Short Visit Grant
Reference Number : 4742
Application submitted : 16/01/2012 10:34:58
Place of visit: Vienna, University of Vienna and OEAW
Period: 14/05/2012 – 24/05/2012

FINAL REPORT
Phonotactic effects on morphological structure – Psycho-computational studies on Italian, French, English and German

(co-granted Dr. C. Celata, Ref. No 4741)

Applicant:

Basilio **Calderone** <basilio.calderone@univ-tls2.fr>
CLLE-ERSS, CNRS & Université de Toulouse Le Mirail
5 Allées Machado, 31058 Toulouse

Research partners:

Wolfgang U. Dressler <wolfgang.dressler@univie.ac.at>
Sylvia Moosmüller <sylvia.moosmueller@oeaw.ac.at>
Katharina Korecky-Kröll <katharina.korecky-kroell@oeaw.ac.at>
Austrian Academy of Science (ÖAW)
Dr. Ignaz Seipel-Platz 2
1010 Vienna
Nikolaus Ritt <nikolaus.ritt@univie.ac.at>
University of Vienna
Dr.-Karl-Lüger-Ring 1
1010 Vienna
Karlheinz Mörth <moerth@oeaw.ac.at>
Austrian Academy of Science (ÖAW)
Institut für Corpuslinguistik und Texttechnologie (ICLTT)
Sonnenfelsgasse 19/8, 1010 Wien

§1. Purpose of the visit

The purpose of the 10-day visit was that of establishing a collaboration with the researchers in Vienna to further develop our PHACTS-based psycho-computational research on the phonotactics-morphology interface (Calderone & Celata, 2012). In particular, two topics were selected before the visit as the most fruitful areas of joint research: (i) the phonotactic processing of German clusters (with an expected extension to English); (ii) German plurals in acquisition and processing. (See Application for further details). The visit was therefore aimed to (a) develop a psycho-computational experiment to test native speakers' processing of different types of German clusters,

and (b) discuss ways of collaboration on German plurals and evaluate the possibility of a cross-linguistic approach in the study of German and Italian plural processing.
In particular, since I am the responsible of the computational part of the project, my specific aims were those of evaluating the possibility of accessing different corpora of spoken or written German (preliminary data on German clusters were collected from the Celex database (Baayen et al. 1995) before our visit), developing a feature-based phonological codification for German to be used in PHACTS-based simulations, running some preliminary simulations with the frequency data for the relevant German clusters and evaluating their output with respect to the behavioral expectations.

§2. Description of the work carried out during the visit

During our 10-day visit, there were two plenary meetings (one at the beginning, the other two days before the end of the visit) and several partial meeting with the members of the group. During the first meeting, our psycho-computational project on the phonotactic-morphology interface was presented with particular reference to the Morphonotactic Hypothesis (Ritt, Dressler & Moosmüller 2012) and to the ways of implementing a series of psycho-computational experiments on German morphonotactic and phonotactic clusters (see point (i) in §1 above). Shortly, morphonotactic clusters are those which result from morphological concatenation and contain a morphological boundary inside, while phonotactic clusters do not contain any morphological boundary (and they are therefore said to be 'lexical'). In German, there are consonant sequences that are exclusively morphonotactic (such as word-final /m#st/ or /x#st/, where # indicates a morphological boundary) as well as sequences that can be either morphonotactic or phonotactic (e.g., /p(#)st/, /k(#)st/).
Concerning the computational part of the project, our PHACTS-based simulation of the development of phonotactic knowledge from lexical and distributional information was presented with some detail to the group during the first plenary meeting. The implications for the refinement of the Morphonotactic Hypothesis were clearly put forward: PHACTS simulations, once completed, may provide support to the hypothesis that the morphonotactic regularities of a language are derived from the speakers' exposure to the speech input, thanks to the inherently relational nature of the lexicon. This hypothesis may be true in particular for relatively highly inflecting languages such as German. On that basis, paradigms of 'multiple learning' were discusses, that is, we discussed the possibility of testing PHACTS' generalizations with respect to the learning of German corpora of different size, nature, and historical period. Thus, the activity of the following days was centered on PHACTS' learning procedures, while in the second part of the visit, some trial simulations with words containing morphonotactic and phonotactic clusters were run. The preliminary results showed that variations in the phonological codification given in input to the system may have strong influence on the patterns of result. Than a feature-based phonological codification for German was developed, with the help of the German phonologists in the group.
Concerning the research line on German plurals (see point (ii) in §1 above), the possibilities of a collaboration with the French laboratory where I am affiliated was discussed during the second plenary meeting. Then the development of a computational model for automatic learning and generalization of the statistical regularities in the application of German plural formation processes was envisaged. The collaboration between the Viennese group and the Toulouse laboratory is being concretized in the form of a joint Austrian-French project proposal which is being submitted during these days (see §4 below for more details).

