NETWORDS EXCHANGE VISIT GRANT SCIENTIFIC REPORT

Description of the project: Towards a formal model for Canonical Typology

Period spent: January 4^{th} to February 28^{th} 2012

Name of the visitor: Géraldine Walther
Univ. Paris Diderot, Sorbonne Paris Cité, LLF, UMR 7110, 75013, Paris, France
geraldine.walther@linguist.jussieu.fr

Name of the host institution: Surrey Morphological Group, University of Surrey
Prof. Greville G. Corbett, g.corbett@surrey.ac.uk
Dr. Dunstan P. Brown, d.brown@surrey.ac.uk

General background

The canonical approach to typology [Corbett, 2003] has recently proven to be particularly fruitful within the field of morphology. With respect to inflection, the notion of *canonicity* corresponds to an abstract and ideal system from which deviation occurring in the world's languages can be measured concretely. As an absolute measuring point, canonicity is meant to allow comparison of different languages w.r.t. how far apart they are from the canonical stage.

However, work in canonical typology did not, at the time of the project proposal, rely on quantitative measures, nor was it associated with a formal model capable of encoding the observable deviation from canonical inflectional behaviour. The aim of the research project was thus to advance the definition of a formal model specifically designed for encoding inflectional phenomena as treated within *Canonical Typology*.

As part of my PhD work, I am developing PARSLI, a formal model based on the notion of *canonicity*. PARSLI stands for "PARadigm Shape and Lexicon Interface" [Walther, 2011]. At the beginning of the visit, it already covered non-canonical inflectional phenomena such as *suppletion* [Boyé, 2006, Bonami and Boyé, 2006], *deponency* [Baerman *et al.*, 2007], *heteroclisis* [Stump, 2006], *defectiveness* [Baerman *et al.*, 2010] and *overabundance* [Thornton, 2010]. Extending the coverage in terms of encoding of non-canonical phenomena was part of the aims of the visit at the University of Surrey.

My model aims at giving a formal representation of both canonical and non-canonical inflectional phenomena and offers the means to provide a quantitative interpretation of the observable deviations from canonicity.

It is also designed for representing both the morphological information stored within the (morphological) lexicon, on the one hand, and the (morphological) grammar, on the other. It provides a description of each lexeme of a given language with regard to its own paradigm's structure. The paradigm structure is, among other things, represented as a pattern of lexically specified *stem zones*[1] and a generalisation thereof in the form of *inflection zones*. It is the paradigm's structure that accounts for the various non-canonical inflectional phenomena mentioned above.

So far, the development of PARSLI had gone hand in hand with medium and large scale lexical resource development [Sagot and Walther, 2010, Walther and Sagot, 2010, Walther *et al.*, 2010] that implements the inflectional properties encodable within PARSLI. These implementations, encoded within a modified version of the Alexina framework [Sagot, 2010], have shown the relevance of using PARSLI as a means for encoding non-canonical phenomena within a complete description with low complexity [Sagot and Walther, 2011, Walther and Sagot, 2011].

1 Purpose of the visit

The aim of the research visit within the Surrey Morphology Group at the University of Surrey was fourfold:

1. Make use of the available typological expertise will help improve the quality of PARSLI w.r.t. to the coverage of all studied non-canonical phenomena and ensure the theoretical soundness of the model by verifying that the way PARSLI represents inflectional phenomena is coherent with the theoretical insights won by advances in Canonical Typology. In return, the improved model would be made available and usable for the typological comparison of languages through precise quantitative measures. It would also enable the quantitative cross-linguistic comparison of non-canonical behaviour, using, for example, PARSLI measures such as the *Heteroclicity Index* and the *Deponency Index*.

2. The Surrey Morphology Group has collected vast amounts of data on non-canonical phenomena. Exposing PARSLI to the available data was meant to provide a means for testing the model and ensure it covers all currently known non-canonical phenomena.

