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 Koskenniemi (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~ kimat(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.ing.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 *jumpø, jumped, jumping* and *jumps*. Other examples of signatures are *e.ed.ing, NULL.s, NULL.ing.s, NULL.er.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.er.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) έγραψα > ε – γραφ – σ(α) [dissimilation]
'I wrote' stem: γραφ
eγrapsa> e – γraf – s(a)
'I wrote' stem: γraf

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 the model to Greek would result in an even lower score, due to the extensive degree of allomorphy of the language (see Karasimos, 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 may be found within the same word.

(2) χορ – ευ – τ – αρ – ουλ – ικ –(ο)
stem –ds[1] –ds –ds –ds –ds –(is)
xoreftaruliko
'little great dancer'
κοιν – ων – ι – ικ – οτητα(ø)
stem – ds – ds – ds – ds – (is)
kinonikotita
'sociability'

ξανά – επ – ανα – λαμβάν(ω)
dp – dp – dp stem
ksanaepanalamvano
'repeat again'
παρά – συν – παρα –στεκ(ομαι)
dp – dp – dp stem
parasimbarastekome
'aid (sb) too much'

Going back to Goldsmith's theory, a signature is a set of suffixes that can be attached to a set of stems. Therefore, one should create signatures of suffixes that combine with other signatures. It is easy to imagine how complex a system with a net of suffix/ prefix signatures 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, Karasimos (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 έτος* (ana etos) 'per annum', *συν τοις άλλοις* (sin tis alis) 'moreover'.

[1] DP = derivational prefix, DS = derivational suffix, IS = inflectional suffix

Only 32% of the prefixes display allomorphic behaviour. This allomorphy is mostly due to certain phonological rules that became inactive in Modern Greek, such as Grassman's Law or the aspiration principle. On the other hand, suffixes constitute a larger set than prefixes. They come in two varieties, inflectional and derivational, both subcategories being quite large for a closed–set, and both exhibiting considerable allomorphy, as 85% of suffixes have allomorphs. The allomorphic changes apply to both stems and suffixes. More specifically, items sharing the same morphological (noun, verb or adjective, inflectional endings) and phonological features (same final character) exhibit similar allomorphic behavior.

(3) a. prefix: ΥΠΟ allomorph: ΥΦ
υπόλογος υφυπουργός
(ipoloγos) (ifipurγos)
'accountable' 'vice minister'
prefix: ANTI allomorph: ΑΝΘ
αντιμέτωπος ανθυγιεινός
(antimetopos) (anθiγiinos)
'opposing' 'unhealthy'
b. suffix: ΤΖΗ(Σ) allomorph: ΤΖΗΔ
ταξιτζής ταξιτζήδες
(taksitzis) (taksitziδes)
'taxi driver' 'taxi drivers'
suffix: AP(Ω) allomorph: API
παρκάρω παρκάρισα
(parkaro) (parkarisa)
'I park' 'I parked' suffix: ar(o)

Melissaropoulou & Ralli (2009) deal with the general principles, which underlie the structural combination of a base with a particular suffix in Standard Modern Greek and some of its dialects. They argue that: a) suffixes select bases of a specific type, b) certain suffixes can be followed by other suffixes, while others are not susceptible to further suffixation, and c) the total number of attested suffix combinations is generally smaller than those theoretically possible.

The first systematic attempt to account for the combinatorial behavior of affixes was made within the framework of strata–oriented models (cf. Siegel 1974, Allen, 1978; Selkirk, 1982; Kiparsky 1982; Mohanan, 1986), according to which the different combinatorial properties of derivational affixes follow, to a great extent, from the position they hold into the different 'lexical strata' ('levels' in Kiparsky's 1982 terms).

Therefore, in the light of evidence provided above, we argue in favor of the main thesis taken by Fabb (1988), Scalise (1994) and Melissaropoulou & Ralli (2009), according to which suffix–driven selectional restrictions are the ones that govern the formation of derivational structures.

