Allomorphy and Greek Nominal Compounds: a computational prediction and analysis[1]

Athanasios Karasimos[1] & Georgios Markopoulos[2]
[1]Academy of Athens, [2]National and Kapodistrian University of Athens
akarasimos@academyofathens.gr, gmarkop@phil.uoa.gr

Abstract

Η υπολογιστική επεξεργασία της αλλομορφίας εξακολουθεί να αποτελεί τεράστια πρόκληση από τις πρώτες συστηματικές προσπάθειες πρόβλεψης της αλλομορφίας με τεχνικές μηχανικής μάθησης. Το μοντέλο MaxEnt προσφέρει έναν στατιστικό τρόπο για να δημιουργήσετε ένα πιθανοτικό μοντέλο για SOI που συνδυάζει διαφορετικά γλωσσικά στοιχεία. Στόχος είναι να προβλέψουμε τις αλλομορφικές αλλαγές στην ονοματική σύνθεση και να δείξουμε την ουσιαστική συμβολή διαφόρων μορφολογικών, φωνολογικών και σημασιολογικών χαρακτηριστικών. Για την αξιολόγηση της αποτελεσματικότητας του μοντέλου μας, χρησιμοποιήθηκε ένα δοκιμαστικό σώμα με ονοματικά σύνθετα που έχουν οποιοδήποτε είδος γραμματικής κατηγορίας ως πρώτο συνθετικό. Δημιουργήσαμε τον ALLOMANTIS, έναν αναλυτή μορφολογικής πρόβλεψης για την ονομαστική αλλομορφία. Η συνολική ακρίβεια του μοντέλου ήταν πάνω από 90%.

Keywords: Μέγιστη Εντροπία, Επιβλεπόμενη μορφολογική μάθηση, Αλλομορφία, Σύνθεση, Δεσμευμένα θέματα, AlloMantIS

1 Introduction

Interest in the topic of stem allomorphy has been renewed by Mark Aronoff's (1994) work, which led to novel descriptions of inflectional and derivational phenomena in work by Booij (1997), Thornton (1997), Pirrelli and Battista (2000a,b), Maiden (2004), Stump (2001), among others. The main aim of Aronoff's work and later research is the notion that the significance of a lexeme is not a single phonological representation, but an array of indexed stems, which may stand in relations ranging from identity through semiregular/ irregular phonological alternation to full suppletion. It is pointed out that, beyond the theoretical challenges of the phenomenon, allomorphy remains a serious problem for morphological parsing that must be solved.

On the other hand, the goal of Computational Morphology is to create programs, which can produce an output that matches as closely as possible the analysis that would be given by a morphologist. More specifically, an Unsupervised Morphology Learning Model (UMLM) only accepts as input huge corpora and tools for analysis, without the use of a lexicon and morphological (or phonological) rules of a particular language (Harris 1955, 1967, Hafer and Weiss 1974, Goldsmith 2001). As part of the criticism of Unsupervised Morphology Learning Models for their failure to deal with Greek allomorphy, Karasimos (2009) has argued that probably only a supervised morphology learning model is more likely to successfully face allomorphy. The computational treatment of allomorphy still is a huge challenge since the first systematic attempts on

[1] This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Reinforcement of Postdoctoral Researchers" (MIS-5001552), implemented by the State Scholarships Foundation (IKY).

predicting allomorphy with machine learning techniques (Rumelhart and McLelland 1986, Pinker and Prince 1988, Ling and Marinov 1993 among others).

2 Allomorphy and nominal compounding

This study is couched in a theoretical framework centered on the morpheme, treating allomorphy as a morphological phenomenon which places derivation on a separate level of the word formation process. Comparing the processes of compounding and derivation through the prism of allomorphy, we can observe various tendencies between languages. There are languages, such as German, where all the allomorphs of a clitic example participate in their production and synthesis, while in other languages, such as Dutch and Greek, the above behavior does not exist.

More specifically, analyzing the data from the nominal compounds of Modern Greek, we discover that all the forms of a morpheme are not fully available depending: (a) on the position within the compound as first or second component, as well as (b) in its form as a stem or a word. For example, the noun "κύμα" ('wave') displays two allomorphs *κυμα~ κυματ* in inflection, but only one allomorph appears as the first component (*κυματ-*), e.g., κυματ-ο-μορφή 'waveform', κυματοθραύστης 'wave breaker', κυματοσυνάντηση 'wave function'. Furthermore, as a second component, the allomorphic pattern changes based on its structure (stem vs. word; see 4.3). The same allomorphic nominal pattern is observed in the derivation. As evidenced in the following subsection, this allomorphic behavior is not random. It is related to the aforementioned constraint and it applies unexceptionally to all nominal compounds.

