Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3861–3867
Marseille, 11–16 May 2020

A Supervised Part-Of-Speech Tagger for the Greek Language of the Social Web

Maria Nefeli Nikiforos, Katia Lida Kermanidis
Department of Informatics, Ionian University
Corfu, Greece
c19niki@ionio.gr, kerman@ionio.gr

Abstract

The increasing volume of communication via microblogging messages on social networks has created the need for efficient Natural Language Processing (NLP) tools, especially for unstructured text processing. Extracting information from unstructured social text is one of the most demanding NLP tasks. This paper presents the first part-of-speech tagged data set of social text in Greek, as well as the first supervised part-of-speech tagger developed for such data sets.

Keywords: supervised part-of-speech tagging, social web language, Greek

1. Introduction

The increasing volume of communication via microblogging messages on social networks (such as Facebook, Twitter, Blogs and YouTube) has created the need for developing efficient Natural Language Processing (NLP) tools, especially for unstructured text processing. It has been observed that the performance of NLP tools developed for structured text processing is significantly reduced when used in unstructured microblogging text (Nand et al., 2014). Therefore, there is a growing demand for information extraction tools from unstructured social text for business intelligence, security, programming etc. However, extracting information in this context is one of the most demanding NLP tasks due to the unusual structure of the text (Nand et al., 2014). As a result, it is necessary to either adapt existing methods for implementation in this context, or to develop new methods that perform well on unstructured microblogging text.

Part-of-speech tagging (POS tagging) is a fundamental part of NLP (Gimpel et al., 2010). A robust POS tagging tool has an important role in most NLP problems and applications, such as syntactic and semantic analysis, machine translation and sentiment analysis (Bach et al., 2018; Liu et al., 2012). The main challenges in POS tagging are: a. creating the tag set, based on which each word will be labeled with a POS tag, which needs to include the specifics of the social network context, b. labeling ambiguous words with the correct tag, and c. creating a labeled data set of sufficient size for the best possible results in machine learning.

Despite the development of certain NLP tools for conventional text in Greek (Papageorgiou et al., 2000; Petasis et al., 2001; Sgarbas et al., 2001), there are not any POS taggers for social text in Greek. The contribution of our research is: a. creating the first annotated data set for Greek social text (2,405 tweets or 31,697 tokens), b. creating the first tag set (22 different tags) including special tags for Twitter language specifics, and c. developing the first supervised POS tagger for such data with significantly high performance (accuracy up to 99.87%). The data set is available for research purposes at this address: `https://hilab.di.ionio.gr/index.php/en/datasets/`

The rest of this paper is structured as follows. Section 2 describes related work, with various approaches and POS taggers for processing microblogging data from social networks written in several languages. Section 3 presents the methodology for the creation, preprocessing and annotation of the data set, as well as the methodology for the creation of the tag set. Section 4 describes various machine learning experiments with the developed POS tagger, and analyzes their results. Finally, Section 5 concludes the paper and discusses future work.

2. Related Work

There are some tools for POS tagging of conventional text in Greek. The most well known are the "Greek POS Tagger" (`https://bit.ly/3438zjl`), developed by the NLP Team of the Department of Informatics, Athens University of Economics and Business, and the "Greek part of speech tagger" (`https://bit.ly/37leoL0`), based on Coleli's, E. thesis. Furthermore, Sgarbas et al. (2001) used POS tagging as a part of morphological analysis for their tests. They used detailed tags. Petasis et al. (2001) developed a word-based morphological analyzer to use as a basic feature of a spelling checker. Each word's context was considered (POS tagger accuracy 95%). Another POS tagger by Papageorgiou et al. (2000) used a data set that consisted of 447,000 tokens from financial reports, news agencies, and instruction manuals, and 584 tags (accuracy 96.28%). In a recent research, Keersmaekers (2019) attempted to automatically annotate the entire corpus of Ancient Greek scrolls. However, there are not any POS taggers for social text in Greek.