2.3. Bound stems

Another case of interest in the morphological parsing of MG is a special type of words containing bound stems. As discussed in Petropoulou (2009), this class of words comprises part of what we call neoclassical compounds in MG, because, like neoclassical compounds in English, they contain a bound element of Ancient Greek origin. Examples are *νηπι–αγωγ(ος)* (nipiaγogos) 'preschool teacher', *παθο–γον(ος)* (paθogonos) 'pathogenic', *δακτυλο–γραφ(ος)* (δaktiloγrafos) 'typist', *σκηνο–θετη(ς)* (skinoθetis) 'director', *τυρο–κομ(ος)* (tirokomos) 'cheese producer', *εντομο–κτον(ο)* (entomoktono) 'insecticide', *μετεωρο–λογ(ος)* (meteoroloγos) 'meteorologist', *καρδιο–παθ(ης)* (karδiopaθis) 'cardiopath', where the elements *–αγωγ(ος)* (–aγogos), *–γον(ος)* (–γonos), *–γραφ(ος)* (–γrafos), *–θετη(ς)* (–θetis), *–κομ(ος)* (–komos), *–kton(o)* (–ktono), *–λογ(os)* (–loγos) and *–paθ(is)* (–paθis) are bound morphemes, since they cannot stand as free words.

According to Giannoulopoulou (2000), following Anastasiadi–Simeonidi (1986), these elements are considered as 'confixes' (Martinet 1979), as they appear to acquire gradually more and more characteristics of suffixes. In these terms, confixes are secreted parts of words (Jespersen 1941, Warren 1990), which have been associated with a new specialized meaning. Examples of confixes cited by Giannoulopoulou (2000), presented here with their extended meanings, are *–λόγος* ((–loγos) 'scientist' as above), *–λογία* ((–loγia) 'science', as in *θεολογία* (θeoloγia) 'theology'), *–γράφος* ((–γrafos) 'writer/recorder' as above), *–γραφία* ((–γrafia) 'science/study', as in *ωκεανογραφία* (okeanoγrafia) 'oceanography'), *–κτόνος* ((–ktonos) 'killer', as above), *–κτονία* ((–ktonia) 'killing' as in *πατροκτονία* (patroktonia) 'patricide'), *–ποιός* ((–pios) 'maker' as in *επιπλοποιός* (epiplopios) 'carpenter/ (lit.) furniture maker'.

On the other hand, Ralli (2008a) supports that these elements are bound stems of a verbal origin and defies the opinion favouring their suffixal character presenting a number of opposing arguments. She claims that these elements: i) can serve as bases to prefixation, e.g. *ipo–logos* ('responsible for one's actions'), *υπερ–μαχος* ('supporter'), ii) carry more concrete meaning in comparison to affixes which have a more functional role, often expressing agentive or instrumental meaning, iii) carry valency information, i.e. information about the obligatory complements of the verbs they derive from, calling for theta–role saturation by the left–hand element in the constructions they appear, and iv) participate in compound structures, which are recognizable both from the presence of the linking vowel –o–, which constitutes a compound marker in Greek (Ralli 2008b), e.g. *πατρ–ο–κτονος* ((patroktonos) 'patricide' (agentive)) and from the recursivity they exhibit in their structures, e.g. *κοινωνι–ο–γλωσσ–ο–λόγος* ((kinonioloγos) 'socio–linguist'), which characterizes compounding.

The structures corresponding to the opposing views presented above for a word involving a bound element such as *βιολόγος* (violoγos) 'biologist' are formulated as follows: a) *βιο–λογος*, where the element *–λογος* is a confix, and b) *βι–ο–λογ(ος)*, where the element *–λογ* 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 *βιολογ–ια* (violoγia) 'biology', *βιολογ–ικ(ος)* (violoγikos) 'biological' and so forth. Confixation in this case, which renders the elements *–λογος* (–loγos) and *–λογια* (–loγia) as separate items belonging to the closed set of confixes, with no apparent morphological association between them, gives rise to the unrelated structures *βιο+–λογος* and *βιο+–λογια*, thus obscuring the obvious morphological relationship between the two items. In these terms, the structure of the word *βιολόγος* (violoγos) is not related to the structure of the word *βιολογία* (violoγia), more than it is related, for example, to the structure of the word *βιογραφία* (vioγrafia) '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 confixes[2] (e.g. *–λογος* (–loγos), *–λογία* (–loγia), *–γράφος* (–γrafos), *–γραφία* (–γrafia), *–κτόνος* (–ktonos), *–κτονία* (–ktonia)) keeping them unrelated to each other.