3. The development of PARSLI also entails a computational component, while the Surrey Morphology Group is also developing complete morphological analyses within the framework of *Network Morphology* which itself relies on the DATR language. Both models are complementary. *Network Morphology* provides complete descriptions and implementations of morphological systems, but it does not target the explicit expression of the notion of canonicity. PARSLI, on the other hand, aims at explicitly representing paradigm shape (ir-)regularities but does not yet include a complete form realisation apparatus. The third objective of the visit therefore was to investigate whether and how features available within the PARSLI model, such as the notion of *inflection zones*, could be useful to extend the *Network Morphology* framework so as to allow it greater expressive power w.r.t. the representation of canonicity. In return, combining PARSLI with the expressive power of DATR should constitute a useful extension of PARSLI,

[1] These stem zones are similar to the *partitions* in [Pirelli and Battista, 2000] or *stem spaces* in [Bonami and Boyé, 2003].

hence allowing it to also include a formal apparatus for the implentation of form-realisation proper.

4. Finally, I share with members of the Surrey Morphology Group an interest in the evaluation of morphological complexity. However, we have been approaching the problem from different angles. The approaches developed within the Surrey Morphology Group, be they qualitative [Baerman, 2011] or quantitative [Brown and Evans, 2010], focus on directly comparing the complexity of different languages, while I have been conducting experiments on comparative complexity of multiple descriptions of one given language, in terms of description compactedness. These experiments have compared ᑭᗩᖇᔕᒪᓀ-based descriptions to other competing descriptions using information theoretic measures [Sagot and Walther, 2011, Walther and Sagot, 2011], and have shown ᑭᗩᖇᔕᒪᓀ's relevance for approaching optimal compactedness levels. The leading idea of this research is that the comparison of specific description schemes is a necessary first step before meaningful cross-linguistic comparison can be devised. The final goal of this project is to further the comparison of the different quantitative approaches to morphological complexity.

2 Description of the work carried out during the visit

2.1 Typological expertise

The typological expertise available at the SMG has been extremely valuable to the improvement of the ᑭᗩᖇᔕᒪᓀ model. In particular, discussions with Greville G. Corbett have resulted in the development of an independent morphosyntactic feature hierarchy, orthogonal to ᑭᗩᖇᔕᒪᓀ's other features. These discussions have also triggered the a formalisation of the notion of morphomic splits [Corbett, 2010] in the ᑭᗩᖇᔕᒪᓀ model, in order to allow a more explicit distinction between motivated and unmotivated morphomic structures [Walther, 2012b].

The soundness of the ᑭᗩᖇᔕᒪᓀ model has also vastly benefitted from general discussions about the theoretical aim underlying the development of this new model.

2.2 Confrontation with real data

Given the different types of work currently being carried out, we have been able to test the ᑭᗩᖇᔕᒪᓀ model on the inflectional systems of new languages.

In particular, a formalisation of Maris Camilleri's description of Maltese verbs built from semitic bases [Camilleri, 2011, Camilleri, 2012] has been developed using ᑭᗩᖇᔕᒪᓀ notions. This formalisation is already associated with an almost complete implementation within the Alexina framework. Formalising and implementing the descriptions of[Camilleri, 2011] and [Camilleri, 2012] has allowed us to uncover generalisations that had been missed in these descriptions before and has therefore both completed and improved the account given on Maltes verbal inflection.

This implementation of Maltese verbal inflection consitutes the first large-scale implemented formalisation of a non-concatenative language using the ᑭᗩᖇᔕᒪᓀ model. Whereas the ᑭᗩᖇᔕᒪᓀ model has proven appropriate for formalising descriptions of semitic verbal inflection, the implementation has resulted in modifications of the Alexina framework. At this stage, the implementation provides 6 754 correctly inflected (validated) forms for 450 verb lemmas based on CVCVC stems. The formalisations of

the other (less numerous) semitic verb forms have also been done; their implementation is ongoing. In parallel, new, competing, analyses, also using ParSli, are being developped. This ongoing work is part of a new, already fairly underway, collaboration (see below).

