Linguistics **I**ssues in **L**anguage **T**echnology – LiLT

Theoretical and Computational Morphology: New Trends and Synergies

Edited by Bruno Cartoni, Delphine Bernhard and Delphine Tribout

CENTER FOR THE STUDY OF LANGUAGE AND INFORMATION

Contents

Introduction

Nowadays, theoretical morphology witnesses a revival due to the emergence of new formalization frameworks, both for inflectional and lexical morphology: canonical morphology, approaches based on analogy, rule-based approaches, to name just a few. All these approaches can potentially shed new light on computational processing of morphology, either for parsing or generation. In parallel, new computational techniques (such as (un)supervised morphological acquisition) and formal frameworks bring a fresh look on morphological phenomena. Both domains tackle more or less implicitly the organization of the lexicon in general

The TACMO workshop, organised within the International Congress of Linguists(held in Geneva in July 2013) aimed at gathering these two facets of morphology – computational and theoretical formalization – in order to foster interactions and to highlight how both approaches benefit from each other.

After a call for abstracts in Summer 2012, we received eighteen proposals that have been evaluated by the Program committee. Eight of them were selected, as representing an appropriate range of the synergies targeted by the workshop.

The present volume gathers long versions of six of these research papers. They address a wide variety of issues, concerning inflection, derivation and compounding, and covering theoretical questions (polyfunctionality, organisation of inflection classes) and computational questions (measuring the complexity of morphology). Methodological questions (acquisition of morphological data in corpora, etc.) and representation and documentation issues of morphological information

are also addressed.

This volume is organized in three sections.

The first section is composed of two articles that present computational methods which help enhance morphological description and the understanding of morphological systems. In his paper "What is grammar like? A usage-based constructionist perspective", V. Kapatsinski presents the relations that hold between usage-based linguistics and computational models for morphology. He claims that the language acquisition model developed within this linguistic approach are suitable for computational morphology model. Supported by various experimental observations, both in computational linguistics and psycholinguistics, he argues that usage-based approaches are more appropriate to account for the various aspects of grammar than generative linguistics could do. The large overview provides an interesting starting point for a future debate on the need for a grammar model that can account for the internal grammar of a linguistic community.

On the other hand, Katharina Ehret's article "Kolmogorov complexity of morphs and constructions in English" focuses on the measurement of morphological and syntactical complexity in English. In an English mixed-genre corpus made of samples of literature, gospel and newspapers, the author measures the contribution of 5 different morphs and 5 different syntactic constructions to the complexity of the language, using the gzip compression programme as an approximation of the Kolmogorov complexity. Through targeted manipulations of the corpus with respect to the 5 morphs and constructions under study, she can assess their contribution to the global complexity, both at the morphological and the syntactical level by distorting the texts morphologically and syntactically and compressing them with gzip. The compression ratio between the distorted text and the original indicates the morphological (or syntactical) complexity, and reveals how much the manipulated morph or construction contributes to it.

Through these experiments, the author shows how compression algorithms can be used in order to measure the contribution of different features to the morphological and the syntactical complexity in English.

The second section focuses on inflectional morphology. In his article "Polyfunctionality and infectional economy" Gregory Stump analyses the polyfunctionality of exponents within and across paradigms in different languages, polyfunctionality being the use of one exponent in order to express different morphosyntactic properties. The formal and theoretical framework of Paradigm Function Morphology allows the

author to postulate three different types of polyfunctionality. The first type is when one exponent is used to express related functions in more than one affix position, as illustrated by the negative prefix ha- within verb inflection in Swahili. In the second type, one exponent is used to express different functions within one lexeme's paradigm, as can be found within the paradigm of latin adjectives. The third type is the case when different categories use the same set of exponents in order to express different but related morphosyntactic properties, as shown by noun and verb inflection in Khanty. G. Stump provides a formal analysis accounting for each type of polyfunctionality, and further concludes that polyfunctionality contributes to the economy of an inflectional system.