§3. Description of the main results obtained

Concerning the computational side of the project, one of the fundamental advancements that has been realized during this visit has to do with the quantitative treatment of the phonotactic regularities of German on the basis of different corpora. This constitutes the basis of the ongoing simulations that are currently under development to investigate the computational effort in the processing of morphonotactic clusters within a PHACTS environment.

In particular, the collaboration of Dr. Karlheinz Mörth of Institut für Corpuslinguistik und Texttechnologie (Austrian Academy of Science) turned out to be crucial for the evaluation of CELEX's characteristics with respect to other corpora. Larger corpora of Standard Austrian and Standard German German were selected to be exploited in future simulations, once a phonological codification will be provided. In addition, some historical corpora of past varieties of Austrian German were evaluated and included in future simulations. However, for the purposes of the current simulations, CELEX continues to be the fundamental resource for the PHACTS's learning phase.

With Prof. Sylvia Moosmüller we defined the phonological grid of features to encode all vowels, diphthongs and consonants of Austrian German according to a binary codification. of the data according to features defining the place and the manner of articulation in the German language.

Some exploratory simulations have also been realized. From the computational point of view, the hypothesis that the same /pst/ sequence is processed differently when it is included in *Obst* than when it is included in *liebst* by German native speakers or children acquiring German as their native language is based therefore mostly on the observation that /p#st/ and /pst/ (with '#' signaling a morphological boundary henceforth) are different in the lexicon for – at least – the following characteristics: (i) their relative frequencies (e.g., there are much more /p#st/-ending than /pst/-ending words in German); (ii) their being or not being included in paradigmatic alternations with other sequences (e.g., the fact that *lieb-st* contrast paradigmatically with *lieb-e*, while *Ob-st* does not contrast with **Ob-e*); (iii) possibly, the relative frequency of their preferential phonotactic environments (such as pre-consonantal vowels). All these factors are likely to interact in the mind of the speaker and should be comprehensively accounted for computationally in order to predict whether and to what extent /p#st/ and /pst/ are truly different objects in language processing. Frequency is of special relevance here, since both type numerosity and frequency of occurrence for the relevant clusters should be considered, differently from the general account, which tends to look after type numerosity only (e.g., Ritt et al. 2012). Thanks to this corpus-based phonological information, and differently from general edit distance algorithms, PHACTS appears to be particularly suitable for modeling the combined role of phonotactic regularity and lexical frequencies in lexical generalizations (Celata et al. 2011). In our preliminary simulations, PHACTS was trained on a phonologically encoded subset of CELEX for German. Then, the system was provided with the following lists of German words: (a) words with lexical /kst/ in final position (e.g., *Text*); (b) words with morphological /k#st/ in final position (e.g., *fliegst*). A vector representation of each word of the lists was created and plotted on the U-matrix conventional representation (Fig. 1). On that basis, PHACTS was asked to output one average vector representation for each of the two different sequences in (a) and (b). This exploratory procedure was the prerequisite for the currently developed simulations on PHACTS's treatment of phonotactic vs. morphonotactic cluster.

§4. Future collaboration with host institution

The visit was the occasion for programming with some detail a joint Austrian-French project regarding the role of phonotactic generalizations in German plural formation, investigated by means of supervised algorithms of classification tasks, that I have already implemented in previous studies on different languages.

This joint initiative of collaboration between the Austrian Academy of Science and the Laboratoire CLLE-ERSS of Toulouse has a computational core overall. A project proposal entitled « La simulation de l'acquisition de la (mor)phonotaxe et des pluriels en Allemand » has been submitted within the framework of the Partenariat Hubert Curien (PHC) – "Amadeus" programme for the support of French-Austrian scientific cooperation. The project gathers 5 researchers from the CLLE-ERSS of Toulouse and 3 researchers from the ÖAW of Vienna, with W.U. Dressler as the scientific responsible for the Austrian group and myself as the scientific responsible of the French group.

