A crash test with Linguistica in Modern Greek: the cases of derivational suffixes and bound stems

Athanasios Karasimos, Evanthia Petropoulou
University of Patras
University Campus, Rion 26500, Greece
E-mail: akarasimos@upatras.gr, evapetro@cc.uoi.gr

Abstract
This paper attempts to participate in the ongoing discussion in search of a suitable model for the computational treatment of Greek morphology. Focusing on the unsupervised morphology learning technique, and particularly on the model of Linguistica by Goldsmith (2001), we attempt a computational treatment of specific word formation phenomena in Modern Greek (MG), such as suffixation and compounding with bound stems, through the use of various corpora. The inability of the system to accept any morphological rule as input, hence the term 'unsupervised', interferes to a great extent with its efficiency in parsing, especially in languages with rich morphology, such as MG, among others. Specifically, neither the rich allomorphy, nor the complex combinability of morphemes in MG appear to be treated efficiently through this technique, resulting in low scores of proper word segmentation (22% in inflectional suffixes and 13% in derivational ones), as well as the recognition of false morphemes.

1. Unsupervised Morphology Learning: A theoretical approach

1.1. An Introduction to Unsupervised Morphology Learning

As opposed to the computational analyses on syntax, computational work on morphology has been relatively scarce. According to Roark and Sproat (2007), the absence of a corpus of morphologically annotated words put a burden on the development of a machine learning morphological system that could rival a morphologically complex analyzer such as the one proposed by Koskenvuo (1983). However, close to the dawn of the new millennium, the interest in statistical models of morphology, particularly of unsupervised (or lightly supervised) morphology–learning from annotated corpora, has rapidly increased. Special attention has been paid to automatic – basically unsupervised – methods for the discovery of morphological alternations. However, allomorphy poses a serious problem for both tasks. By treating allomorphy, the goal is to find related morphological forms of the same word, such as κύμα and κύματα (kima– kimata(a)) (‘wave’), which are not the product of any phonological and morphological rules.

Since most of recent research has been carried out within the field of unsupervised morphological learning, we will focus our discussion and criticism on this system, and specifically on the theory of Minimum Length Description (MLD) proposed by Goldsmith (2001) [other recent works in the same direction are Yarowsky and Wicentowski 2001, Schone and Jurafsky 2001, Creutz and Lagus 2002]. Goldsmith’s (2001) theory and the implementation of his program Linguistica are based on the framework of Rissanen’s (1989) MLD. His article is not the first work on unsupervised morphology learning, as there are three other approaches by previous researchers. Nevertheless, this work is certainly the mostly cited, and is considered to be the standard model compared to other systems.

1.2. Goldsmith’s Minimum Length Description (2001)

Goldsmith’s system starts with a very large corpus of annotated texts and produces a range of signatures along with words that belong to these signatures. A Signature is a set of affixes (prefixes or suffixes) that combine with a given set of stems (Goldsmith, 2001; Roark and Sproat, 2007). An example suffix signature in English could be NULL.ed.in.g.s, which combines with the stems jump, laugh, walk, talk, etc., all of which take the signature’s suffixes in order to create words, such as jumps, jumped, jumping and jumps. Other examples of signatures are e.ed.in.g.s, NULL.s, NULL.in.g.s, NULL.est.ly, etc.

A closer look at the signatures reveals that the sets are not always complete. Usually the past tense suffixes are absent, even for regular verb stems. For example, Roark and Sproat (2007:120) point out that the signature NULL.ing.s proposed by Goldsmith (2001: 179), that includes stems such as blow, broadcast, drink, feel does not display the –ed suffix, since the verbs are irregular in their past tense form. However, the –ed suffix is also absent from stems such as bomb and farm, which, although regular in their past tense form (bombed and farmed), unfortunately did not occur in the corpus! Goldsmith discusses in general terms some problems with signatures and notes that his system is incapable of handling alternations (e.g. allomorphs), such as feel/ felt, since it deals only with affixation.