The article "Semi-separate exponence in cumulative paradigms. Information-theoretic properties exemplified by Ancient Greek verb endings" by Paolo Milizia looks at a theoretical issue with quantitative information extracted from a corpus. Two concurrent interpretations may be given for Ancient Greek thematic imperfect endings: the fully cumulative interpretation and the semi-separate interpretation, which analyses some endings as made of two morphs. The aim of the study is to show how frequency information derived from a corpus (Ancient Greek Dependency Treebank) can help choose one interpretation. The main observation is that semi-separate exponence corresponds to low frequency paradigm cells which leads the author to formulate an hypothesis according to which the inflectional system tends to favor equiprobable exponents In order to verify this hypothesis, he relies on information-theoretic properties and shows that syncretism and semi-separate exponence lead to a more equiprobable distribution of exponents. Several cross-linguistic examples are detailed to support this claim: non-singular numbers of the Ancient Greek verb endings, dual number of the personal pronoun of classical Arabic, conjugation of transitive verbs in Nganasan. P. Milizia also describes and accounts for some contradictory examples which could give the impression that the principle of equilibrium in morphological encoding should be rejected. He also analyses a diachronic example (Proto-Indo-European to Ancient Greek verb endings) which is consistent with his proposals.

The last section tackles another important facet of morphological research that also witnesses important renewal in the past years: derivational morphology. The first paper of this section, "Démonette, a French derivational morphosemantic network" by Nabil Hathout and Fiammetta Namer, describes the creation of a large lexical resource based on the principle of derivational morphology.

They present Démonette, a derivational morpho-semantic network for the French language, that contains morphological and semantic information about lexemes.
The Démonette lexicon is built by merging Dérif, a morphosemantic analysis tool, and Morphonette, a morphological network. This new resource, anchored in the Word-based theoretical framework of morphology, follows a formal network architecture in which morphological relations are established between base and derivatives and between indirectly related words. The framework also contains semantic information: the lexemes are labelled with semantic types, and links between related lexemes are labeled with concrete and abstract bi-oriented definitions.
The proposed resource offers two important advantages. First, it provides a large-scale database of evidence within the same theoretical framework of word-base morphology. Second, it is highly profitable for many NLP application that can take advantage of the semantic and morphological information encoded in this resource.
In their article "Evaluative prefixes in translation: From automatic alignment to semantic categorization", Marie-Aude Lefer and Natalia Grabar use evidence from translations into English to study the semantics of French evaluative prefixes. These prefixes can convey information about quantity (big / small) or about quality (good / bad). The aim of Lefer and Grabars study is to provide new insights on French evaluative prefixes by looking at their occurrences in a French-English parallel corpus. The assumption here is that translations into English – and in particular periphrastic translations – will help disambiguate the meaning of the prefixes and thus allow their semantic categorization. In a first step, sentences containing evaluative prefixes are detected and extracted with their translation into English. Then, French prefixed words are aligned with their translation into English. Two different alignment methods are compared: GIZA++ and a tailor-made program relying on several heuristics. The latter has a better recall and was thus used in the experiments, after manual validation of the alignments. The prefix sur- is studied in more detail, as it is the most frequent evaluative prefix in the corpus. This makes it possible to identify recurrent periphrastic constructions for sur- in its excess meaning, which can in turn be used to disambiguate the meaning of less frequent prefixes. This corpus study helps refine categorizations proposed in previous research and shows how morphology, contrastive linguistics, translation studies and NLP can be combined in a constructive manner. While focusing on a rather precise issue, the article describes a methodology which we believe can be used in other contexts, in order to perform

translation-based semantic analyses of morphological phenomena.

Interestingly, most of these research papers make extensive use of corpus data, either to support their claims or to uncover more evidences of certains morphological phenomena. Some of them make even use of multilingual data. Even if some of these research papers are more focused on the elaboration or the improvement of a theoretical frameworks, all of them have in common a corpus-based approach, and for some of them, the objective of building morphological resources that will provide large-scale descriptions of morphological phenomena. All these aspects reflect largely the current trends in computational and theoretical morphology.

Scientific committee of the Edited Volume

Acknowledgments

We would like to thank the local committee of the "International Congress of Linguists" (ICL) for their help with the organisation of the workshop on "Theoretical and Computational Morphology". We also would like to thank the members of the scientific committee for their reviewing work.