Athanasios N. Karasimos. *Computational Processing of Allomorphy in word derivation of Modern Greek.* University of Patras, 2011.

[akarasimos@upatras.gr, akarasimos@gmail.com]

Keywords

allomorphy, derivation, constraints, (un)supervised morphology learning, maximum entropy, finite-state automata, AlloMantIS

1 Introduction

Allomorphy, as defined by Lieber (1982: 27), is the study of morpheme variants which share such lexical information as semantic representation and argument structure but which differ unpredictably and arbitrarily in the phonological forms and in the morphological environments in which they occur (e.g., *kima* – *kimat* ‘wave_N’, *pern* – *pir* – *par* ‘take’). This dissertation presents a systematic theoretical view of analysis and processing of stem and affixal allomorphy in derivation. Here, it is also shown successfully that it is possible to treat and computationally parse allomorphy in Greek derivation.

Research on stem allomorphy was renewed by Aronoff (1994), whose work led to novel approaches to inflectional and derivational phenomena in morphological research by Booij (1997), Thornton (1997), Pirrelli & Battista (2000a, 2000b), Ralli (2000, 2007), Stump (2001), and Maiden (2004), among others. Aronoff’s main position, also followed by other morphologists, is that the signans (the formal side) of a lexeme is not a single phonological representation, but an array of indexed stems, which may stand in relations to one another ranging from identity through regular phonological alternation or arbitrary change to full suppletion.

On the other hand, there is a strong tendency among phonologists, more specifically by supporters of *Optimality Theory* (OT), to describe and analyze allomorphy as a (morpho-)phonological phenomenon and to treat it as such (see Drachman, Kager & Malikouti-Drachman 1995, Mascarò 1996, 2007, Thornton 1997, Drachman 2006, *inter alia*). In phonological research couched in OT, all arbitrary and predictable variations of morphemes are treated primarily with phonological criteria. Usually the selection of the appropriate form of morphemes depends on the different properties of the syllables of the material, the position of a tone or stress, and the phonological composition of suffixes. For example, in French, the adjective *beau* ‘beautiful’

has two allomorphs (*beau ~ bel*); when it is followed by a word with an initial vowel, the second allomorph (*bel*), as *bel ami* (**beau ami*) is chosen, because the selection is based on phonological properties of specific syllables (Mascaro 2007).

In Computational Morphology, there have only been a few attempts to investigate the issue of allomorphy (Krieger, Pinker & Nerbonne 1993, Goldsmith 2001, Handl et al. 2009, Kohonen, Virpioja & Klami 2009, Virpioja & Kohonen 2009). Moreover, without avoiding the pitfalls of main theoretical problems, researchers usually deal with the phonological alternations of morphemes, and often they exclude groups of allomorphs and morphemes from their analysis due to the absence of a proper treatment. In this dissertation, we combine the advantages of each computational model (finite state automata and (un)supervised morphology learning) for the immediate and effective processing of allomorphy. Any effort to implement a computational model has some disadvantages, but we try to overcome these with similar and related strategies.

The aim of the dissertation is to cover the linguistic phenomenon of allomorphy both on a synchronic theoretical level as well as on a computational level of realistic and practical tactics for computational processing. More specifically, I define allomorphy, set its boundaries, and analyze the behavior of allomorphs. Theoretically, I deal with cases of pure allomorphy, excluding the ones which do not comply with the definition of Lieber (1982). Similarly, on the computational level, I propose a feasible solution of a structured model (see Koskenniemi 1983, Sgarpas 1996, Markopoulos 1998) and strategies based on machine learning of morphology (Jurafsky & Martin 2000, Goldsmith 2001, Schone & Jurafsky 2001, Roark & Sproat 2007).