1.3. MLD Model Critism

As it will be demonstrated in the next section, this kind of allomorphic alternation can be an enormous problem, if one tries to apply an Unsupervised Morphology Learning Model (UMLM) for example to the Greek language, which exhibits a high degree of complex allomorphy in every word formation process (inflection,
derivation, compounding). The combinability of derivational suffixes and bound stems deteriorates the problem even more.

As Roark and Sproat (2007:123) correctly point out, Goldsmith’s method is “the de facto gold standard for work on unsupervised acquisition of morphology”. However, this system is still a far cry from perfection. As already observed, an UMLM does not use morphological and phonological rules, does not have a pre-built lexicon, and obviously does not take advantage of any linguistic (more specifically morphological) theory or framework. It only tries to split words on the basis of huge corpora. Several researchers complain that Goldsmith’s method does not exploit semantic and syntactic information. This criticism echoes the psycholinguistic approach and its objection to the fact that children and adults access other information besides the set of stems and affixes. Considering the fact that even morphological rules or theories are left out of the model, it would perhaps be too much to anticipate the use of semantic and structural information.

The failure to correctly segment words into actual morphemes is due to the lack of morphological and phonological rules, the non-use of Lexical Phonology and the occurrence of rare, marked and irregular cases. This can happen on both the orthographical and phonological levels of word transcription:

(1) ἐγραψα > ἐ – γραφ – σ(α) [dissimilation]
‘I wrote’ stem: γραφ
εγραψα> ε – γραφ – σ(α)
‘I wrote’ stem: γραφ

Parsing failure is more frequent in morphologically rich languages, such as Greek, Finnish, Swedish, Hungarian and Turkish. The high productivity of compounding and derivation complicates things more, introducing the factor of affix combinability. According to Kurimo et al (2007), the highest score of an UML model evaluation for Finnish and Turkish was 65% and 64% respectively, and the lowest score was 3% and 2%, in spite of the fact that Kurimo’s system was partly assisted by supervised morphology. One would expect that the application of Goldsmith’s theory to Greek would result in an even lower score, due to the extensive degree of allomorphy of the language (see Karasinos, 2001; Ralli, 2005, 2007), as well as the complex combinatorial properties of affixes and bound stems. Melissaropoulou (2007a, 2007b) and Melissaropoulou & Ralli (2008) note that in Greek, a sequence of as many as five derivational suffixes in a row can become; the selection restrictions and combinational choices of derivational suffixes and bound stems render the creation of these signatures almost impossible or completely defective.

2. Greek derivational affixes vs. Bound stems

2.1. Allomorphy and short overview of previous work

As already pointed out, allomorphy can be serious problem for UML models and an issue that almost no one in computational morphology tries to solve or even discuss. Allomorphs are different forms of the same morpheme that share lexical information, but differ unpredictably and arbitrarily in their phonological form and in the morphological environment, where they appear. Allomorphy is a central issue in morphology; however apart from a few exceptions it has never become the focus of attention, particularly within the generative grammar framework. As Ralli (2006: 2) claims “the reason for such neglect is mainly the fact that allomorphy is usually considered as nothing more than the absence of uniformity, resulting either from historical processes or from borrowing”.

Lieber (1982), Carstairs (1987), Booij (1997), and Ralli (1994, 2000, 2005, 2006) provide a thorough treatment of allomorphy proposing various analyses and raising several interesting points; their approaches deal with the problem from a morphological point of view. In particular, Ralli shows that the systematic allomorphic behavior of a number of Greek stems affects the organization of paradigms in a significant manner. Additionally, Karasinos (2001) provides a wide range of examples in all three word-formation processes, inflection, derivation and compounding, and shows how important allomorphy can be in the Greek language.

2.2. Derivational prefixes and suffixes

Affixes, depending on their position with respect to a stem/root, are distinguished into prefixes and suffixes. The prefixes are a small group of morphemes, the majority of which used to belong to the class of prepositions of Ancient Greek; some of them still participate in lexicalized phrases, such as ανά · ἐσο (ana etos) ‘per annum’, σὺν τοῖς ἄλλοις (sin tis alis) ‘moreover’.

