CONFERENCE REPORTS

EACL 2009 Workshop on Computational Approaches to Semitic Languages, Athens, Greece, 31 March 2009

WIDO VAN PEURSEN AND CONSTANTIJN SIKKEL PESHITTA INSTITUTE LEIDEN, TURGAMA PROJECT

In combination with the Twelfth Conference of the European Chapter of the Association for Computational Linguistics (ACL), Athens, 30 March – 3 April 2009, the ACL Special Interest Group for Computational Approaches to Semitic Languages organized a one-day workshop. The various contributions to this workshop gave a state-of-the art impression of work that is going on in the application of Computational Linguistics to various Semitic languages.[1]

Four papers dealt with morphology. Semitic languages, with their rich morphology, provide interesting material for computational morphological analysis. In the study of Semitic morphology, as in many other areas of computational linguistics, ample use is made of finite state automata. There are several advantages to the use of finite state automata (FSAs). They can handle a number of phenomena relatively easily, they are broadly available, and they have been studied intensively and tested and improved accordingly. However, some typical features of Semitic languages pose serious challenges to the application of FSAs. A FSA reads an input tape in a linear way, and goes through it character by character, from the beginning to the end. This approach assumes that morphemes are built up of concatenations of characters and that words are built up of concatenations of morphemes and that morphological analysis concerns the segmentation of words into morphemes. This approach works well for languages with complex concatenative morphological patterns, such as Turkish, but is not very suitable for languages that have typical non-concatenative features, including the Semitic languages. For example, the Syriac verb form *qabbel*, 'he received', is a combination of the pattern C_1-a-C_2-C_2-e-C_3 and the root QBL. In this example the verbal stem Pael is realized by the gemination of the second root letter and the a-e vowel pattern,

[1] The workshop proceedings can be found at http://staff.um.edu.mt/mros1/casl09.

rather than by a morpheme that can be isolated in a segmentation process.

In the workshop various papers addressed the question of how to deal with these non-linear phenomena when using FSAs. They elaborated upon the work that has been done over the last years. One of the pioneers in this field is the general editor of Hugoye, George A. Kiraz.[2] In a number of publications, including his monograph *Computational Nonlinear Morphology*,[3] Kiraz has developed a model that deals with the root-template pattern by allotting multiple grammatical/lexical layers to separate tapes.[4]

Michael Gasser gave a presentation of his work on Finite State morphology for Ethiopic Semitic languages, especially Amharic and Tigrinya. These languages pose a number of additional challenges compared to other Semitic languages. Due to prefixes and suffixes that function as negation, relativizers, accusative markers, prepositions, conjunctions, and sometimes even auxiliaries, the verb forms can receive even more complex morphological structures than those known from other Semitic languages. Other distinctive features of these languages are that gemination is not only grammatical but also lexical and that some roots take a vowel, rather than only consonants. Gasser showed how FSAs can be applied to these languages in an effective way by augmenting the transitions between the states with feature-weights. The feature-weight constraints make it possible to handle the phenomenon that the occurrence of a certain element in a word depends on the occurrence of an element later on in the word, that is, to describe long-distance dependencies.

Mans Hulden presented a method by which multi-tape automata, used to build analysers for root-and-pattern morphology,

---

[2] Other scholars who have worked on the application of FSAs to Semitic languages are Kenneth R. Beesley, Lauri Karttunen and Shuly Wintner (see also note 4).

[3] George Anton Kiraz, *Computational Nonlinear Morphology. With Emphasis on Semitic Languages* (Studies in Natural Language Processing. Cambridge: Cambridge University Press 2001).

[4] See also his article 'Multitiered Nonlinear Morphology Using Multitape Finite Automata: A Case Study in Syriac and Arabic', *Computational Linguistics* 26/1 (2000) 77–105. For the application of Finite State Automata to Modern Hebrew see Yael Cohen-Sygal and Shuly Wintner, 'Finite-State Registered Automata for Non-Concatenative Morphology', *Computational Linguistics* 31/1 (2006) 49–82. The journal *Computational Linguistics* is now online available at http://www.mitpressjournals/org/loi/coli.

can be simulated using standard finite-state methods and toolkits. This has the advantage that one can make use of the existing tools for single-tape automata or finite-state transducers. He tested his approach on a limited implementation of Arabic verb morphology.

François Barthélemy presented Karamel, which is a system for the development of morphological descriptions compiled in finite-state machines. In this system embedded units are employed for obtaining tree structures expressing the relationships between tapes. The tree-structure is also useful to define the scope of the feature structures which are used as an abstract representation of the analysed forms. He gave a demonstration of how this system can be applied to the Akkadian verb.

Although the approaches of Gasser, Hulden and Barthélemy differ in details, their common aim is to deal with the non-concatenative structure of Semitic languages within a Finite State framework. Progress is made in the field, although some linguistic phenomena that abound in Semitic languages, such as reduplication and metathesis, still cannot be dealt with in this framework in a satisfying way.