Another starting collaboration with Enrique Palancar concerns the verbal inflection of the Oto-Manguean language Tlatepuzco Chinantec. We started developping a formalisation and implementation of the verbal inflection of Tlatepuzco Chinantec based on a thorough description of 790 verbs (collected from [Merrifield and Anderson, 2007]) by Enrique Palancar. Tlatepuzco Chinantec is remarkable for its intricate system of multiple exponence used for expressing verbal inflection. Moreover, since interaction of diverse exponents (tone change, ablaut changes and variable stress patterns) is observable on the verbal stems, this work should help improve and clarify the way ParSli deals with phenomena of multiple exponence and in particular on multiple exponence affecting the stems. It should also result in modifications of the Alexina framework which so far, does not contain a specific encoding for multiple exponence.

Other inflectional systems, such as Russian and Polish nouns, have been studied, but have not found their way into a complete implemented ParSli description yet. However, the interaction between inflectional classes and stress patterns in Russian [Brown *et al.*, 1996, Brown and Hippisley, 2012] nouns and the feature splits occuring in subgenders for Polish nouns [Brown, 1998, Brown and Hippisley, 2012] have been identified as interesting test data for the ParSli model.

2.3 Computational issues regarding ParSli and *Network Morphology*

Comparisons between *Network Morphology* and ParSli and a more in depth study of DATR have shown that an implementation of a morphological description as formalised within ParSli can be achieved using DATR. We have observed that the hierarchy system underlying DATR and vastly used in *Network Morphology* can be adapted to handle ParSli notions. As a result, ParSli can now be implemented within the DATR language, thereby benefitting from another complete form realisation apparatus more commonly used within the morphological community.

In order to facilitate the transfer between Alexina and DATR implementations of ParSli formalised descriptions, we have started the development of an automatic converter taking a large scale implementation of a ParSli description within Alexina as an input and outputting a DATR implementation of the same description. DATR descriptions are usually manually built and do not make use of large-scale resources. The new DATR implementation thus takes advantage of the large-scale property of Alexina lexica for which automatic enrichement tools exist [Sagot, 2005, Sagot, 2007]. The advice given by Dunstan P. Brown as well as his help with the improvement of the DATR output have been extremely valuable. At this stage, the converter has been finished but for a few details in the generated DATR files. This work should be completed shortly.

The building of the converter was not part of the initial aims of the visit. Observations on DATR and Alexina implementations have however naturally led towards its development. Unfortunately this additional task has taken up time, so that it has not be possible so far to initiate the work consisting in including ParSli notions into *Network Morphology*.

2.4 Reflections on morphological complexity

During the first two weeks of the visit, I have been able to attend the workshop on morphological complexity organised by the SMG at the British Academy in London (January 13th to 15th), and to present my own work on morphological complexity at the Surrey Linguistic Circle on January 17th [Walther, 2012a]. The presentation has been followed by many fruitful discussions resulting in the intent to further collaborate on these matters. One concrete outcome of these discussions is the organisation of a workshop on computational approaches to morphological complexity, to be held in October 2012. Aside from that, regular meetings on computational complexity issues regarding morphology are intended to take place, also involving Benoît Sagot (Alpage, INRIA & Paris 7) and possibly other regular collaborators of Dunstan P. Brown's.

Moreover, as a side-result of above mentioned development of various formalised descriptions of Maltese verbal inflection developed within PARSLI and implemented in Alexina, new experiments should soon be possible on competing formalised descriptions.

3 List of future collaborative projects and publications

The visit has paved the way for a number of informal collaborations, some of which have aleady started. Four main projects have emerged, presented in this document according to the amount of work already initiated.

3.1 Maltese verb morphology

Participants: *Maris Camilleri (University of Surrey) and Géraldine Walther (LLF: CNRS & Paris 7)*

Status: Ongoing.

Content: Formal representation of Maltese verb morphology within the PARSLI model; development of a lexical ressource for Maltese verbs.

Expected outcomes: An electronic lexical resource for Maltese verbs; new exhaustive analyses with complete implementations; one journal paper in preparation.

3.2 Electronic Lexical Ressources for Chinantec Verb Morphology

Participants: *Enrique Palancar (University of Surrey) and Géraldine Walther (LLF: CNRS & Paris 7)*

Status: Starting.

Content: Formal representation of Chinantec verb morphology within the PARSLI model; development of lexical ressources for Chinantec languages.