One of the first POS taggers for social text in English is that of Gimpel et al. (2010). They developed a POS tagger for data from Twitter, by creating an annotated data set (1.8K tweets or 26.4K tokens), and a tag set (25 tags) that includes special tags for Twitter language specifics (username, hyperlink, email address, hashtag and emoticon). Prior to their tests, they made some adjustments to reduce misspellings and inconsistent capitalization. Their tool had better performance than the Stanford tagger for such data (90% accuracy). Huang et al. (2016) developed a tagger for locating linguistic variants by USA region, following a

similar methodology (data set of 924M tweets or 7.8B tokens). Their goal was to extract the lexical attributes of Twitter users of each region, by aggregating and smoothing them. The difficulty to address abbreviations and misspellings in such occasional and short messages was noted. Liu et al. (2012) proposed a cognitive system for normalizing non-standard tokens of social text before applying NLP techniques, which automatically converted tokens into formal English words. Their system was evaluated at both the word and message levels, by using 4 data sets of SMS and tweets. They tried to implement letter transformation patterns that humans use to decipher tokens, resulting in an approximately 90% successful conversion of words in all data sets. Popescu and Pennacchiotti (2010) used a POS tagger for tweets in English and created 3 models to identify controversial events. Their data set consisted of 73.3K tweets posted by 104.7K celebrity accounts. The models was the result of supervised machine learning algorithms, based on lexicons of ambiguous words and slang, and a lexicon of 100K annotated English words. In a follow-up survey (Popescu et al., 2011) they used these models and the EventAboutness tool to automatically track events and their connections with specific celebrities on Twitter, as well as the public reaction to them. Foster et al. (2011) evaluated and improved the statistical dependency parser, Malt. They used a data set consisting of 519 annotated sentences from Twitter, but replaced the Twitter language specifics with code words to smooth the data. In the initial trials with Malt, there was a 20% decrease in accuracy. After a reinforcement with a phrase structure parser, accuracy improved by 4%. Nand et al. (2014) developed a POS tagger for microblogging text using Hidden Markov Models. They used 3 data sets of tweets. The Stanford tagger was used as a benchmark for performance evaluation. They argued that their tool was suitable for real-time applications, despite the significantly reduced performance (less than 80% accuracy) for several tags.

One of the first POS taggers for social text in German is that of Rehbein (2013), and it used Conditional Random Fields (CRFs) and word clustering (unsupervised learning), as well as an annotated data set (1.4K tweets or 20.8K tokens). Expanding the conventional tag set for German, they added tags for the Twitter language specifics (65 tags in total), (89% accuracy). Nooralahzadeh et al. (2014) developed a POS tagging model for social text in French, following a similar methodology. They used an existing data set of 1.7K sentences or 38K words from Facebook, Twitter and medical forums, and a conventional corpus as benchmark. They used the conventional tag set for French, with the addition of tags for the social text specifics (accuracy for words from 88% to 92%, accuracy for sentences from 45% to 52%). Neunerdt et al. (2013) created a new corpus (36K tokens), and compared and evaluated 4 POS taggers, based on their performance on 4 types of social text (blog comments, chat messages, YouTube comments and comments on websites). They used the conventional tag set for German (54 tags), and tagged the Twitter language specifics by defining certain annotation rules (accuracy from 84% to 93%).

Zalmout et al. (2018) developed a neural morphological tagging and disambiguation model for Egyptian Arabic, based on previous models for Modern Standard Arabic. They used a 410M-word corpus, consisting of blog comments and social media text, and a 160K-word annotated corpus in Modern Standard Arabic. Their annotation process focused on morphological features and context, and various normalization techniques were applied (22% relative error reduction in POS tagging). They argued that morphological, syntactic and phonetic variations of each dialect or geographical region may reduce significantly the accuracy of NLP models. Bach et al. (2018) proposed a POS tagger for Vietnamese social text. They created an annotated corpus (4.1K sentences or 38.4K tokens) from Facebook. They used CRFs and compared the tool's performance with conventional Vietnamese taggers. Extending the conventional tag set, they add 5 emoticon tags, 1 tag for all punctuation marks, 1 tag for all foreign words, and 1 tag for all unknown words. Their results are 88.26% tagging accuracy in supervised learning scenarios, 88.92% tagging accuracy in semi-supervised learning scenarios, and a 12% improvement over vnTagger (the most modern and widely used POS tagger for conventional text in Vietnamese).

3. Data Set

The data which were collected for the purpose of this research consist of a set of tweets written in Greek and posted on Twitter in April, 2019. Our data was searched, collected and stored using Twitter's Standard search API (`https://bit.ly/2XrNNrl`). The following parameters and filters were set while searching the API: a. a filter to ignore retweets, to avoid duplicates, b. a parameter to specify the language of the tweets, only including the ones in Greek, and c. a parameter to specify the time frame of their posting, to result in a data set of adequate size. The collected data set consisted of 2,578 tweets in total.