1 DP = derivational prefix, DS = derivational suffix, IS = inflectional suffix
structures. Allen, 1978; Selkirk, 1982; Kiparsky 1982; Mohanan, framework of strata–oriented models (cf. Siegel 1974, those theoretically possible. of attested suffix combinations is generally smaller than susceptible to further suffixation, and c) the total number followed by other suffixes, while others are not and some of its dialects. They argue that: a) suffixes select base with a particular suffix in Standard Modern Greek principles, which underlie the structural combination of a 'park' 'I parked' suffix: ar(o) (parkaro) (parkarisa) b. suffix: TZH(Σ) allomorph: TZHΔ 'taxi driver' 'taxi drivers' 'park' 'I parked' suffix: ar(o) Melissaropoulou & Ralli (2009) deal with the general (3) a. prefix: YIHO allomorph: ΥΦ υπόλογος υπολογιστής υπολογιστής 'taxi driver' 'taxi drivers' suffix: AP(Ω) allomorph: API παρκάρο παρκάρισμα (parkaro) (parkarisa) ‘accountable’ ‘vice minister’ ‘opposing’ ‘unhealthy’ b. suffix: TZH(Σ) allomorph: TZHΔ ταξίτης ταξιτζήδες (taksitizis) (taksitziideshow) ‘taxi driver’ ‘taxi drivers’ suffix: AP(Ω) allomorph: API παρκάρο παρκάρισμα (parkaro) (parkarisa) ‘park’ ‘I parked’ suffix: ar(o) Melissaropoulou & Ralli (2009) deal with the general the prefixation, e.g. ipo–logos (‘responsible for one’s actions’), –γράφη (‘writer/recorder’ as above), –κομοσ), –κτονος (–ktos) ‘killer’ as above), –παθη (‘science/study’, as in –γλωσσ (‘science’, as in –λογος (‘theology’). Such as (–loğos) ‘‘science’’, as in –λογος (‘‘science’’, as in –λογος (‘‘science’’, as in –λογος (‘‘science’’, as in –λογος (‘‘science’’, as in ‘supporter’), ii) carry more concrete meaning in in the constructions they appear, and iv) participate in the presence of the linking vowel –o–, which constitutes a compound marker in Greek (Ralli 2008b), e.g. παρ–ο–κτονος (παρακτονος) ‘patricide’ (agentive) and from the recursivity they exhibit in their structures, e.g. κοινονι–γλòσσα–ο–λογος (κοινονιολογος) ‘socio–linguist’), which characterizes compounding. The structures corresponding to the opposing views presented above for a word involving a bound element such as βιολογος (viologos) ‘biologist’ are formulated as follows: a) βιο–λογος, where the element –λογος is a conﬁx, and b) βι–ο–λογος, where the element –λογος is a bound stem. Although, there is seemingly no signiﬁcant difference between the two structures, the implications they have for the computational treatment of words containing these elements, are signiﬁcant. This stems
from the fact that as Ralli (2008a) has noticed, words containing bound elements, regularly serve as bases for the formation of derivatives, through suffixation, selecting suffixes from a closed set and giving rise to words such as βιολογία (biología) ‘biology’, βιολογικός (viologikós) ‘biological’ and so forth. Confixation in this case, which renders the elements –λόγος (–logos) and –γραφά (–grafas), as separate items belonging to the closed set of confixes, with no apparent morphological association between them, gives rise to the unrelated structures βιολογία and βιολογικός, thus obscuring the obvious morphological relationship between the two items. In these terms, the structure of the word βιολογία (viología), more than it is related, for example, to the structure of the word βιογραφία (viografía) ‘biography’ sharing with both of them only the same initial stem and a different confix. In computational terms, this would require the insertion of all possible confixes2 (e.g. –λόγος (–logos), –γραφά (–grafas), –γραφία (–grafía), –κόσμος (–kósmos), –κοσμία (–kosiá) keeping them unrelated to each other.

On the other hand, the ‘bound stem’ view gives rise to the structure βιο–λόγος(ος), which then, according to Ralli (2008a) serves as a base for the derivation of the word βιολογία (biology). In computational terms, this would require the insertion of all bound elements with verbal origin, along with the possible suffixes they may receive, namely the –ια (–ia), –ικ– (–ik–), –ειο (–io), –σασ (–issa), –πια (–pia), all of which are common suffixes in MG attaching to other bases apart from compounds with bound elements (e.g. κατικικ–ια (katikiia) ‘residence’, φιλικ–ος (filikos) ‘friendly’, Ασιάτισσα (Asiatissa) ‘female Asian’ etc.). Apart from the obvious economy of the ‘bound stem’ solution, it serves for greater accuracy in the morphological analysis obtained, as it preserves the morphological relationships between words.