2 The Study

2.1 Theoretical Approach of Allomorphy

This study is couched in a theoretical framework centered on the morpheme, treating allomorphy as a morphological phenomenon which places derivation at a separate level of the word formation process. Therefore, I adopt here Ralli's framework (1994, 1999, 2000, 2005, 2008) and Lieber's (1982) theory of the Lexicon and allomorphy and thereby modify them. Allomorphs are defined with morphological criteria, and I introduce the term *phonomorph*, for phonologically-driven allomorphy. In addition, I classify allomorphy into various categories based on its variant forms. Moreover, I introduce the *rules*

of *allomorphy treatment*, which is a model of registry information (inflection, derivation, and compounding) for adjectival, nominal, and verbal allomorphs, given that the morpholexical rules in Lexicon (Lieber 1982) include only inflectional information for allomorphs. These rules are placed outside of the Lexicon, in the area of Grammar and assign to each allomorph the proper information of their morphological environments of appearance. Based on Greek data, I also define the *Principle of Allomorphic Behavior*, according to which all morphemes exhibit the same allomorphic behavior, i.e., the presence or absence of allomorphy in all word formation processes without dependence on any specific process (Karasimos 2009). Finally, I propose the *constraint of single allomorphy selection*, according to which a correlation between derived words with a common base is ensured by the participation of only one form of the lexeme-base. This constraint applies to all nominal and adjectival bases and to ex-contracted verbs without exceptions, usually satisfying the optimal syllabic structure (CV).

2.2 Computational Processing of Allomorphy

In order to build a concrete model for the computational processing of allomorphy, I conducted two experiments, one with *Linguistica*, an unsupervised morphology learning implementation by Goldsmith (2001) and one with modified *AMIS*, a maximum entropy supervised model technique (Berger, Della Pietra & Della Pietra 1996).

For the first experiment (*Linguistica*), three different corpora were employed, C1 (55,867 tokens) from the newspaper *Makedonia*, C2 (30,907 words) a target-made corpus, and C3 (281,821 tokens) from a personal sci-fi novel. The goal was to test this morphological parser in Greek and to see how capable it is of determining allomorphy. The results from C1 were quite disappointing (7.31 % accuracy). The results from C3 were slightly better, but the accuracy was still very low (11.49%). On the other hand, the results from C2 were more specific and clear, although the accuracy was also low (23.41%). The system managed to detect several inflectional paradigms, few derivational suffixes, and some bound stems. Additionally, only two types of allomorphy were recognized, only one of which was correct, i.e., the *karavi* ~ *karav*-type allomorphs. The results from the Greek corpora do not require a closer study, as the errors form the rule rather than the exception. These errors typically stem from the collapse of two or more suffixes into one, the systematic inclusion of stem-final material into a set of (spurious) suffixes, and the failure to segment all words actually containing the same stem in a consistent fashion, the most

important problem being the fact that *Linguistica* failed to correctly treat allomorphy.

For the second experiment (AMIS) I modified a syntactic – maximum entropy – parser to predict allomorphic changes of nominally derived words with the help of a training corpus. A training corpus of 4,677 inflected nouns was created, combining various characteristics, such as: inflectional class, allomorphic class, syllables, genres, stress, etc. This model was trained and gave as an output a wide array of positive or negative weights. Based on these weights, I built *AlloMantIS*, an AMIS analyzer for allomorphy prediction, which was tested on a testing corpus of 2,755 derived nominal words. In the first attempt, *AlloMantIS* managed to predict the allomorphic changes of 86.49% words, and in the second attempt it scored 91.43%. The reason for accuracy improvement was the reverse syllabification of words¹. It is important to point out that if the training corpus also included derived nominal words, the prediction accuracy would be more than 96%.

The results of the experiment led to the creation of a combined model of computational processing of allomorphy. The philosophy underpinning the proposed framework is to draw the basic layout for a computational treatment of allomorphy, inasmuch as it is an issue that commonly escapes the attention of researchers. The majority of approaches fail to confront allomorphy, and this study is inspired by Wicentowski's (2002, 2004) strategy, which shows that the combination of supervised and unsupervised models has comparatively better results in inflectional morphology than each approach, taken individually.