3.1. Preprocessing

After filtering duplicates, 88 tweets were removed, and after filtering corrupted data, 85 tweets were removed (3.41% and 3.29% of the original data set, respectively). As a result, the final data set consists of 2,405 tweets or 31,697 tokens (single words). This data set size is common in related work for similar data (microblogging text) and similar tasks (POS tagging) (Gimpel et al., 2010; Neunerdt et al., 2013; Rehbein, 2013; Nooralahzadeh et al., 2014; Bach et al., 2018).

According to Zalmout et al. (2018), morphological, syntactic and phonetic variations of each dialect or geographical region may reduce significantly the accuracy of NLP models. In order to achieve high performance of machine learning algorithms with a significantly diverse data set (Papageorgiou et al., 2000; Gimpel et al., 2010; Liu et al., 2012), tokens of selected categories were replaced with code-words prior to the annotation process.

Hyperlinks (1,356 tokens) were replaced with the code-word "HTTP". 1,197 tweets contained hyperlinks. Emojis and emoticons (342 tokens) were replaced with the code-word "EMOT". When the same emoji or emoticon occurs more than once in a row, only one code-word was used to

replace the entire sequence. 287 tweets contained emojis or emoticons.

Punctuation marks were replaced with the code-word "ΣΗΣΤ". When the same punctuation occurs more than once in a row (e.g. ??, !!!), only one code-word was used to replace the entire sequence (Table 1). 27 dashes were removed, as they functioned as a joint between words (e.g. παιδί-θαύμα (whiz kid), ωμέγα-3 (omega-3)). Mathematical and other symbols were replaced with the code-word "SYMB" (Table 2). Symbols which were part of a word were removed (e.g. K@@A). Time, quantitative and numeric expressions were replaced with the code-word "NUMB" (Table 3). Such expressions are date, time, prices, temperatures, distances, and scores of any form (e.g. 1., 4/10, 2018-2019, 23:30, 2,5, 14.7C, 68-66, 0.0) (Anastasiadi-Symeonidi and Kyriakopoulou, 2015).

Punctuation	Frequency	Tweets
Period (.)	1,735	864
Ellipsis (...)	157	155
Comma (,)	594	453
Interpunct (·)	10	5
Colon (:)	281	263
Dash (-)	203	171
Parenthesis (( ))	191	90
Bracket ([ ])	38	22
Quote (“ ”)	125	64
Guillemet (≪ ≫)	218	102
Apostrophe (’)	58	35
Exclamation mark (!)	1,506	382
Semicolon (;)	270	195
Question mark (?)	176	97

Table 1: Total number of occurrence of each punctuation mark and number of tweets containing one or more punctuation marks.

Symbol	Frequency	Tweets
Plus (+)	5	5
Asterisk (*)	3	3
Slash (/)	8	8
Equals sign (=)	4	4
At sign (@)	11	5
Percent (%)	15	12
Euro sign	3	3

Table 2: Total number of occurrence of each mathematical or other symbol and number of tweets containing one or more symbols.

Expression	Frequency	Tweets
Date or time period	33	28
Time	10	9
Temperature	2	2
List	3	2
Decimal	11	8
Score	11	11

Table 3: Total number of occurrence of each time, quantitative and numeric expression and number of tweets containing one or more expressions.

3.2. Tag Set

Following data normalization, a tag set was created, which was used for the annotation process. It consists of 22 tags, including special tags for Twitter language specifics, and it is fully shown in Table 4. This tag set was created considering grammars for the Greek language (Triantafyllidis, 1990; Tzevelekou et al., 2007), and Wiktionary for Greek (`https://bit.ly/2ylRINe`). Tags related to specific categories of tokens that are found exclusively within the social network Twitter ("ΥΣ", "H" (Latin), "E" (Latin), "U") were created based on the tag set of Gimpel et al. (2010).

