detmar Meurers
Professor of Computational Linguistics and head of the Theoretical Computational Linguistics group at the Department of Linguistics, University of Tübingen

On systematically characterizing learner language: a computational and corpus linguistic perspective

Learner corpora as collections of language produced by second language learners have been systematically collected since the 90s, and with readily available collections such as the ICLE for English and FALKO for German there is a growing empirical basis on which theories of second language acquisition and the interlanguage systems can be informed. Yet, as soon as the research questions go beyond the acquisition of vocabulary and constructions with unambiguous surface indicators, corpora must be enhanced with linguistic annotation to support efficient retrieval of the data that is relevant for such research questions. In contrast to the different types of linguistic annotation schemes which have been developed for native language corpora, the discussion on which linguistic analysis and annotation is meaningful and appropriate for learner language is only starting.

When formulating linguistic generalizations, one generally relies on a long tradition of linguistic analysis that has established an inventory of categories and properties to abstract away form the surface strings. In this talk, we will see that traditional linguistic categories are not necessarily an appropriate index into the space of interlanguage realizations and their systematicity, which research into second language acquisition aims to capture. Complementing the language explicitly given in the corpus, we also consider the need for information about the task which resulted in the corpus and the learners who produced it for interpreting and annotating learner data.

Kari Tenfjord
Professor at the Department of Linguistics, University of Bergen. Leader of the ASKeladden project in the framework of which ASK, The Norwegian Language Learner Corpus, was developed.

The ASK corpus - a learner corpus of Norwegian as a second language; design, annotation and search interface

In my talk I will present the ASK corpus which contains 1739 texts written in Norwegian as a second language and personal data about the learners. The texts are written essays from two different tests measuring language performance at two different levels (supposed to be at or above level B1 or B2). A reassessment of the texts according to CEFR level descriptions, were performed in 2009. The texts and personal data are marked up in XML according to the TEI Guidelines. Error coding is done manually using a relatively simple system developed for ASK. To compensate for this simple system, the texts are grammatically tagged using an automatic tagger developed for standard Norwegian, “The Oslo-Bergen tagger”. The latest version of the ASK corpus is accessible in a newly designed and implemented corpus management platform (Corpuscle).

Jirka Hana
Senior Researcher at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague

Building a Learner Corpus of Czech; Lessons Learned

Czesl is a corpus of texts produced by non-native speakers of Czech. To adequately annotate the deviations in Czech word-order and a fairly complex inflectional morphology, the corpus uses two tiers of annotation, each tier correcting different types of errors. Links between the tiers allow capturing errors in word order and complex discontinuous expressions. Errors are not only corrected, but also classified. We combine (1) a grammar-based taxonomy of errors in spelling, morphology, morphosyntax, lexicon and style, and (2) a formal error classification based on surface alternations. The annotation scheme was tested on a data set of approx. 175,000 words with fair inter-annotator agreement results. In addition to presenting the current state of the corpus, I will discuss various decisions (design, workflow, technical) made during its creation and revisit them with the advantage of hindsight.

Organising committee
Dr. Maria Tzevelekou, Research Director, mtzevelekou@ilsp.gr
Dr. Maria Giagkou, Associate Researcher, mgiagkou@ilsp.gr
Dr. Vicky Kantzou, Associate Researcher, vkan@ilsp.gr
Dr. Spyridoula Stamouli, Associate Researcher, pstam@ilsp.gr

Institute for Language and Speech Processing-“Athena” Research Centre