On the other hand, the 'bound stem' view gives rise to the structure *βι–ο–λογ(ος)*, which then, according to Ralli (2008a) serves as a base for the derivation of the word *βιολογία* (*βι–ο–λογ+ια)*. 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)*, *–ισσα (–issa)*, *–ρια* (*–ria)*, all of which are common suffixes in MG attaching to other bases apart from compounds with bound elements (e.g. *κατοικ–ια* (katikia) 'residence', *φιλ–ικ(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 *–λογ (–loγ)*, *–γραφ (–γraf)*, *–κρατ* (*–krat)*, *–δοτη* (*–δoti)*, *–δετη (–δeti)*, *–γον (–γon)*, *–γεν (–γen)*, *–μαθ (–maθ)*, *–μαν (–man)* etc. along with their derivatives formed with the nominalising suffixes *–(e)ia*, *–(e)io*, *–issa*, *–ria* (e.g. *archeologʏ–ia* ('archaeology'), *emoδot–ria* ('female blood donor'), *kosmoγonia* ('cosmogony'), *vivlioδet–eio* ('bookbinding site')) and verbs ending in *(o)* arising from conversion (e.g. limokton(o) '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.

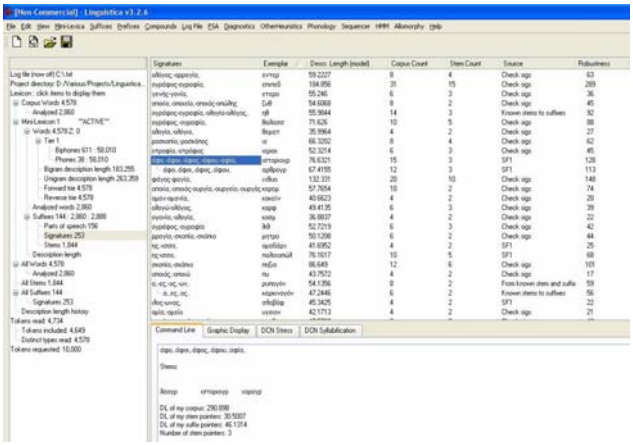

Figure 1: The interface of software *Linguistica*

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.<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

[2] The collection of confixes provided by Giannoulopoulou (2000) is not exhaustive, consisting only of a part of elements that could be classified as confixes, which may mean that potential confixes might have to satisfy a number of criteria in order to enter this class of items. This would leave out a significant number of elements, which would have to be treated in other terms.

academy and the suffixes NULL and <*y*>*ies*.

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 *παιδι~παιδ*–type allomorphs!

3.4. Results

The application of Linguistica to our data produced the following top ten signatures are: (i.) NULL.δες.δων, (ii.) άρεις.άρετε.άρισα.άρουμε. άρω, (iii.) άρα.ες.ης.ικός.ων, (iv.) άτων.ατάκι.ατάρα, (v.) ά.άκι.ου.ο.ων, (vi.) ά.άδες.άς, (vii.) ές.εδάκι, (viii.) ά.άς.ατζή, (ix.) νες.νων and (x.) NULL.είς. The first signature is composed of noun stems with a δ–allomorph (μεζέ, κουβά, μαμά), the second includes foreign stems, which form verbs with –άρ(ω) (σκορ, σοκ, σκαν) and the fifth is only combined with neutral nouns belonging to the sixth inflectional class (βουνό, μωρό, νερό).