Tag	Description	Examples
ΚΝΟ (Greek)	Common noun	νερό, τιμεσ
ΚΡΟ (Greek)	Proper noun	Θήβα, Τόνια
Ρ (Greek)	Verb/ active participle	ειναι, μιλώντας
ΕΘ	Adjective/ passive participle	περιορισμένος, απλη
ΕΡ (Greek)	Adverb	κάτω, ναι
ΑΡ (Greek)	Article	η, ένας, στο
ΑΝ (Greek)	Pronoun	μου, εσύ, κάτι
ΣΥΝ	Conjunction	κ, όταν, άρα
ΠΡ	Preposition	από, αντί, για
Μ (Greek)	Particle	σαν, ως, θα
ΕΦ	Interjection	χαχα, μπράβο, α
ΣΣ	Punctuation mark	ΣΗΣΤ
ΣΥΜ	Mathematical/ other symbol	SYMB
ΑΡΘΜ	Numeral	όγδοος, δεκάδα, 2
ΕΚ (Greek)	Expression	NUMB
Ξ	Foreign word	reunion, οκ
ΣΥΝΤΜ	Abbreviation	ΕΕ, πχ, κλπ
H (Latin)	Hashtag	#πρωταπριλια, #Greece
U	@at-mention	@aegeanews
E (Latin)	Emoticon	EMOT (Latin)
ΥΣ	Hyperlink	HTTP
Δ	Miscellaneous	άστα, σκκσκς

Table 4: Tag set.

In order to avoid reduced performance phenomena during the annotation and machine learning processes, that could occur due to an overly extensive tag set, certain assump-

tions were made to limit its size (Gimpel et al., 2010). As the number of participles is proportionately small considering the volume of the data set, it was decided not to create a separate tag for the participles. Passive participles were therefore categorized as adjectives and active participles as verbs. As a consequence, the passive participles are tagged with "ΕΘ" and the active participles with "Ρ".

Since this research does not focus on the semantics of the data, the following assumptions were made: a. all punctuation marks are equivalent and are, therefore, tagged with a collective tag, "ΣΣ", b. all articles (definite, indefinite, or prepositional) are equivalent and are, therefore, tagged with a collective tag, "ΑΡ" (Greek), c. all pronouns (personal, possessive, reflexive, definite, indefinite, demonstrative, relative or interrogative) are equivalent and are, therefore, tagged with a collective tag, "ΑΝ" (Greek), and d. all adverbs (of place, time, way, quantity, certainty, hesitation, or negativity) are equivalent and are, therefore, tagged with a collective tag, "ΕΡ" (Greek).

It was also assumed that the adverbs "σαν" (like) and "ως" (as) are tagged with a collective tag, "Μ" (Greek), along with particles. Finally, as an exception to Triantafyllidis (1990), "δε(ν)" (neither, nor) and "μη(ν)" (non) are considered as particles and not as adverbs, as most of modern Greek grammars (Tzevelekou et al., 2007) consider them as particles.

3.3. Annotation

After data normalization and tag set creation, the annotation process started as follows. The data set was partitioned so that it would not consist of whole tweets, but of the 31,697 tokens that compose them.

At first, a team of 6 domain experts acted as annotators. Each annotator POS tagged the tokens assigned to them by the researchers, consulting the grammars Triantafyllidis (1990), Tzevelekou et al. (2007), and Wiktionary for Greek (`https://bit.ly/2ylRINe`). It has been observed that context assists the annotators in POS tagging of tokens, thereby making the choice of the correct tag easier and less ambiguous, and thus fewer errors to be corrected during the checking phase. So, the annotators knew the context of each token during the annotation process. Additionally, the context is used as a basis for several attributes of the training and test sets for machine learning experiments.

In order to avoid frequent errors that were observed during the annotation process, certain assumptions were made to provide specific guidelines for annotators in ambiguous and special cases. Tenses of verbs, such as present perfect, past perfect etc., are formed with an auxiliary verb, e.g. "έχω κάνει" (have done). Therefore, these two words (auxiliary and main verb) are considered and tagged as two separate verbs ("Ρ" (Greek)). Also, the conjunction "για να" (to) consists of two tokens, so they are tagged separately as preposition ("ΠΡ") and conjunction ("ΣΥΝ"), respectively. As foreign words ("Ξ") are considered words that are written in any language other than Greek, e.g. "reunion", "Trump", as well as any words written in Greek but are of foreign origin, e.g. "ακαου", "οκ", "μπαρ" (account, ok, bar). Only proper nouns are excluded, e.g. "Ντράγκι", "Λίβερπουλ", "Τομ" (Draghi, Liverpool, Tom). A first name followed by a last name is tagged with two tags as proper nouns ("ΚΡΟ" (Greek)). Only names of the form "Λ. Σισιλιάνος" are excluded, with the first name tagged as abbreviation ("ΣΥΝΤΜ") and the last name as proper noun ("ΚΡΟ" (Greek)). Numerals also include adjectives and nouns related to age, time, or quantity, e.g. "16χρονος", "εξηντάχρονος", "δεκάλεπτο", "20λεπτο", "10άδα", "ντουζινα" (16-year old, 60-year old, ten minutes, twenty minutes, ten, dozen). Therefore, they are tagged with "ΑΡΘΜ". Finally, the words "όλος, -η, -ο" (all) are usually tagged as adjectives ("ΕΘ"), since when they describe a noun, there is always a definite article before the noun, e.g. "όλος ο κόσμος" (all the world) (Tzevelekou et al., 2007). In any other case, they are tagged as pronouns ("ΑΝ"(Greek)).