Expanding Lieber's (1982) definition, we define allomorphy as the study of the different variants of a lexeme, which share lexical information and semantic representation. However, they differ simultaneously in their phonological form unpredictably and arbitrarily due to the application of some diachronic phonological or morphological rule. We argue that it is a process in which the morphological environment and the choice of the appropriate allomorph are characterized by regularity and predictability (Karasimos 2011).

Ralli (2000, 2007) underlines that allomorphy participates in the core morphology and without exception in all word formation processes. She suggests that it is one of the main features of verbal and nominal categorizing into inflectional classes. For example, ποιητη~ποιητ ('poet', 2nd class), καφε~ καφεδ ('coffee', 3rd class) and βημα~ βηματ ('step', 8th class). Following Lieber's (1982) and Ralli's (2000, 2007) theoretical model we do not consider as allomorphs any kind of changes resulting from phonological rules (*phonomorphs*, as ράβ-ω – έ-ραψ-α, 'I sew – I sewed'), free variants (ψάλτ-ες vs. ψαλτάδ-ες, 'chanter'), and suppletions (είμαι – υπήρξα, 'I am – I was').

3 Maximum entropy and morphology learning

Maximum Entropy aims to determine the set of statistics that can capture the behavior of a random process, i.e. the feature selection of our training data. Then given all these statistics, the second objective is to include these features in a precise process model –a model that can predict the future exported processing– i.e. the final choice of this model. According to the supporters of MELA all the known and unknown, regular and irregular words are treated together with the same strategy, since they are another feature in the general model of probability. This strategy offers great potentials to treat

allomorphy, which is considered as something irregular, as a marginal synchronic junk pile and a relic

The MaxEnt framework offers a mathematically sound way to build a probabilistic model for Subject-Object Identification (SOI) which combines different linguistic features. Dell' Orletta et al. (2007) research uses constraints on the prediction of subject and object in Italian and Czech by resorting to the technique of Maximum Entropy. Inspired by their results, we attempt to test a model for the Greek allomorphy in nominal compounding. Our goals are to predict the allomorphic changes and to show the essential contribution of various morphological, phonological and semantic characteristics. The aim of this model is to identify the weight of these characteristics that are directly dependent on allomorphy, to help design a predictive model. This model is not only destined for nominal compounding allomorphy, but also for nominal inflectional and derivational allomorphy.

4 The AMIS experiment for nominal compounding and allomorphy

4.1 Introduction

A great challenge of Natural Language Processing applications dealing with morphologically-rich languages, such as Greek, German, Dutch, Norwegian, Swedish or Danish is the successful processing of their compound words. “These languages are very productive in the creation of new compounds, as they may concatenate several words together into a single typographic word at any time” (Escartín 2014: 3340). For this demanding task, the MaxEnt framework offers a mathematically sound way to build a probabilistic model for *SOI* which combines different linguistic cues. Our goals are to predict the allomorphic changes and to show the essential contribution of various morphological, phonological and semantic characteristics. The aim of this model is to identify the weights of these characteristics that are directly dependent on allomorphy, to help design a predictive model. This model is not only destined for nominal stem allomorphy, but also for nominal derivational allomorphy. Our model is based on AMIS, which is a parameter estimator for maximum entropy models (Berger, Della Pietra and Della Pietra 1996), is freeware and benefits from linguistic feature sets (Yoshida 2006); given a set of events as training data, the program outputs parameters that optimize the likelihood of the training data. AMIS is a parameter estimator for maximum entropy models. Given a set of events as training data, the program outputs parameters that optimize the likelihood of the training data.

4.2 Morpho-phonological interpretations as feature sets

Karasimos’ (2011) research revealed that nominal allomorphy ‘represents’ usually relics of non-active phonological and morphological rules and changes in a Greek language. Therefore, we make the assumption that also the nominal compounds give the necessary information to a system with minimal supervision to predict the appearance (or not) and type of allomorphy. We maintain that the stochastic models seem to be more suitable to satisfy the requirements of a model with linguistic feature sets. These characteristics are functions type-f_{xn} (λ,Σ), where a particular item χ_i is tested for the word-attribute λ, which is included in a feature set Σ. For this MaxEnt model, we chose different types of features that contain morphological, phonological and

semantic dimensions of the distributions of nominal allomorphy (in allomorphic classes *ACx*).