Therefore, we compiled a corpus consisting of about 7000 words, each containing one of the 54 bound stems with verbal origin found in MG, such as –λόγος (–logos), –γραφά (–grafas), –κριτ (–krit), –δησι (–dési), –πέτη (–péti), –γόν (–gon), –γεν (–gen), –μαθ (–mat), –μαν (–man) etc. along with their derivatives formed with the nominalising suffixes –(e)ια, –(e)ιο, –issa, –(e)ia (e.g. archeologi–ια (‘archaeology’), emodot–ια (‘female blood donor’), kosmopoiia (‘cosmology’) viilooöt–eio (‘bookbinding site’)) and verbs ending in (o) arising from conversion (e.g. limoikton(ο) ‘starve’).

3. The Linguistica Experiment

3.1. About Linguistica

Linguistica is a program designed to explore the unsupervised learning of natural language, with primary focus on morphology (word–structure). In the case of unsupervised learning of morphology, Linguistica explores the possibilities of morpheme–combinations for a set of words, based on no internal knowledge of the language from which the words are drawn.

Segmentation is the first task of this process; the program figures out where the morpheme boundaries are in the words, and then decides which of them are stems, affixes and so forth. Most of Linguistica’s functionality, at this point, goes into making these decisions. For our experiment, we used the 3.2.6 version (March 2009) for Windows XP.

Figure 1: The interface of software Linguistica

3.2. Find Allomorphy with Linguistica

It is referred that Linguistica is capable of determining a limited amount of allomorphy in stems. In many languages (including English), stem final material is deleted in front of certain suffixes. For example, stem–final –e is deleted in English before a number of suffixes: love, but lov–ing and not love–ing; sane, and sanity, not sane–ity. Goldsmith treats this as allomorphy, although it is not.

The strategy of Linguistica is to reanalyze material that had been previously included in a suffix as part of the stem, and provide the information that other suffixes must delete that material, when it appears before them. Goldsmith (2001) illustrates this with the following example: the words love, loves, loved, and loving, which had been analyzed as lov + signature e.ed.es.ing, will be reanalyzed with the stem love and the suffixes NULL, ed, s, and ing. The suffixes –ed and –ing will be informed that they are capable of deleting the preceding e, and this is indicated by placing an e in angle brackets before the prefix, thus: <e>ing and <e>ed. Thus the new signature for love is NULL.<ed.<e>ed.<e>ing.s, and this signature correctly deals both with stems that end in –e and those that do not.

Additionally it is pointed out that Linguistica treats y–final nouns and verbs in the same way: academy/academies are treated as if based on the stem
3.3. Our experiment corpora

As already put forward, our hypothesis is that *Linguistica* would appear to have major problems in analyzing a corpus of Greek words. In order to test this, three text corpora were created ad hoc; the first had 60,000 tokens (28,000 words) from a newspapers corpus, the second had 8,500 words with carefully selected lemmas and entries (words with same inflectional and derivational suffixes, groups of common prefixed words, etc) and the third was a unpublished science fiction novel with 200,000 words written by the first author of the present work. The results from the first corpus were quite disappointing (3% accuracy). The results from the third corpus were slightly better, but the level of accuracy was still quite low (6%). On the other hand, the results from the second corpus were more specific and clear, although the level of accuracy was also quite low. The system managed to detect several inflectional paradigms, few derivational suffixes and some bound stems. Additionally, only two allomorphy types were recognized, only one of which was correct, i.e. the παιδ–παίδ–type allomorphs!