During the annotation process certain observations were made about the data set, considering the nature of the language of the social web and its specifics. Tweets often contain features of oral speech, e.g. "ναιιιιιιι", "Γουστάρωωωωωω", "Χαααααχαχα" (yeaaaah, Likeee, Hahaha), especially when they are posted from users who are not public figures (news agencies, government officials, celebrities etc.). Also, it is common to glue two or more words, e.g. "Γιαυτό", "αστο" (for that, leave it), resulting in a token consisting of two or more words instead of one ("Δ" tag is assigned). On the other hand, public figures tend to use more formal and conventional language. The use of idioms and dialects, such as Cretan, Cypriot and Pontic Greek, is also common. Tokens that either come from these dialects or are idioms are tagged as foreign words ("Ξ"), to moderate the diversity of the data set (Papageorgiou et al., 2000; Gimpel et al., 2010; Liu et al., 2012; Zalmout et al., 2018), in order to achieve high performance of machine learning algorithms. Finally, regarding the parts of speech of the tokens, the following were observed: a. words such as "καληνύχτα", "καλημέρα" etc. (goodnight, good morning) are usually used as nouns ("ΚΝΟ" (Greek)) instead of interjections ("ΕΦ"), b. the most common tag is "Ρ" (Greek), (verb), which was assigned to a total of 4,149 tokens, c. the least common tag is "ΣΥΜ" (mathematical or other symbol), which was assigned to a total of 43 tokens, and d. the most frequent token, with 3,692 appearances, is the codeword "ΣΗΣΤ" (punctuation mark).

Based on the level of detail of the aforementioned guidelines, the cases of disagreement between the annotators were extremely rare. Unanimous decisions were reached in all these cases after discussion. When every partition of the data set was annotated by the annotators, the final, fully and correctly annotated, data set is created (31,697 annotated tokens).

4. Experiments

In order to perform experiments with machine learning algorithms, training and test sets need to be created from the annotated data set. The training set and test set consists of 31,697 examples (one example for each token). The following attributes are defined for each focus word (token), based on its context: a. the third preceding neighbor of the focus word and its tag, b. the second preceding neighbor of the focus word and its tag, c. the preceding neighbor of

Tag	Frequency
U	1,696
AN (Greek)	2,317
P (Greek)	4,149
ΣΣ	3,692
EP (Greek)	1,521
KNO (Greek)	3,960
ΠΡ	1,115
ΕΘ	1,461
ΣΥΝ	1,937
AP (Greek)	3,343
ΥΣ	1,345
ΑΡΘΜ	466
M (Greek)	907
KPO (Greek)	1,596
Ξ	782
Δ	123
EK (Greek)	68
ΣΥΜ	43
E (Latin)	338
ΣΥΝΤΜ	217
ΕΦ	301
H (Latin)	320

Table 5: Frequency of each tag.

the focus word and its tag, d. the next neighbor of the focus word and its tag, e. the second next neighbor of the focus word and its tag, f. the third next neighbor of the focus word and its tag, and g. the suffix of the focus word (last 3 characters). All of these attributes, except of the suffix, may has "NULL" (tags) or blank (words) values if the focus word has no context. Additionally, there is an attribute for the tag assigned to each token during the annotation process, "token's tag". This is also the label (predicted class) of each example for the machine learning algorithms. So, there are 13 attributes in total.