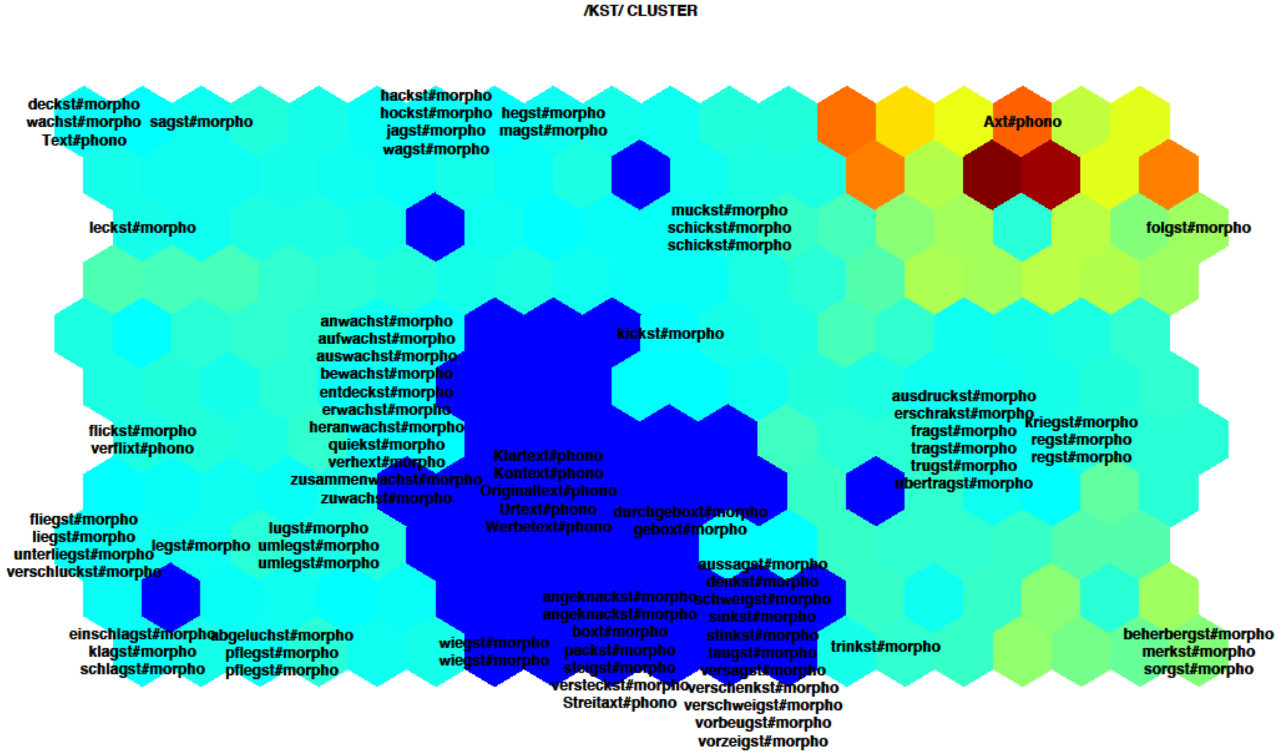

Figure 1. /kst/ and /k#st/ lexical items in PHACTS-based representation.

A new occasion of discussion will come from the workshop “Theory and Evidence in the Study of Phonotactcs” organized by K. Dziubalska-Kołaczyk within the 43rd Poznań Linguistic Meeting (8-10 Septeber 2012), in which Prof. Dressler’s team is in the list of participant. My talk (in collaboration of C. Celata) will concern the emergence of phonotactic clusters in the German language

I am also the co-organizer of an international workshop (“Phonotactic grammar: theories and models”, http://linguistica.sns.it/phonotactics/home.html) to be held in September 2012 in Cortona, Italy. Profs. Dressler and K. Dziubalska-Kołaczyk are among the keynote speakers of the event.

§5. Projected publications/articles resulting or to result from the grant

Several collaborative articles are expected to result from this ESF grant; they will be submitted to peer-review conferences and journals.

References

Calderone, Basilio and Chiara Celata, 2012. PHACTS about activation-based word similarity effects. *Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss*, pages 33–37, Avignon, France, April 24, 2012.

Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers. 1995. The CELEX lexical database (release 2). Linguistic Data Consortium, Penn.

Ritt, Nikolaus, Wolfgang U. Dressler and Sylvia Moosmüller. 2012. The dynamics of morphonotactics. Manuscript, University of Vienna.