Our characteristics are 10 which are different from the initial model (for more information, see Karasimos 2011). For obvious reasons some of these characteristics are empty for the training data, since inflected and derived nouns do use a linking element or the inflected nouns do not include any information about derivational suffixes, bound stems or headness.

i. *Allomorphic Class* (8 classes of different nominal allomorphic behavior), as the main characteristic that is under survey to discover the connection with the other characteristics,
ii. *Inflectional Class* (8 classes based on Ralli's (1994) model; two for masculine nouns, two for feminine nouns and four for neutral nouns),
iii. *Genre* (masculine, feminine, neutral)
iv. *Linking element* (yes, no, alternative),
v. *Derivational suffix* (yes, no),
vi. *Bound stem* (ye, no),
vii. *Syllables* (up-to-6 syllables),
viii. *Stress* (3 levels – ultimate, penultimate, antepenultimate),
ix. *Last characters* (up-to-4 characters),
x. *Headeness* (no, endocentric, exocentric, dvandva)

4.3 Training and test data

This Greek model of maximum entropy was trained on a corpus of 4,677 inflected nouns (neither derived nor compound nouns) and 2,755 derived nous, a sufficient sample of all eight inflectional classes and a significant sample of all available nominal derivational suffixes. Training data contain inflected nouns (stem and inflectional suffixes) and derived nouns (stem and nominal derivational suffixes) that are synchronically morphological transparent. Based on electronic version "Λεξικό της Κοινής Νεοελληνικής"[2] of Triantafyllidis, the printed lexicon "Χρηστικό Λεξικό της Νεοελληνικής Γλώσσας" of Academy of Athens, the printed lexicon "Λεξικό της Νέας Ελληνικής Γλώσσας" (5th Edition, 2019) by Mpampiniotis, and on the neologisms list from ΝεοΔημία corpus (Χριστοφίδου κ.ά. 2013). Ηλεκτρονικό πρόγραμμα Νεολογισμών και Ορολογίας ΝΕΟΔΗΜΙΑ: Παρουσίαση και Προκλήσεις. Στο Α. Χριστοφίδου (εκδ.) *Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών (Δημιουργία και Μορφή στη Γλώσσα)*, Τεύχος 12ο, σσ. 198-243. Αθήνα: Ακαδημία Αθηνών – Κέντρον Ερεύνης Επιστημονικών Όρων και Νεολογισμών.), all the nouns were manually imported and every feature of the model was checked with the help of these dictionaries. From our data only 34,5% of nouns do not display allomorphy; therefore, the amount of allomorphs is quite high in the Greek language. AMIS produced weights for more than 38,000 features. It is expected that a model with more than 15,000 features for weights is quite heavy statistically, since the combinationality of syllables and characters increased exponentially the size of our sets.