The tool we used to conduct the experiments with machine learning algorithms, the collection of results and the comparison of the models is the RapidMiner Studio Educational (`https://bit.ly/2OaBX1N`). For all experiments, 80% of the data set was used as training set and 20% as test set.

For the first experiments, the Naive Bayes algorithm (supervised learning) was implemented and applied. Laplace correction was used in order to smooth the conditional probabilities. The produced model has 99.87% accuracy, which is high, compared to related work (Section 2). Precision, recall and F1 score for each label are high. More specifically, the following are observed: a. precision ranges from 99% to 100% for all labels, except for "H" (Latin) which has the lowest value (96.60%), b. recall ranges from 99% to 100% for all labels, with "U" and "ΣΥΝΤΜ" having the lowest values (99.41% and 99.43%, respectively), and c. F1 score ranges from 99% to 100% for all labels except for "H" (Latin) which has the lowest value (98.27%). From the values described above, it seems that the parts of speech that are easier to identify are punctuation marks ("ΣΣ"), miscellaneous ("Δ"), expressions ("EK" (Greek)), symbols ("ΣΥΜ"), emoticons ("E" (Latin)) and interjections ("ΕΦ"), as their values of precision, recall and F1 score are up to 100%. This probably occurs because the labels "ΣΣ", "EK" (Greek), "ΣΥΜ" and "E" (Latin) are always exclusively assigned to the same tokens ("ΣΗΣΤ", "NUMB", "SYMB", "EMOT"), due to data preprocessing. For labels "Δ" and "ΕΦ", it is probably due to the fact that they are either often attributed to the same or morphologically similar tokens, or have similar parts of speech as context.

For the last experiments, the ID3 algorithm (supervised learning) was implemented and applied. The attributes were converted from text to nominal, in order to produce the model. Certain parameters needed to be defined: a. information gain was defined as the split criterion (no pruning), with the minimal gain defined to 0.01, b. the minimal number of node examples for splitting was set to 4, and c. the minimal leaf size was set to 2. The produced model has 99.44% accuracy, which is high, compared to related work (Section 2). Precision, recall and F1 score for each label are high. More specifically, the following are observed: a. precision ranges from 98% to 100% for all labels, with those of "ΕΘ" and "KNO" (Greek) having the lowest values (98.03% and 98.62%, respectively), b. recall ranges from 96% to 100% for all labels, with those of "H" (Latin) and "ΑΡΘΜ" having the lowest values (96.09% and 97.32%, respectively), and c. F1 score ranges from 97.50% to 100% for all labels, with those of "H" (Latin) and "ΕΘ" having the lowest values (97.61% and 98.03%, respectively). From the values described above, it seems that the parts of speech that are easier to identify are punctuation marks ("ΣΣ"), expressions ("EK" (Greek)), symbols ("ΣΥΜ") and emoticons ("E" (Latin)), as their values of precision, recall and F1 score are up to 100%. This probably occurs because the labels "ΣΣ", "EK" (Greek), "ΣΥΜ" and "E" (Latin) are always exclusively assigned to the same tokens ("ΣΗΣΤ", "NUMB", "SYMB", "EMOT"), due to data preprocessing. Hashtags ("H" (Latin)) is the most difficult part of speech to identify for both models, as its values of precision, recall and F1 score are slightly lower (precision lower than 96.60%, F1 score lower than 98.27%). Twitter hashtags are in the form of "#text", where text can be any token with any part of speech, and it is common to place hashtags to replace the corresponding tokens without "#", e.g. "μεγάλη #απογοήτευση" instead of "μεγάλη απογοήτευση" (great frustration). This has also been observed by other researchers with data sets from Twitter (Gimpel et al., 2010; Rehbein, 2013; Nand et al., 2014; Nooralahzadeh et al., 2014; Bach et al., 2018). Some even choose to tag the token with the label it would normally have, ignoring the presence of "#". The rest of them assign "H" to every token starting with "#", as in the present study. It is also worth noting that there is not much decline in the values of precision, recall and F1 score for most labels that are not Twitter specifics. For example, there are not frequent errors in classification when it comes to nouns ("KNO", "KPO" (Greek)) and adjectives ("ΕΘ"), while other Greek POS taggers often assign adjective tag to nouns, and vice versa ((Petasis et al., 2001)).