3.4. Results

The application of *Linguistica* to our data produced the following top ten signatures are: (i) NULL.δες.δων, (ii.) άρες.άρετε.άριστα.άρουμε. άρο, (iii.) άρα.ες.τις.ικός.ον, (iv.) άτον.ατάκα.ατάμα, (v.) ά.σκι.λευ.ο.όν, (vi.) ά.άδες.άς, (vii.) ες.εδάκι, (viii.) ά.δες.άζη, (ix.) νες.νον and (x.) NULL.ες. ες.

The results are derived through the application of an advanced system with heuristics (see Goldsmith, 2001). Goldsmith points out that the overall sketch of the morphology of English and other European languages comes out quite normal in its outlines. Nevertheless, the results from the English experiments, when studied closely, show that there are some parsing errors. The author of *Linguistica* tried quite successfully to fix these errors with additional heuristics and to evaluate them using the MLD measure. However, the results from the Greek corpora do not demand a closer study, since the errors form the rule rather than the exception. These errors may be organized in the following ways:

a. The collapsing of two or more suffixes into one: for example, here we find the suffix –ικός (–ίκοσ); in most corpora, the equally spurious suffix –ευτικός (–εύτικοσ) is found.

b. The systematic inclusion of stem–final material into a set of (spurious) suffixes. In Greek, for example, the high frequency of stem–final –τ ( κύματα– (kimata)) can lead the system to the analysis of a set of suffixes as in the spurious signature τος.τα.τον or τακι.ταρα.

c. The inclusion of spurious signatures, largely derived from short stems and short suffixes, and the question related to the extent of the inclusion of signatures based on real, but overapplied, suffixes. For example, –ς (–s) is a real suffix of Greek, but not every word ending in –ος (–us) should be analyzed as containing that suffix.

d. The failure to segment all words actually containing the same stem in a consistent fashion: for example, the stem χωρ with the signature ος.οι.ων (οις.οι.ων) is not related to χωρ with the signature ουν.ευκ.εκ.εξει.ετε. 

e. Stems may be related in a language without being identical. The stem αυ is may be identified with the signature α.αται.αται and the stem α may be identified with the signature μαται.μαται, but these stems should be morphologically related.

f. The system has never identified the linking vowel –ο of the bound stems as a separate element. It was always attached either to the first component (γλωσσο–) or to the bound stem (–ολόγος) without any systematic treatment.

g. *Linguistica* failed to treat allomorphy correctly.

3.4.1. Prefixation

The analysis of prefixes in Greek should not pose a serious problem for *Linguistica*, since there are very few and with limited allomorphy. The system managed to create signatures like συν.αντι sin.anti (ἐργατικός) (εργατικος), ἐνζυμο (enzoymo), εισφορά (isfora), αντι.κατα (anti.kata) {βάλλω (vallo), θέτω (θετo)}, συν (sin) {θέτo (θετo), τρέχo (trexo), ἄγω (αγω)}, which contain true prefixes. Nevertheless, as mentioned above, signatures with two prefixes combined were also created, such as συν.συναρ (sin.sinapo) {τηρo (tiro), τηρητικός (tiritis), τηροίμα (tirume)}, συν.συναρμ (sin.sinarmo) {δηλώνo (dilwono), δηλατικός (dilatomitos)} and αντι.συναρμ (anti.sinarmo) {γράφωμα (grapfome), γεγραμμένος (gegrammenos)}. Additionally, the system failed to relate prefixes with common characters like α (α–) and αν (αν–), κατα (κατ–) and καται (καται–) or the most changeable prefix συν (sin) {συν– (sim–), συν– (sin–), συν– (sin–), συν– (sin–), συν– (sic–)}, since the system does not incorporate any phonological rules, such as deletion and assimilation. Moreover, it was very common in spurious signatures to include some of the first characters of the stem in the prefixes (i.e. συνδ– (sinδ–), συναρ– (sinar–), συνθήκ– (sinthik–), συναρμ– (sinar–)) or to mislabel part of stems as prefixes (γτ– (gt–), ετε– (ete–), τρε– (tre–)). Finally, *Linguistica* could not detect any allomorphic behaviour of prefixes and of course it failed to relate them with other true forms of the same prefix, for example κατα (κατ–) and καται (καται–).
(trexo) ‘run’) and o.o.o.o.v.a.[ακι] (βουνό (yuno) ‘mountain’), μορό (moro) ‘baby’, νερό (nero) ‘water’). Except for three other signatures, the rest of them (62) were spurious. There is an average number of signatures with combined suffixes (usually a derivational with an inflectional), such as όρεων[αρσεν.]άρσεως.ορισμος.ορισμού, ατζή.ατζής.ατζήδος.ατζήθην or ευπτώκης.κυνικοί (χορός (xoros) ‘dance’, δήμος (dimos) ‘municipality’). It was a very common mistake to create suffixes by including the last character of the stem; for example γα.ξα (ανοι (ani), τολι (tilli), διαλε (diale)) or ινα.να (γλυκα (glyka), πικρα (pikra), λευκα (lefka)).