[2] http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/

πίνακας
0 IC_second!1 Genre_masculine!1 Syllables_three!1 Stress_antipenultimate!1 Syllable1_πι!1 Syllable2_να!1 Syllable3_κας!1
1 IC_second!2 Genre_masculine!2 Syllables_three!2 Stress_antipenultimate!2 Syllable1_πι!2 Syllable2_να!2 Syllable3_κας!2
0 IC_second!3 Genre_masculine!3 Syllables_three!3 Stress_antipenultimate!3 Syllable1_πι!3 Syllable2_να!3 Syllable3_κας!3
0 IC_second!4 Genre_masculine!4 Syllables_three!4 Stress_antipenultimate!4 Syllable1_πι!4 Syllable2_να!4 Syllable3_κας!4
0 IC_second!5 Genre_masculine!5 Syllables_three!5 Stress_antipenultimate!5 Syllable1_πι!5 Syllable2_να!5 Syllable3_κας!5
0 IC_second!6 Genre_masculine!6 Syllables_three!6 Stress_antipenultimate!6 Syllable1_πι!6 Syllable2_να!6 Syllable3_κας!6
0 IC_second!7 Genre_masculine!7 Syllables_three!7 Stress_antipenultimate!7 Syllable1_πι!7 Syllable2_να!7 Syllable3_κας!7
0 IC_second!8 Genre_masculine!8 Syllables_three!8 Stress_antipenultimate!8 Syllable1_πι!8 Syllable2_να!8 Syllable3_κας!8
καραβόσχοινα
0 IC_sixth!1 Genre_neutral!1 Syllables_five!1 Stress_antipenultimate!1 LinkingElement_yes!1 Headness_endo!1 DerivSuffix_no!1
0 IC_sixth!2 Genre_neutral!2 Syllables_five!2 Stress_antipenultimate!2 LinkingElement_yes!2 Headness_endo!2 DerivSuffix_no!2
0 IC_sixth!3 Genre_neutral!3 Syllables_five!3 Stress_antipenultimate!3 LinkingElement_yes!3 Headness_endo!3 DerivSuffix_no!3
0 IC_sixth!4 Genre_neutral!4 Syllables_five!4 Stress_antipenultimate!4 LinkingElement_yes!4 Headness_endo!4 DerivSuffix_no!4
0 IC_sixth!5 Genre_neutral!5 Syllables_five!5 Stress_antipenultimate!5 LinkingElement_yes!5 Headness_endo!5 DerivSuffix_no!5
0 IC_sixth!6 Genre_neutral!6 Syllables_five!6 Stress_antipenultimate!6 LinkingElement_yes!6 Headness_endo!6 DerivSuffix_no!6
0 IC_sixth!7 Genre_neutral!7 Syllables_five!7 Stress_antipenultimate!7 LinkingElement_yes!7 Headness_endo!7 DerivSuffix_no!7
1 IC_sixth!8 Genre_neutral!8 Syllables_five!8 Stress_antipenultimate!8 LinkingElement_yes!8 Headness_endo!8 DerivSuffix_no!8
κυματοθραύστης
0 IC_second!1 Genre_masculine!1 Syllables_five!1 Stress_penultimate!1 LinkingElement_yes!1 Headness_exo!1 DerivSuffix_yes!1
1 IC_second!2 Genre_masculine!2 Syllables_five!2 Stress_penultimate!2 LinkingElement_yes!2 Headness_exo!2 DerivSuffix_yes!2
0 IC_second!3 Genre_masculine!3 Syllables_five!3 Stress_penultimate!3 LinkingElement_yes!3 Headness_exo!3 DerivSuffix_yes!3
0 IC_second!4 Genre_masculine!4 Syllables_five!4 Stress_penultimate!4 LinkingElement_yes!4 Headness_exo!4 DerivSuffix_yes!4
0 IC_second!5 Genre_masculine!5 Syllables_five!5 Stress_penultimate!5 LinkingElement_yes!5 Headness_exo!5 DerivSuffix_yes!5
0 IC_second!6 Genre_masculine!6 Syllables_five!6 Stress_penultimate!6 LinkingElement_yes!6 Headness_exo!6 DerivSuffix_yes!6
0 IC_second!7 Genre_masculine!7 Syllables_five!7 Stress_penultimate!7 LinkingElement_yes!7 Headness_exo!7 DerivSuffix_yes!7
0 IC_second!8 Genre_masculine!8 Syllables_five!8 Stress_penultimate!8 LinkingElement_yes!8 Headness_exo!8 DerivSuffix_yes!8

Table 1 | Sample from training and test data (!NUMBER is the corresponding allomorphic class and 0/1 in each line is the true/false value for the proper allomorphic class)

To evaluate the effectiveness of our model, a testing corpus with nominal compounds that have any kind of stem, at least one nominal derivation suffix (in the rightest part of the word) and an inflectional suffix, was created. This second corpus contains 2,884 carefully selected nominal compounds from the aforementioned sources and 671 neoclassical nominal bound stems from Πετροπούλου's (2012) doctoral dissertation list. We created ALLOMANTIS v2[3], an updated morphological prediction analyzer for nominal allomorphy, which takes an input imported data from our training corpora on AMIS. ALLOMANTIS replaces each word characteristic with the proper weight given by the training corpus from AMIS. The analyser multiplies the weights of all attributes for each candidate allomorphic class and proceeds with the one with the largest result of multiplication; according to the model of maximum entropy, this is the winner and is identified by the ALLOMANTIS as the proper allomorphic class.