The main priority here is to create a computational combinational model which implements the basic features of the theory and exploits the results of the experiments. The proposed rules of allomorphic transformation are geared towards the implementation of morpholexical rules (redundancy rules), which computationally produce allomorphs from a basic morpheme; typically, these rules are very similar to inactive morphological and phonological rules. The final result of the rules is to associate allomorphs to a corresponding morphological environment via a morphological context filter, which assigns the proper information to each allomorph of a morpheme. Simultaneously, based on the results of the AMIS experiment, I introduce a predictive model of allomorphic “regularity,” which fairly accurately guesses the allomorphic behavior

¹ For my first attempt word syllabification was from left to right (e.g., *karavi* > Syl1_ka, Syl2_ra, Syl3_vi), and for the second attempt word syllabification was from right to left (e.g., Syl1_vi, Syl2_ra, Syl3_ka).

of derived words, both of stems and derived suffixes. Overall, the model predicts the allomorphy of a morpheme, the rules of allomorphic transformation producing all the forms of a morpheme, and it assigns to the morpheme morphological-context information.

3 Conclusion

Computational Morphology is a rapidly growing area of linguistics. Unsupervised and Supervised Morphology Learning models are a recent approach to morphological analysis problems. This dissertation investigates allomorphy, a morphological phenomenon which occurs with the same frequency in every word formation process in Greek. I present here a concrete theoretical analysis, which helped facilitate the processing of allomorphy computationally. On the whole, the results obtained from the two experiments provide additional support to build a combinational computational model of processing allomorphy. In other words, the results of a maximum entropy model provided evidence that it is possible to predict allomorphs, although their “alternations” are arbitrary and not produced by any morpho-phonological rule. Finally, it should be pointed out that only supervised morphology learning models with rules and imported human knowledge can serve as the basis for the computational treatment of the morphological phenomena of derivation in Modern Greek and morphologically rich languages.

References

- Aronoff, Mark. 1994. *Morphology by Itself: Stems and Inflectional Classes*. Linguistic Inquiry Monograph 22. Cambridge, MA: MIT Press.
- Berger, Adam, Stephen Della Pietra & Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1), 37-71.
- Booij, Geert. 1997. Allomorphy and the Autonomy of Morphology. *Folia Linguistica* XXXI (1-2):. 25-56.
- Drachman, Gaberell. 2005. A note on ‘shared’ allomorphs. *Journal of Greek Linguistics* 6, 5-38.
- Drachman, Gaberell, René Kager, & Angeliki Malikouti-Drachman. 1995. Greek Allomorphy: an Optimality Account. In M. Dimitrova-Vulchanova & L. Hellan (eds.), *Papers from the First Conference on Formal Approaches to South Slavic Languages*, 345-361.
- Goldsmith, John. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27(2): 153-198.
- Handl, Johannes, Besim Kabashi, Thomas Proisl & Carsten Weber. 2009. JSLIM - Computational Morphology in the Framework of the SLIM Theory of Language. In Cerstin Mahlow & Michael Pioteowski (eds.), *State of the Art in Computational Morphology*, 10-27. Berlin: Springer Verlag.