Wrong predictions were identified with RapidMiner Stu-

Label	Precision	Recall	F1 score
U	100.00%	99.41%	99.70%
AN (Greek)	99.73%	99.78%	99.75%
P (Greek)	99.94%	99.97%	99.95%
ΣΣ	100.00%	100.00%	100.00%
EP (Greek)	99.84%	99.67%	99.75%
KNO (Greek)	99.91%	99.94%	99.92%
ΠΡ	99.89%	99.89%	99.89%
ΕΘ	99.83%	99.91%	99.86%
ΣΥΝ	100.00%	99.81%	99.90%
AP (Greek)	99.93%	99.89%	99.90%
ΥΣ	99.91%	99.91%	99.91%
APΘM	100.00%	99.73%	99.86%
M (Greek)	99.45%	100.00%	99.72%
KPO (Greek)	99.92%	99.77%	99.84%
Ξ	99.68%	99.84%	99.75%
Δ	100.00%	100.00%	100.00%
EK (Greek)	100.00%	100.00%	100.00%
ΣΥΜ	100.00%	100.00%	100.00%
E (Latin)	100.00%	100.00%	100.00%
ΣΥΝΤΜ	100.00%	99.43%	99.71%
ΕΦ	100.00%	100.00%	100.00%
H (Latin)	96.60%	100.00%	98.27%

Table 6: Values of precision, recall and F1 score for each label (Naive Bayes).

Label	Precision	Recall	F1 score
U	99.05%	99.78%	99.41%
AN (Greek)	99.35%	99.30%	99.32%
P (Greek)	99.82%	99.85%	99.83%
ΣΣ	100.00%	100.00%	100.00%
EP (Greek)	99.18%	99.59%	99.38%
KNO (Greek)	98.62%	99.46%	99.03%
ΠΡ	99.89%	99.78%	99.83%
ΕΘ	98.03%	98.03%	98.03%
ΣΥΝ	99.81%	99.48%	99.64%
AP (Greek)	99.70%	99.78%	99.73%
ΥΣ	99.91%	100.00%	99.95%
APΘM	99.73%	97.32%	98.51%
M (Greek)	99.18%	99.86%	99.51%
KPO (Greek)	99.36%	97.89%	98.61%
Ξ	99.20%	98.88%	99.03%
Δ	100.00%	97.96%	98.96%
EK (Greek)	100.00%	100.00%	100.00%
ΣΥΜ	100.00%	100.00%	100.00%
E (Latin)	100.00%	100.00%	100.00%
ΣΥΝΤΜ	100.00%	98.85%	99.42%
ΕΦ	100.00%	99.17%	99.58%
H (Latin)	99.19%	96.09%	97.61%

Table 7: Values of precision, recall and F1 score for each label (ID3).

dio Educational. Articles ("AP" (Greek)) are incorrectly classified as pronouns ("AN" (Greek)). This seems to occur for articles that are morphologically similar (same attribute "3-char suffix") with certain types of pronouns. Additionally, at-mentions ("U") are incorrectly categorized as hashtags ("H" (Latin)). This seems to occur when the at-mention is the last word of the tweet and does not have any neighboring words after it. Conjunctions ("ΣΥΝ") are incorrectly classified either as adverbs ("EP" (Greek)), or as particles ("M" (Greek)). This seems to occur when the conjunction is the first word of the tweet and has no previous neighboring words. Adverbs ("EP" (Greek)) are incorrectly classified either as particles ("M" (Greek)), or as pronouns ("AN" (Greek)), or as prepositions ("ΠΡ"), or as verbs ("P" (Greek)). This seems to occur for adverbs which have the same suffix (same attribute "3-char suffix") with some particles, pronouns, prepositions and verbs. Pronouns ("AN" (Greek)) are incorrectly classified either as particles ("M" (Greek)), or as articles ("AP" (Greek)), or as prepositions ("ΠΡ"), or as common nouns ("KNO" (Greek)). This seems to occur for pronouns that have the same suffix (same attribute "3-char suffix") as some particles and articles, or the pronoun is the first word of the tweet and has no previous neighboring words, so it is confused with hyperlinks or common nouns.

5. Conclusion

The increasing volume of communication via microblogging messages on social networks has led to a growing demand for efficient NLP tools, especially for unstructured text processing. Extracting information from unstructured social text is one of the most demanding NLP tasks (Nand et al., 2014), (Bach et al., 2018).