AC5	**Positive weights**		**Negative weights**	
	Syllable4_λης	7,27E+02	Syllable2_πα	3,78E-01
	Syllable2_δες	1,02E+02	Syllable1_α	3,74E-01
	Syllable1_γιορ	5,62E+01	Syllable2_πι	3,66E-01
	Syllable3_νης	4,24E+01	BoundStem_yes	3,45E-01
	Syllable3_πης	4,11E+01	LinkingElement_no	3,44E-01
	Syllable3_για	3,69E+01	Syllable1_λα	3,42E-01
	Syllable3_κης	2,74E+01	Character2_υ	2,23E-01
	Syllable2_ριας	2,68E+01	Syllable3_κας	1,40E-01
	Syllable2_σπο	2,00E+01	Stress_antipenultimate	4,99E-02
	Syllable4_γας	1,89E+01	Stress_penultimate	8,16E-02
AC8	**Positive weights**		**Negative weights**	
	Syllable2_λης	4,59E+01	Syllable2_ρα	3,00E-01
	Syllable2_ης	4,54E+01	Character4_ν	2,96E-01
	Syllable2_γης	4,12E+01	DerivationalSuffix_no	2,93E-01
	Syllable2_ρης	2,98E+01	Character3_σ	2,34E-01
	Syllable2_ντζε	2,79E+01	Character4_ρ	2,03E-01
	Syllable1_ερ	2,63E+01	Character4_λ	1,84E-01
	Syllable2_τερ	2,35E+01	Stress_ultimate	1,37E-01

[3] The blend ALLOMANTIS is a combination of 'αλλομορφία' (allomorphy) and 'μάντης' (seer, prophet) and the capital letters refer to the initials of the maximum entropy program AMIS.

	Syllable1_μπου	1,70E+01	Boundstem_yes	1,36E-01
	Syllable2_της	1,69E+01	Character4_α	1,10E-01

Table 2 | Sample of positive and negative weights of our AMIS features.

4.4 Results

The overall accuracy of the model was 81.37% with the failure rate up to 18.63%. A detailed analysis of the model for each allomorphic class is shown that the weaknesses were between two specific allomorphic group. More than 90% was achieved in several classes, as in the AC1 94.44%, AC2 96.11% and AC4 90.9%, whereas the two classes with the lowest percentage was AC5 (50.18%) and AC8 (26.85%), with the latter rates considered to be a strong flaw of (from) the average success.

To improve the system, we tried a more rational approach to achieve a better performance. In the previous experiment of the AlloMantIS, we observed an improvement when we reversed the syllables numbering. while in the updated version we followed the stress strategy for spelling, i.e, the ultimate syllable was numbered as first, the penultimate as second and so forth. The result of the upgraded version of ALLOMANTIS was the rise of the correct prediction (Recall: 93,76%, Precision: 95,02%). Indeed, the first four allomorphic classes reached 100%, but AC8 remained in a tragically low threshold (31.22%), as well as AC5 with 58.32% of erroneous estimations, since both of them are similar cases of loan nominal components that have a slightly different allomorphy in inflection (two allomorphs, i.e. ντενεκε~ ντενεκεδ 'tin' vs. three allomorphs, i.e. μαναβη~ μαναβηδ~ μαναβ 'grocer). Actually, we made a third attempt with our data by minimizing the characteristics dataset into three features (stress, 3 last syllables and inflectional class). It was impressive the system prediction performance (Recall: 90,82%, Precision: 92,11%), since this kind of word annotation can be built automatically.

5 The alternative route: Compound splitting from Translation-verse

It raises the question how we are going to deal with the allomorphy of the nominal first components. In natural language processing applications (particularly in MT), a rather non-compositional, morpheme-based approach prevails, since compound structures are processed through splitting and merging, without however overshadowing the validity of compositional theories. Henrich and Hinrichs (2011) report that there are a number of morphological tools available that include compound splitting, such as GERTWOL (Haapalainen and Majorin 1995), SMOR (Schmid et al. 2004), ASV Toolbox (Witschel and Biemann, 2005), BananaSplit 3, and Morfessor (Creutz and Lagus, 2002). For the GermaNet project, they created a hybrid combined compound splitter (its performance was almost 95% correct prediction) that takes into account all knowledge provided by the individual compound splitters, but that also takes into account some domain knowledge about German derivation morphology and compounding.