Table 1: Signatures of inflectional and derivational suffixes

Goldsmith tried to fix this problem by advancing the heuristics and applying the feature “short–length for non–stems”; however, the treatment of one–character cases, if there was a V–deletion allomorphy (i.e. (xoros) ‘dance’, όρεων[αρσεν.]άρσεως.ορισμος.ορισμού, ατζή.ατζής.ατζήδος.ατζήθην or ευπτώκης.κυνικοί (χορός (xoros) ‘dance’, δήμος (dimos) ‘municipality’). It was a very common mistake to create suffixes by including the last character of the stem; for example γα.ξα (ανοι (ani), τολι (tilli), διαλε (diale)).

3 We consider as true signatures, the signatures that contain real suffixes. Of course, some signatures did not contain all the inflectional paradigms of a noun or a verb.

The system failed to relate any of the stems. Also the statistical analysis of both corpora reveals that only 4% of the allomorphs were detected by Linguistica. These results are similar to those of Kurimo et al (2007) for Finnish and Turkish; moreover, the hypothesis of Linguistica’s inadequacy in dealing with Greek allomorphy expressed by Karasimos (2008) was experimentally tested and found to be valid.

3.4.4. Compounds and Bound stems

Linguistica could not analyze any compounds. Its strategy and architecture is to extract suffixes and prefixes even for languages with rich morphology. English corpora that were tested with it contained very few one–word compounds and a significant group of neoclassical compounds; the authors do not show that this system treated them correctly. Unfortunately the three Greek test corpora cannot provide any serious conclusions for Greek compounds, since the results were totally haphazard. As a rule, the system was unable to recognize any of the compound’s components and failed to analyze many of them.

As we already mentioned the inability to feed the system with any rules or structural information means that, despite our preferred morphological analysis of the words involving bound elements, the analysis obtained by the system would not necessarily be the desired one, which was the case for one example. Specifically, among the signatures produced by the analysis of our ‘bound–stem’ corpus, we found ‘real’ suffixes, such as the derivational –ία (e.g. θεολογία–ία (θεολογία) ‘theology’), –σιο (e.g. ανθοποιοί–σιο (ανθοποιοί) ‘flower shop’), –της (e.g. αιμοδοθής–της (αιμοδόθης) ‘blood donor’), –αρεία (e.g. παλαιοποιολογική–αρεία (παλαιοποιολογικά) ‘female antique seller’), the nominal inflectional (ος) (e.g. βοτανολόγος (βοτανολογώς) ‘botanologist’), (ης) (e.g. πατριαρχής (πατριαρχίς) ‘patriarch’), and the verbal inflectional (ω) (e.g. χρησιμοποιώντας (χρησιμοποιώντας) ‘sound record’). However, we also found sequences like –άλογος (–άλογος), –ειοτής (–ειοτής), –αφότοις (–αφότοις), –κυκλοφόρο (–κυκλοφόρο), –κακοίς (–κακοίς), –εύζοις (–εύζοις), –ευπροσδοκητής (–ευπροσδοκητής), –ευπτώκης (–ευπτώκης), –ευπλοκοτής (–ευπλοκοτής), –ευπροσδοκήτης (–ευπροσδοκήτης), –ευπροσδοκήτης (–ευπροσδοκήτης), (etc. As we can see, the accuracy of the system was 13% for derivational infixes and 22% for inflectional suffixes.