- Jurafsky, Daniel & James Martin. 2000. *Speech and Language Processing (An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition)*. Upper Saddle River, NJ: Prentice Hall Inc.
- Karasimos, Athanasios. 2009. Comments on Unsupervised Morphology Learning Model: The case of Greek Allomorphy. Paper presented at the 19th International Symposium in Theoretical and Applied Linguistics (ISTAL19). 3-5 April 2009, Thessaloniki.
- Kohonen, Oskar, Sami Virpioja, & Mikaela Klami. 2009. Allomorfessor: Towards Unsupervised Morpheme Analysis. In *Evaluating systems for Multilingual and Multimodal Information Access, Lecture Notes in Computer Science* 5706: 975-982. Springer Publications.
- Koskeniemi, Kimmo. 1983. *Two-level morphology: a general computational model for word-form recognition and production*. Ph.D. Thesis. Helsinki: University of Helsinki.
- Krieger, Hans-Ulrich, Hannes Pirker, & John Nerbonne. 1993. Feature-based Allomorphy. *Proceedings of the 31st annual meeting of the Association for Computational Linguistics*, pp. 140-147. Morristown, NJ: Association for Computational Linguistics. .
- Lieber, Rochelle. 1982. Allomorphy. *Linguistic Analysis*. 10(1): 27-52.
- Maiden, Martin. 2004. When lexemes become allomorphs – On the genesis of suppletion. *Folia Linguistica* XXXXCIII(3-4): 227-256.
- Mascarò, Joan. 1996. External allomorphy and contractions in Romance. *Probus* 8: 181-205.
- Mascarò, Joan. 2007. External allomorphy and lexical representation. *Linguistic Inquiry* 38(4): 715-735.
- Pirrelli, Vito, & Marco Battista. 2000a. The paradigmatic Dimension of Stem Allomorphy in Italian Verb Inflection. *Italian Journal of Linguistics (Rivista di Linguistica)* 12(II): 307-380.
- Pirrelli, Vito, & Marco Battista. 2000b. On the interaction of paradigmatic and syntactic stem alternation in Italian conjugation. *Acta Linguistica Hungarica* 47(1-4): 289-314.
- Ralli, Angela. 1994. Feature Representations and Feature-Passing Operations in Greek Nominal Inflection. *Proceedings of the 8th Symposium on English and Greek Linguistics*, pp. 19-46. Thessaloniki: English Department of Aristotle University of Thessaloniki.
- Ralli, Angela. 1999. Inflectional Features and the Morphological Module Hypothesis” In Athanasios Kakouriotes (ed.) *Working Papers in Theoretical and Applied Linguistics* 6, pp. 111-142. Thessaloniki: English Department of Aristotle University of Thessaloniki.
- Ralli, Angela. 2000. A Feature-based Analysis of Greek Nominal Inflection. *Glossologia* 11-12: 201-228.
- Ralli, Angela. 2007. On the Role of Allomorphy in Inflectional Morphology: Evidence from the Dialectal Varieties of Lesbos, Kydonies and Moschonisia. In G. Sica (ed.) *Open problems in Linguistics and Lexicography*. Milano: Polimetrika.
- Roark, Brian & Richard Sproat. 2007. *Computational Approaches to Morphology and Syntax*. Oxford: Oxford University Press.
- Schone, Patrick & Daniel Jurafsky. 2001. Language-independent Induction of Part of Speech Class Labels using only Language Universals. In *Proceedings of the Workshop “Text Learning: Beyond Supervision” of UCAI 2001*, pp. 53-60.
- Stump, Gregory. 2001. *Inflectional Morphology*. Cambridge: Cambridge University Press.
- Thornton, Anna-Maria. 1997. Stem allomorphs, suffix allomorphs, interfixes or different suffixes? On Italian derivatives with antesuffixal glides. In Geert Booij, Sergio Scalise & Angela Ralli (eds.) *Proceedings of the 1st Mediterranean Meeting of Morphology*, pp. 86-98.
- Virpioja, Sami & Oskar Kohonen. 2009. Unsupervised Morpheme Discovery with Allomorfessor. In *Working Notes for the CLEF 2009 Workshop*. Corfu, Greece.
- Wicentowski, Richard. 2002. *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework*. Ph.D. Thesis. Baltimore, MD: The Johns Hopkins University.

- Wicentowski, Richard. 2004. Multilingual noise-robust supervised morphological analysis using the WordFrame model. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*, pp. 70-77. Morristown, NJ: Association for Computational Linguistics.
- Μαρκόπουλος, Γεώργιος. 1998. *Υπολογιστική επεξεργασία του Νεοελληνικού Ονόματος*. Διδακτορική διατριβή. Αθήνα: Εθνικό Καποδιστριακό Πανεπιστήμιο.
- Ράλλη, Αγγελική. 2005. *Μορφολογία*. Αθήνα: εκδόσεις Πατάκη.
- Ράλλη, Αγγελική. 2008. *Η Σύνθεση λέξεων: διαγλωσσική μορφολογική προσέγγιση*. Αθήνα: εκδόσεις Πατάκη.
- Σγάρμπας, Κυριάκος. 1997. *Τεχνικές αυτόματης μορφοσυντακτικής ανάλυσης της Ελληνικής*. Διδακτορική διατριβή. Πάτρα: Πανεπιστήμιο Πατρών.