The splitting of compound words into their constituents is "borrowing" strategy from machine translation that can actually provide some reasonable solutions. Different theoretical approaches have been developed throughout the years on compound splitting, based either on the frequencies of the substrings in particular corpora (Koehn and Knight 2003), or on linguistic knowledge through, for example, part-of-speech constraints (Stymne 2008). The split compounds follow in the categories below:

a) Correct splits (words that should be split and were split correctly)
b) Correct non splits (words that should not have been split and were not split)
c) Wrong non splits (words that should have been split but were not split)
d) Faulty splits (words that should be split but were split wrongly)
e) Wrong splits (words that should not have been split but were split)

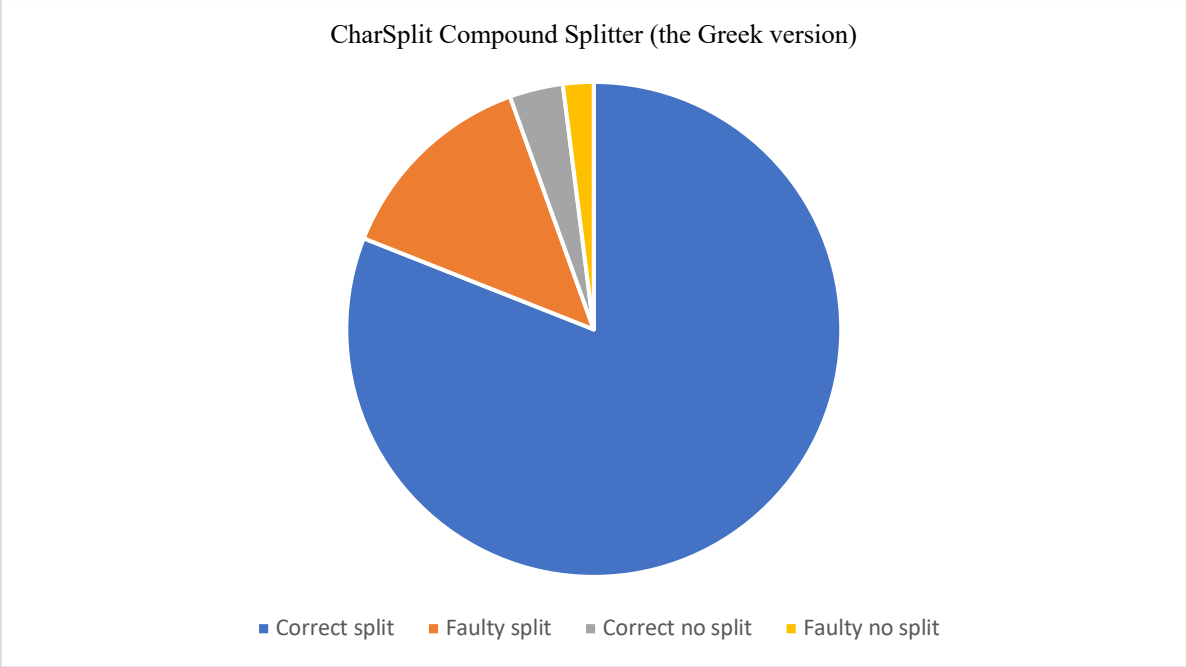
The important factor of the linking element may shift the balance between success and failure. It can be used as a parsing boundary that should split easily the two components. For these reasons the need of training data from inflected nouns (absence of derivation affixes) is crucial.

Using corpus from AMIS experiment, we created a sub-corpus with 1,125 nominal compounding data but without the features, but with the necessary morphological parsing (both first and second component are nominal, linking element [yes]). We chose randomly 925 compounds as training data and the rest (200) were the testing data. We modified Tuggener's (2016) CharSplit[4], whose method achieved ~95% accuracy for head detection on the Germanet compound test set. CharSplit compound splitter returns a list of all possible splits, ranked by their score, e.g[5].

[(0.6157458641452111, 'κουκλο', 'σπιτο'),
(-0.3245885423784691, 'κουκλ', 'οσπιτο'),
(-4.3845574213895133, 'κου', 'κλοσπιτο'),...] κουκλόσπιτο ('dollhouse')
[(0.8845962378456548, 'μπακαλο', 'γατος'),
(-0.2400556587892472, 'μπακαλ', 'ογατος'),
(-1.0045581132387926, 'μπακα', 'λογατος'),...] μπακαλόγατος ('stock boy')
[(0.7131548946548794, 'γυρο', 'λογος'),
(0.1256244876462498, 'γυρ', 'ολογος),
(-0.731596723556389, 'γυρολο', 'γος'),...] γυρολόγος ('peddler, chapman')