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* (–ikos); in most corpora, the equally spurious suffix *–ευτικός* (–eftikos) 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* (κύματ–α (kimata)) can lead the system to the analysis of a set of suffixes as in the spurious signature *τος,τα.των* 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 *–ους* (–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 *χορ* with the signature *ος.οι.ους* is not related to *χορ* with the signature *ευω.ευεις.ευει*.etc.
e. Stems may be related in a language without being identical. The stem *αιμ* may be identified with the signature *α.ατα.ατο* and the stem *αι* may be identified with the signature *ματακι.ματαρα*, but these stems should be morphologically related.
f. The system has never identified the linking vowel –o– of the bound stems as a separate element. It was always attached either to the first component (γλωσσ**ο–**) or to the bound stem (–**ο**λόγος) 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 {εργατικός (ergatikos), ένζυμο (enzimo), εισφορά (isfora)}*, αντι.κατα* (anti.kata) {βάλλω (valo), θέτω (θeto)}*, συν* (sin) {θετώ (θeto, τρέχω (trexo), άγω(αγο)}, which contain true prefixes. Nevertheless, as mentioned above, signatures with two prefixes combined were also created, such as *συν.συνεπι* (sin.sinepi) {τηρω (tiro), τηρητής (tiritis), τηρούμαι (tirume)}*, συν.συνυπο* (sin.sinipo) {δηλώνω (δilono), δηλωτικός (δilonotikos)} and *αντι.συνυπο* (anti.sinipo) {γράφομαι (γrafome), γεγραμμένος (γeγramenos)}. Additionally, the system failed to relate prefixes with common characters like *α–* (a–) and *αν–* (an–)*, κατα–* (kata–) and *κατ–* (kat–) or the most changeable prefix *συν–* (sin) {*συμ– (sim–), συγ– (siγ–), συλ– (sil–), συρ– (sir–), συσ–(sis–)*}, 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–), συνθηκ– (sinθik–), συναρμ– (sinarm–)*) or to mislabel part of stems as prefixes (*γλ–* (γl–), *λευ–* (lef–)). 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–* (kata–) and *καθ–* (kaθ–), *υπο–* (ipo–) and *υφ–*(if–).

3.4.2. Suffixation

The suffixal system of the Greek language is quite complex; *Linguistica* succeeded in creating some inflectional paradigms like the verbal present *ω.εις.ει.ουμε.ουτε.ουν* (γράφω (γrafo) ‘write’, τρέχω

(trexo) ‘run’) and *ο.ου.ων.α.[ακι]* (βουνό (vuno) ‘mountain’, μωρό (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 *άρεις.αρει.αρω.αρουμε.αρισα*, *ατζη.ατζης.ατζηδες.ατζηδων* or *ευτικός.ευτικοί* (χορός (xoros) ‘dance’, δήμος (δimos) ‘municipality’). It was a very common mistake to create suffixes by including the last character of the stem; for example *γα.ξα* (ανοι (ani), τυλι (tili), διαλε (δiale)) or *ινα.να* (γλυκα (γlika), πικρα (pikra), λευκα (lefka)).

ω.εις.ει.ουμε.ετε.ουν SIGNATURE (SG)
γραφ, τρεχ, δεν, βαζ, καν
ε.ινος.ο.οι.ος.ου.ους.ων SG
ανθρωπ, κακτ, βαλτ
ο.ου.α.ων.ακι SG
βουν,νερ, μωρ, κακ, ποτ
άνθηκα.άνθηκες.αίνομαι.αίνουμε.αίνω.ανθείς.ανθώ
λευκ, γλυκ, μωρ
γα.ξα SG
ανοι, διαλε, κοιτα, τυλι
ριού.ριων.ρί SG
καλαμα, ποτη, σαμα, σφυ

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 suffixes and prefixes is an important issue that causes many difficulties for a UML system. Finally, as claimed in our hypothesis, *Linguistica* failed to detect suffixal allomorphy, since the system did not relate the suffixes and usually failed to analyze them (45% failure). Therefore, it identified suffixes such as *αρω.αρισα* instead of *αρ~αρι (αρ<ι>), ατζής.ατζήδων.ατζήδες* instead of *τζη~τζηδ (τζη<δ>)* etc. As we can see, the accuracy of the system was 13% for derivational suffixes and 22% for inflectional suffixes[3].