Expected outcomes: Electronic lexical resources for Tlatepuzco Chinantec verbs; possibly similar resources for other Chinantec languages and joint publications.

The verb lexicon for Tlatepuzco is possibly to be completed with comparable resources for other Chinantec languages to be used for comparative purposes. These lexica should be useful as resources within the ES/I029621/1 project "Endangered

Complexity: Inflectional classes in the Oto-Manguean languages" (Principal Investigators: Dunstan P. Brow, Matthew Baerman, Greville G. Corbett; Researcher: Enrique Palancar).

Moreover, possible further collaborations resulting from this joint work might include an investigation of the canonicity in Chinantec verb paradigms as defined in the framework of Canonical Typology and formalised and measured with PARSLI. If carried out, this work is likely to result in a joint publication on the formalisation of Chinantec verbal inflection focusing on the (non-)canonicity of their paradigms.

3.3 Computational Issues in the Implementation of Morphological Descriptions

Participants: *Dunstan P. Brown (University of Surrey) and Géraldine Walther (LLF: CNRS & Paris 7)*

Status: Regular ongoing, yet informal, discussions.

Content: Comparison of Alexina and DATR implementations.

Expected outcomes: A converter from Alexina to DATR (almost finished) and a converter from DATR to Alexina; continuing discusions regarding theoretical and computational issues related to the notion of default.

3.4 Quantitative Approaches of Morphological Complexity

Participants: *Dunstan P. Brown (University of Surrey), Benoît Sagot (Alpage: INRIA & Paris 7) and Géraldine Walther (LLF: CNRS & Paris 7)*

Status: Occasional discussions, regular meetings intended.

Content: Comparison of various quantitative approaches to morphological complexity.

Expected outcomes: Organisation of a workshop in October 2012; joint publications intended, but no specific work in progress yet.

References

[Baerman, 2011] Baerman, Matthew. 2011. Defectiveness and homophony avoidance. *Journal of Linguistics*, **47**, 1–29.

[Baerman *et al.*, 2010] Baerman, Matthew, Greville G. Corbett, and Dunstan Brown. 2010. *Defective Paradigms*. Oxford, UK: Oxford University Press. Proceedings of the British Academy 145.

[Baerman *et al.*, 2007] M. Baerman, G. G. Corbett, D. Brown and A. Hippisley, Eds. 2007. *Deponency and Morphological Mismatches*. Oxford University Press.

[Bonami and Boyé, 2003] Bonami, Olivier and Gilles Boyé. 2003. Supplétion et classes flexionnelles dans la conjugaison du français. *Langages*, **152**, 102–126.

[Bonami and Boyé, 2006] Bonami, Olivier and Gilles Boyé. 2006. Deriving inflectional irregularity. In *Proceedings of the 13th International Connference on HPSG*, p. 39–59, Stanford, USA: CSLI Publications.

[Boyé, 2006] Boyé, Gilles. 2006. Suppletion. In K. Brown, Ed., *Encyclopedia of Language and Linguistics (2nd ed.)*, volume 12, p. 297–299. Oxford, Royaume Uni: Elsevier.

[Brown *et al.*, 1996] Brown, Dunstan, Greville G. Corbett, Norman Fraser, Andrew Hippisley, and Alan Timberlake. 1996. Russian noun stress and network morphology. *Linguistics*.

[Brown and Hippisley, 2012] Brown, Dunstan and Andrew Hippisley. 2012. *Network Morphology: A Defaults-based Theory of Word Structure*. Cambridge University Press.

[Brown, 1998] Brown, Dunstan P.. 1998. Defining 'subgender': virile and devirilised nouns in polish. *Lingua*, **104**, 187–233.

[Brown and Evans, 2010] Brown, Dunstan P. and Robert Evans. 2010. Inflectional defaults and principal parts: an empirical investigation. In S. Müller, Ed., *CSLI Proceedings of the HPSG10 Conference*, p. 234–254, Université Paris Diderot, Paris 7, France. 17th International Conference on Head-Driven Phrase Structure Grammar HPSG10.