The distinction between lemmas and word forms made by the splitter has not been taken into consideration to allow for a proper grouping between the first components and if it possible to identify the allomorphs (including the linking element). The results show that the highest score generally is the correct split for the NN compounds, regards to precision, when running splitting tasks in a corpus with proper training data annotation. Nevertheless, the effort to lemmatize the first nominal component with a second nominal component of the same lemma had several issues due the absence of any training data with all the inflected forms and allomorphs. As far as corpus size is concerned, it can be acknowledged that it is very small and in the case of corpus-based compound splitters, it does have an impact in the overall scores. In future work, we expect the output of this modified splitter will be improved with more NV and NA compounds and a solution of the linking element detection. Additionally, the allomorphy detection for the first nominal components is definitely not satisfactory, but this is strongly connected with the absence of annotated Greek corpus with inflectional features. Our sub-corpus was quite small and simple; nevertheless, for a modified splitter version, the allomorphy prediction was above the baseline.

[4] https://github.com/dtuggener/CharSplit
[5] No attempt to identify the linking element for this compound splitter modification.



Graph 1 | The results from the test data based on four major categories

	Precision	Recall	Accuracy
NN Compounds	**84.87%**	71.29%	82.14%
NN Bound stems	**87.36%**	80.92%	86.37%
Connected forms	**48.63%**	32.33%	40.29%

Table 3 | Evaluation of the performance of the splitter and the effort of forms-to-lemma connection

6 Concluding remarks

It is noteworthy that our model was trained by a corpus of inflected and derived nouns (not created by the process of derivation and compounding) and evaluated by a corpus with nominal compounds, since we tried to make our task more difficult. This choice was not arbitrary based on Karasimos' (2011) argument that the nominal derivational suffixes display similarities with nominal stems/ roots, participate in the same inflectional classes and thus exhibit the same allomorphic behaviour. ALLOMANTIS correctly predicted allomorphy for more than 95% of the nominal compounds of the testing corpus. It is expected that if ALLOMANTIS is trained with a corpus of inflected and derived nouns, then the prediction accuracy rate will be much higher. It was considered necessary in this primary testing stage of our model to provide a minimal help from the training corpus. Moreover, we 'borrowed' a machine-translation technique to split the NN compounds and connected the forms to detect the allomorphs. The modified version of CharSplit proved quite efficient since the precision was above 85% considering the small training and testing sub-corpora.

Extending this reasoning means that certain morphological phenomena or processes can be a result of a combinatorial analysis of morphological features that are (sometimes) assisted from data of other language (phonology, semantics, etc). In these experiments, it is inferred how essential the existence of morphologically annotated corpora is for the effective conduct of morphological experiments in Greek. We have

shown that a (supervised) probabilistic model applied to a corpus with quite rich annotated words can extract some basic principles that can be the keystone to construct a computational model to process the "unpredictable" and hard-to-deal phenomenon of allomorphy. The results of the third attempt provide some significant and promising results that automatically annotated morphological corpora can provide all the necessary information for quite successful parsing of Modern Greek data.

References

Aronoff, Mark. 1994. *Morphology by Itself: Stems and Inflectional Classes.* Linguistic Inquiry Monograph 22. Cambridge, MA: MIT Press.

Berger, Adam, Stephen Della Pietra, and 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.

Creutz, Mathias and Lagus Krista. 2002. "Unsupervised discovery of morphemes." In Pierre Isabelle, Eugene Charniak and Dekang Lin (Eds.) *Proceedings of the ACL-2002 Workshop on Morphological and Phonological Learning*, 21-30. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Dell' Orletta, Felice, LENCI Alessandro, MONTEMAGNI Simonetta & PIRRELLI Vito 2007. Corpus-based modeling of grammar variation. In A. Sansò (ed.) Language Resources and Linguistics Theory, pp. 38-55.[Materiali Linguistici 59]. Milano: Franco Angeli.

Escartín, Carla Parra. 2014. "Chasing the Perfect Splitter: A Comparison of Different Compound Splitting Tools." In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3340–3347, Reykjavik, Iceland: European Language Resources Association (ELRA).

Goldsmith, John. 2001. "Unsupervised Learning of the Morphology of a Natural Language." *Computational Linguistics* 27(2):153-198.

Hafer, Margaret & Stephen Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10, pp. 371-385.

Haapalainen, Mariikka, and Ari Majorin. 1995. *Gertwol: Ein System zur automatischen Wortformerkennung deutscher Wörter*. Technical report, Lingsoft Inc.