3.4.3. Stems

Linguistica presented a common behaviour in the analysis of nominal stems. First of all, only nominal allomorphs of the παντι–type were detected. In the other cases, if there was a V–deletion allomorphy (i.e. καρδιώ–καρδί (καρδία–καρδί) ‘heart’), the system detected only the V–deleted stem (καρδί)–considering the deleted vowel as a suffix. Moreover, if there was a C–insertion allomorphy (i.e. κύμα–κύματα (kima–kimat) ‘wave’), the system recognised the final consonant of the allomorphs as the initial consonant of the suffixes (κύμα). Additionally, there were a few signatures with spurious suffixes that contained the last two characters of the stem, such as νας.γα.νας.ναν (σολή (soli), πυρή (piri), ανά (eoa), λίμε (lime)) and γα.ξα (ανοι (ani), τολι (tilli), διαλέ (diale)).
attached, as those mentioned above, gave rise to a great number of ‘correct’ stems of MG like mitθ– (‘myth’), okean– (‘ocean’), selin– (‘moon’), musik– (‘music’), xart– (‘paper’), sišir– (‘iron’) or stem allomorphs like ḍramat– (‘drama’), xromat– (‘colour’), bavmat– (‘miracle’), stromat– (‘matress’), nimat– (‘thread’) and so on, appearing as right hand elements in the words provided. However, also as stems were recognized sequences that are like compound stems, such as kriptoγraf–, sismoloγ–, karokinoyon–, vivlioklop–, texnoloγ–, plutokrat–, due to the recognition of true derivational and inflectional suffixes that we mentioned above.

As a result, we should note that the system did not manage to recognize any of the bound stems such as –loγ, -γraf, –kton, –matθ, –krat and so on, neither the linking element –o–, as proposed by the preferred morphological analysis for the words involving bound elements in MG. As we mentioned above, this fact was basically due to the lack of any morphological input to the system, which could lead the morphological analysis towards a particular direction.

4. Conclusions

Computational Morphology is a rapidly growing area of linguistics. Unsupervised Morphology Learning Theory is a recent approach to morphological analysis problems, and seems to work well for languages with poor inflectional morphology, although any attempt to use this theory in morphologically rich languages, such as Finnish and Turkish, could be characterized at least as mediocre (Kurimo et al., 2006, 2008). We claim that a system without: a.) prior human–designed analysis of the grammatical morphemes of a language, b.) some identifying stems and affixes and c.) pre–imported morphological and phonological rules for correct parsing, is bound to fail. A system that builds lexica based on a common sequence of phonemes without proper rules is unable to treat successfully the complex combinations of derivational suffixes and bound stems. As already shown, the phenomenon of allomorphy in Greek is very extensive. Allomorphy participates with the same frequency in every word formation process. A natural question to ask is whether a UML model is able to analyze processes and successfully treat suffixes and bound stems. We have presented a considerable amount of data with allomorphs and shown the complexity of the allomorphic changes, the combinability of derivational affixes and the normality of bound stems. Since the insertion of processing rules for allomorphy is not allowed in a UML model, the goal of correct parsing will never be attained. From a more theoretical point of view, our work has nothing to do with the current question: does a young speaker learn a language and segment the morphemes the way that a UML does? Thus, we would like to point 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 and compounding in Modern Greek.

5. Acknowledgements

We would like to thank prof. Angela Ralli and the researchers of the Laboratory of Modern Greek Dialects, Department of Philology for their helpful comments.