The present work described the design and development of a novel data set for POS tagging microblogging data in Greek, as well as its application and evaluation on real Greek Twitter data. The contribution of our research is: a. creating the first annotated data set for Greek social text (2,405 tweets or 31,697 tokens), b. creating the first tag set (22 different tags), including special tags for Twitter language specifics, and c. developing the first supervised POS tagger for such data with significantly high performance (accuracy up to 99.87%).

An issue that needs to be addressed by future researchers who intend to use the tool developed in this work is that of overfitting. This issue arises when the result of an analysis is largely dependent on (over-adapted) to a particular set of data, resulting in the inability to adapt to more or less specific data. Therefore, it is pending examining machine learning algorithms in different data sets and training them with more data to address this issue.

The development of more POS taggers for social text in Greek is of great interest, and has great potential for future research. Additionally, such taggers could be modified to tag data sets consisting of microblogging text in Greek, derived from different social networks. This POS tagger could also be a benchmark for the development of syntactic and semantic analysis tools for social text in Greek (e.g. for sentiment analysis). Finally, it could be used for more sophisticated linguistic analysis of tweets; analysis of the linguistic variants of tweets, by region, like Huang et al. (2016), or to identify controversial events, like Popescu and Pennacchiotti (2010), and to automatically track events and their connections with specific Twitter accounts, as well as the public reaction to them, like Popescu et al. (2011).

6. Acknowledgements

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 579, Acronym: Let's Talk!).

7. Bibliographical References

Anastasiadi-Symeonidi, A. and Kyriakopoulou, P., (2015). *Automatic word processing*, chapter 2. Hellenic Academic Libraries Association, Athens. (in Greek).

Bach, N. X., Linh, N. D., and Phuong, T. M. (2018). An empirical study on pos tagging for vietnamese social media text. *Computer Speech & Language*, 50:1–15.

Foster, J., Cetinoglu, O., Wagner, J., Le Roux, J., Hogan, S., Nivre, J., Hogan, D., and Van Genabith, J. (2011). # hardtoparse: Pos tagging and parsing the twitterverse. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2010). Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding us regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Keersmaekers, A. (2019). Creating a richly annotated corpus of papyrological greek: The possibilities of natural language processing approaches to a highly inflected historical language. *Digital Scholarship in the Humanities*.

Liu, F., Weng, F., and Jiang, X. (2012). A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 1035–1044. Association for Computational Linguistics.

Nand, P., Perera, R., and Lal, R. (2014). A hmm pos tagger for micro-blogging type texts. In *Pacific Rim International Conference on Artificial Intelligence*, pages 157–169. Springer.

Neunerdt, M., Trevisan, B., Reyer, M., and Mathar, R. (2013). Part-of-speech tagging for social media texts. In *Language Processing and Knowledge in the Web*, pages 139–150. Springer.

Nooralahzadeh, F., Brun, C., and Roux, C. (2014). Part of speech tagging for french social media data. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1764–1772.

Papageorgiou, H., Prokopidis, P., Giouli, V., and Piperidis, S. (2000). A unified pos tagging architecture and its application to greek. In *LREC*.

Petasis, G., Karkaletsis, V., Farmakiotou, D., Samaritakis, G., Androutsopoulos, I., and Spyropoulos, C. (2001). A greek morphological lexicon and its exploitation by a greek controlled language checker. In *Proceedings of the 8th Panhellenic Conference on Informatics*, pages 8–10.

Popescu, A.-M. and Pennacchiotti, M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1873–1876. ACM.

Popescu, A.-M., Pennacchiotti, M., and Paranjpe, D. (2011). Extracting events and event descriptions from twitter. In *WWW (Companion Volume)*, pages 105–106.

Rehbein, I. (2013). Fine-grained pos tagging of german tweets. In *Language Processing and Knowledge in the Web*, pages 162–175. Springer.

Sgarbas, K. N., Fakotakis, N. D., and Kokkinakis, G. K. (2001). A straightforward approach to morphological analysis and synthesis. *arXiv preprint cs/0112010*.

Triantafyllidis, M. (1990). *Modern Greek Grammar*. Textbook Publishing Organization, Athens. (in Greek).

Tzevelekou, M., Kantzou, V., and Stamouli, S. (2007). *Basic Greek Grammar*. National and Kapodistrian University of Athens, Athens. (in Greek).

Zalmout, N., Erdmann, A., and Habash, N. (2018). Noise-robust morphological disambiguation for dialectal arabic. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 953–964.