Harris, Zellig. 1955 [1970]. From phoneme to morpheme. *Language* 31, pp. 190-222.

Harris, Zellig 1967 [1970]. Morpheme boundaries within words: Report on a computer test. *Transformations and Discourse Analysis Papers* 73**.** University of Pennsylvania.

Henrich, Verena, and Erhard Hinrichs. 2011. "Determining Immediate Constituents of Compounds in GermaNet." In *Proceedings of Recent Advances in Natural Language Processing*, 420–426. Hissar, Bulgaria: INCOMA Ltd. Shoumen.

Karasimos, Athanasios. 2009. "Comments on Unsupervised Morphology Learning Model: The case of Greek Allomorphy." Paper presented at the *19*th *International Symposium in Theoretical and Applied Linguistics* (ISTAL19). 3-5 April 2009, Thessaloniki.

Karasimos, Athanasios (2011). Computational Processing of Allomorphy in word derivation of Modern Greek. *Journal of Greek Linguistics 11*, pp. 286-292. Leiden: Brill Publications.

Koehn, Philipp and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In 10th Conference of the European Chapter of the Association for Computational Linguistics. Budapest, Hungary: Association for Computational Linguistics.

Lieber, Rochelle. 1982. "Allomorphy." *Linguistic Analysis* 10(1):27-52.

Ling, C. X. & Marin Marinov. 1993. Answering the connectionist challenge: a symbolic model of learning the past tenses of English verbs. *Cognition 49* (3):235-290.

Maiden, Martin. 2004. "When lexemes become allomorphs – On the genesis of suppletion." *Folia Linguistica* XXXXCIII(3-4):227-256.

Πετροπούλου, Ευανθία. 2012. *Σύνθεση με δεσμευμένο θέμα στην Αγγλική και τη νέα Ελληνική : θεωρητική ανάλυση και υπολογιστική επεξεργασία*. Διδακτορική διατριβή. Πάτρα: Πανεπιστήμιο Πατρών.

Pinker, Steven & Alan Prince. 1988. On language and connectionism Analysis of a parallel distributed processing model of language acquisition. In S. Pinker & J. Mehler (Eds.) *Connections and Symbols (Special Issue: Cognition)*, pp. 73-193. Cambridge, MA: MIT Press.

Pirrelli, Vito, and 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, and 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." In A. Kakouriotis (Ed.) *Proceedings of the 8th Symposium on English and Greek Linguistics,* 19-46. 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 Lesvos, Kydonies and Moschonisia." In G. Sica (Ed.) *Open problems in Linguistics and Lexicography*, 123-153. Milano: Polimetrica.

Roark, B. and Sproat, Richard. 2007. *Computational Approaches to Morphology and Syntax*. Oxford: Oxford University Press.

Rumelhart, David & James McClelland. 1986. On learning the past tense of English verbs. In J. McClelland & D. Rumelhart (Eds.) *Parallel Distributed Processing*, Volume 2, pp. 216-271. Cambridge, MA: MIT Press.

Stymme, Sara. 2008. German compounds in factored statistical machine translation. *International Conference on Natural Language Processing*. Berlin, Heidelberg:
Springer.

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 G. Booij, S. Scalise and A. Ralli (Eds.) *Proceedings of the 1st Mediterranean Meeting of Morphology,* edited by 86-98. Patras: University of Patras.

Tuggener, Don. 2016. *Incremental Coreference Resolution for German*. University of Zurich, Faculty of Arts.

Χριστοφίδου, Αναστασία, Αφεντουλίδου, Βασιλική, Καρασίμος, Αθανάσιος & Ειρήνη Δημητροπούλου. 2013. Ηλεκτρονικό πρόγραμμα Νεολογισμών και Ορολογίας ΝΕΟΔΗΜΙΑ: Παρουσίαση και Προκλήσεις. Στο Α. Χριστοφίδου (εκδ.) *Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών (Δημιουργία και Μορφή στη Γλώσσα)*, Τεύχος 12ο, σσ. 198-243. Αθήνα: Ακαδημία Αθηνών – Κέντρον Ερεύνης Επιστημονικών Όρων και Νεολογισμών.

Yoshida, Kazuhiro. 2006. AMIS – A maximum entropy estimator for feature forests. AMIS Manual. Department of Computer Science, University of Tokyo. Available at: http://www-tsujii.is.s.u-tokyo.ac.jp/amis/manual.html.