6. References

Allen, J., Hunnicutt, M. S. & D. Klatt (1987). From Text to Speech: the MITalk System. Cambridge: Cambridge University Press.
Anastasiad–Symeonidi, A. (1986). Neology in Common Modern Greek [In Greek], Thessaloniki.
Booj, G. (1997). Allomorphy and the Autonomy of Morphology, Folia Linguistica XXX/1–2: 25–56.
Carstairs, A., 1989. Allomorphy in inflection. London: Croom Helm.
Creutz, M. & K. Lagus (2002). Unsupervised discovery of morphemes. Proceedings of the ACL–02 Workshop on Morphological and Phonological Learning, 21–30.
Giannoulopoulou, G. (2000). Morphosemantic comparison of Affixes and Confixes in Modern Greek and Italian [in Greek]. Thessaloniki.
Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language, Computational Linguistics 27, vol 2: 153–196.
Koskenniemi, K. (1983). Two–Level Morphology: A General Computational Model for Word–form Recognition and Production, Proceedings COLING ’84, 178–181.
Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E. & M. Saraclo (2006). Unsupervised segmentation of words into morphemes – Challenge 2005: An Introduction and Evaluation Report. Journal of Proceedings ICSLP 2006.
Kurimo M., Mathias C.& M. Varjokallio (2008). Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard Morpho Challenge 2007 In A. Nardi & C. Peters (eds.) Working Notes of the CLEF 2007 Workshop.
Lieber, R. (1982). Allomorphy. Linguistics Analysis 10(1): 27–52.
Martinet, A. (1979). Grammaire Fonctionelle du Francais. Paris: Didier.
Melissaropoulou D. & A. Ralli (2009, submitted in Morphology). Structural combinatorial properties of Greek derivational suffixes. Paper presented at the 13th International Morphology Meeting (Vienna, February 3-6 2008), Workshop on Affix Ordering
Melissaropoulou, D. (2007b, in print). Remarks on the combinability of derivational suffixes in Greek and its dialectal variation. In Proceedings of the 3rd International Conference on Modern Greek Dialects and Linguistic Theory (Nikosia, 14-16/06/2007).
Petropoulou, E. (2009). On the parallel between Neoclassical compounds in English and Modern Greek. In A. Ralli (ed.) Patras Working Papers in Linguistics, Vol.1. Special Issue: Morphology. Centre of Modern Greek Dialects. Department of Philology. University of Patras.
Ralli, A. (1994). Feature Representations and Feature–Passing operations in Greek Nominal Inflection. Proceedings of the 8th Symposium on English and Greek Linguistics: 19–46. Thessaloniki: English Dept. Aristotle University of Thessaloniki.

\footnote{i.e. without their inflectional ending as they normally appear in compounds.}
Ralli, A. (2000). A feature–based analysis of Greek nominal inflection, *Glossolojia* 11–12: 201–228.
Ralli, A. (2006). On the role of Allomorphy in inflectional Morphology: Evidence from Dialectal variation. *Advances of Language Studies* 1: 1–32.
Ralli, A. (2008a). Greek Deverbal Compounds with Bound Stems. *Journal of Southern Linguistics* 29 (1/2): 150-173.
Ralli, A. (2008b). Compound Markers and Parametric Variation. *Sprachtypologie und Universalienforschung* (STUF) 61(1): 19-38.
Ralli, A. (2009). Hellenic Compounding. In R. Lieber & P. Stekauer (eds.) *The Oxford Handbook of Compounds*, 453-464. Oxford: Oxford University Press.
Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific Publishing Co.
Roark B. & R. Sproat (2007). *Computational Approaches to Morphology and Syntax*. Oxford: Oxford University Press.
Schone, P. & D. Jurafsky (2001). Knowledge–free induction of morphology using latent semantic analysis. *Proceedings of the 4th Conference on Comoputational Natural Language Learning* (CoNLL): 67–72.
Warren, Beatrice (1990). The importance of combining forms. In Wolfgang Dressler et al. (eds) *Contemporary Morphology*. Berlin: Mouton de Gruyter.
Yarowsky, D. & R. Wicentowski (2001). Minimally supervised morphological analysis by multimodal alignment. *Proceeding of the 39th Annual Meeting of the Association for Computational Linguistics* (ACL): 207–216.
Карахімос, Α. (2001). Η αλλομορφία στην κλίση και τη σύνθεση της Ελληνικής Γλώσσας, Πτυχιακή εργασία, Πανεπιστήμιο Πατρών.
Μελισσαροπούλου, Δ. (2007α). Μορφολογική περιγραφή και ανάλυση του μικρασιάτικου ιδιώματος της περιοχής Κιδωνίων και Μοσχονήσιον: η παραγωγή λέξεων, Διδακτορική διατρήση, Πανεπιστήμιο Πατρών.
Ράλλη, Α. (2005). Μορφολογία. Αθήνα: Πατάκη.