[Camilleri, 2011] Camilleri, Maris. 2011. Island morphology: Morphology's interactions in the study of stem patterns. *Linguistica*, **51**, 65–84. Internal and External Boundaries of Morphology.

[Camilleri, 2012] Camilleri, Maris. 2012. Stem patterns across maltese verbal paradigms. Communication at the 15th International Morphology Meeting (IMM15). January 9th-12th 2012. Vienna, Austria.

[Corbett, 2003] Corbett, Greville G.. 2003. Agreement: the range of the phenomenon and the principles of the surrey database of agreement. *Transactions of the philological society*, **101**, 155–202.

[Corbett, 2010] Corbett, Greville G.. 2010. Morphomic splits. Presented at the workshop 'Perspectives on the Morphome'. University of Coimbra, 29-30 October 2010.

[Merrifield and Anderson, 2007] Merrifield, William R. and Alfred E. Anderson. 2007. *Diccionario Chinanteco de la diáspora del pueblo antiguo de San Pedro Tlatepuzco, Oaxaca*. Serie de vocabularios y diccionarios indígenas "Mariano Silva y Aceves". Mexico DF: Summer Linguistic Institute., 2nd edition edition. Accessible at: `http://www.sil.org/mexico/chinanteca/tlatepuzco/S039a-DiccChinTlatepuzco-cpa.htm`.

[Pirelli and Battista, 2000] Pirelli, Vito and Marco Battista. 2000. The paradigmatic dimension of stem allomorphy in italian verb inflection. *Italian Journal of Linguistics*, p. 307–380.

[Sagot, 2005] Sagot, Benoît. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 ((c) Springer-Verlag), Proceedings of TSD'05*, p. 156–163, Karlovy Vary, Czech Republic.

[Sagot, 2007] Sagot, Benoît. 2007. Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Proceedings of the 3rd Language & Technology Conference (LTC'05)*, p. 423–427, Poznań, Poland.

[Sagot, 2010] Sagot, Benoît. 2010. The Le*fff*, a freely available, accurate and large-coverage lexicon for French. In *Proceedings of the 7th Language Resource and Evaluation Conference*, Valletta, Malta.

[Sagot and Walther, 2010] Sagot, Benoît and Géraldine Walther. 2010. A morphological lexicon for the Persian language. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, Valletta, Malta.

[Sagot and Walther, 2011] Sagot, Benoît and Géraldine Walther. 2011. Non-canonical inflection : data, formalisation and complexity measures. *Communications in Computer and Information Science*, **100**. Systems and Frameworks for Computational Morphology.

[Stump, 2006] Stump, Gregory T.. 2006. Heteroclisis and paradigm linkage. *Language*, **82**, 279–322.

[Thornton, 2010] Thornton, Anna M.. 2010. Towards a typology of overabundance. Presented at the Décembrettes 7, Toulouse, France.

[Walther, 2011] Walther, Géraldine. 2011. Measuring morphological canonicity. *Linguistica*, **51**, 157–179. Internal and External Boundaries of Morphology.

[Walther, 2012a] Walther, Géraldine. 2012a. Assessing description compactedness : A comparative study of 4 competing formalisations for french verbal inflection. Communication ar the Surrey Linguistic Circle, University of Surrey, Guildford, UK, January 17th 2012.

[Walther, 2012b] Walther, Géraldine. 2012b. Formally encoding morphomicity. Poster presentation at the 15th International Morphology Meeting (IMM15). January 9th-12th 2012. Vienna, Austria.

[Walther and Sagot, 2010] Walther, Géraldine and Benoît Sagot. 2010. Developing a Large-Scale Lexicon for a Less-Resourced Language: General Methodology and Preliminary Experiments on Sorani Kurdish. In *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*, Valetta, Malta.

[Walther and Sagot, 2011] Walther, Géraldine and Benoît Sagot. 2011. Modélisation et implémentation de phénomènes non-canoniques. *Revue TAL*, **52**(2/2011). Vers la morphologie et au-delà.

[Walther *et al.*, 2010] Walther, Géraldine, Benoît Sagot, and Karen Fort. 2010. Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish. In *Proceedings of the 29th International Conference on Lexis and Grammar*, Belgrad, Serbia.