Our own presentation concerned some of the problems we have to face in the linguistic and philological analysis of the Peshitta.[5] We addressed the question of how we can establish a verbal paradigm on the basis of ancient Syriac manuscripts. Taking the so-called orthographic variants in the Leiden Peshitta edition as our point of departure, we discussed the question of how we should deal with forms that do not agree with the paradigm that is found in the traditional grammars (e.g., the perfect 3rd person masculine singular with a Waw, i.e. QBLW 'he (!) received'), and how we can infer the paradigm from the sources, rather than imposing the paradigm found in the grammars upon the sources (which happens if we take the exceptions as 'orthographic' variants that are even not worthy of being mentioned in the critical apparatus of the edition).

Our inductive approach—trying to deduce the grammar from the extant textual witnesses—and the rule-based approaches used in a Finite State framework presented in the other papers—which usually aimed at the development of efficient parsers or generators—show an interesting complementarity. In the study of Syriac we have on the one hand the classical philological questions that require an inductive approach—in which grammatical forms and

[5] For more details about our research see the Turgama project website at http://www.hum.leiden.edu/religion/research/research-programmes/turgama.jsp.

functions cannot be taken for granted, but have to be established on the basis of corpus analysis—and on the other hand such a large number of texts that automatic or semi-automatic parsing processes will be rewarding in the long term.[6]

Another paper dealing with morphology was presented by Khaled Shaalan, Hitham M. Abo Bakr and Ibrahim Ziedan. Their contribution concerned the automatic vocalization of texts in Modern Standard Arabic. Since the distribution of vowels is partly lexically determined (e.g., the stem-vowels of a certain word) and partly context-determined (e.g., the case ending of a word depending on its syntactic function), they proposed a hybrid approach in which lexical and contextual information are combined.

For specialists in the linguistic and philological study of ancient Semitic texts, such as the authors of this report, the morphological analysis is the most familiar application of computer science in Semitic studies. However, the workshop contained other fascinating contributions dealing with concept discovery, information retrieval, speech technology and machine translation.

Elad Dinur, Dmitry Davidov and Ari Rappoport presented a paper on concept discovery. A concept is defined as a group of words that share a significant aspect of their meaning. Thus the words 'dog', 'puppy' and 'cat' belong to the concept of 'pets'. There are various indications that words possibly belong to the same concept, such as the 'X and Y' collocation ('cats and dogs'). However, the discovery of such collocations in Semitic languages is complicated by the Semitic word structure in which the conjunction and other elements such as prepositions are written together with the next word. Thus to discover the pair 'cats and dogs' in a Modern Hebrew phrase 'the-cat and-the-dog' or 'for-cats and-for-dogs', one first has to segment the words into their various parts and to identify the conjunction, the articles, and the one-letter prepositions before one can isolate the two nouns 'cat' and 'dog'.

Lamia Tounsi, Mohammed Attia and Josef van Genabith presented a paper on the automatic acquisition of grammatical resources based on Lexical Functional Grammar (LFG). They showed how the methodology of automatic treebank-based acquisition of LFG dependency structures can be adapted and employed in a way that accounts for the morphological complexity and syntactic structures of Arabic.

---

[6] Initiatives to develop instruments for automatic POS tagging are being developed at the Center for the Preservation of Ancient Religious Texts at the Brigham Young University, see http://cpart.byu.edu.

Fadi Biadsy, Julia Hirschberg and Nizar Habash held a paper on the use of phonotactic modelling for the dialect identification of spoken Arabic. They presented a model that was able to identify speakers of Egyptian, Levantine, Gulf and Iraqi Arabic dialects.

Lahsen Abouenour, Karim Bouzoubaa and Paolo Rosso discussed the challenges of Arabic Information Retrieval systems, focusing on Question/Answer systems, going from question analysis, through passage retrieval, to answer extraction. For example, to receive an answer to the question 'Where is the city of Marrakech?', the recognition of the keywords 'city' and 'Marrakech' is not sufficient, because a passage that tells that Marrakech is a city does not answer the question. Hence additional morphological and semantic information is needed to identify the correct answer.

Jakob Elming and Nizar Habash addressed the question of how Phrase-based Statistical Machine Translation (PSMT) can be applied to translations from English to Arabic. Since PSMT uses local experience, it is locally strong. But the reordering that is needed to translate English sentences into Arabic (e.g., the verb has to be moved to the first position in the clause) and the non-local conditions that determine these reorganizations (e.g., in a relative clause the verb does not take the initial position) requires considerable syntactic reordering.

We can conclude that fascinating and exciting developments are taking place in the computational analysis of Semitic texts and languages. The wide range of Semitic studies, from second-millennium BCE Akkadian cuneiform tablets to the variations of modern languages and dialects spoken by millions of people is reflected in the wide range of computational approaches to Semitic languages, from detailed philological analysis to speech recognition and machine translation. The workshop gave a valuable overview of these developments.