3.4.3. Stems

Linguistica presented a common behaviour in the analysis of nominal stems. First of all, only nominal allomorphs of the *παιδί*–type were detected. In the other cases, if there was a V–deletion allomorphy (i.e. *καρδιά~καρδι* (karδia~karδi) ‘heart’), the system detected only the V–deleted stem (καρδι–) considering the deleted vowel as a suffix. Moreover, if there was a C–insertion allomorphy (i.e. *κύμα~κυματ* (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.να.νες.νων* (σωλη (soli), πυρη (piri), αιω (eo), λιμε (lime)) and *γα.ξα* (ανοι (ani), τυλι (tili), διαλε (δiale)). 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 indeed. Specifically, among the signatures produced by the analysis of our ‘bound–stem corpus’, we found ‘real’ suffixes, such as the derivational *–ία* (e.g. *θεολογ–ία* (θeologγia) ‘theology’), *–είο* (e.g. *ανθοπωλ–ειο (*anθopolio) ‘flower shop’), *–τη(ς)* (e.g. *αιμοδό–της* (emoδotis) ‘blood donor’), *–ισσα* (e.g. *παλαιοπώλ–ισσα* (paleopolissa) ‘female antique seller’), the nominal inflectional *(ος)* (e.g. *βοτανολόγ(ος)* (votanologγos) ‘votanologist’), *(ης)* (e.g. *πατριάρχ(ης)* (patriarchis) ‘patriarch’), and the verbal inflectional *(ω)* (e.g. *ηχογραφ(ώ)* (ixografo) ‘sound record’). However, we also found sequences like *–ολόγος* (–ologγos), *–ολογία* (–ologγia), *–ογράφος* (–γrafos), *– ογραφία* (–γrafia), *–ομανής* (–omanis), *–ομανία* (–omania), *–οποιία* (–opiia), *–οποιείο* (–opiio), *–οτρόφος*) (–otrofos), *–οτροφία* (–otrofia), *–όφιλος* (–ofilos), *–ορραγία* (–orraγia), *–ογονία* (–oγonia), *–οστάτης* (–ostatis), *–οφαγία* (–ofaγia), *–οκτονία* (–oktonia), which basically look like confixes with the linking element attached to them. At the same time, and for no obvious reason, among the signatures, we found sequences like *–φάγος* (–faγos), *–φαγία* (–faγia), *–παθής* (–paθis), *–σκοπία* (–skopia), *–σκόπιο* (–skopio), *–ούχος* (–uxos), *–γενής* (–γenis), *–γονία* (–γonia), *–μαθής* (–maθis), *–άρχης* (–arxis), *–φόρος* (–foros), *–πρεπής* (–prepis), *–τέχνης* (–texnis), which are also confix–like but without the element –o– attached. Results like these, imply that the system did not manage to recognize neither the linking element as a separate entity, nor the derivational or inflectional suffixes attached to the final bound elements.

Reasonably enough, the recognition of a great number of confix–like sequences with the linking element

[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.

attached, as those metioned above, gave rise to a great number of 'correct' stems[4] of MG like *miθ–* ('myth'), *okean–* ('ocean'), *selin–* ('moon'), *musik–* ('music'), *xart–* ('paper'), *siδir–* ('iron') or stem allomorphs like *δramat–* ('drama'), *xromat–* ('colour'), *θavmat–* ('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–*, *karkinoγon–*, *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,–maθ, –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.

[4] i.e. without their inflectional ending as they normally appear in compounds.

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.

Anastasiadi–Symeonidi, A. (1986). *Neology in Common Modern Greek* [In Greek]. Thessaloniki.

Booij, G. (1997). Allomorphy and the Autonomy of Morphology, *Folia Linguistica XXXI/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. Saraclar (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.

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.
Καρασίμος, 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 Κυδωνίων και Μοσχονησίων: η παραγωγή λέξεων*, Διδακτορική διατριβή, Πανεπιστήμιο Πατρών.
Ράλλη, Α. (2005). *Μορφολογία*. Αθήνα: Πατάκη.