Extraction and normalization of IR indexing terms and phrases in a highly inflectional language

Panagiotis Gakis | ORCID: 0000-0003-3847-9777
University of Peloponnese, Kalamata, Greece
gakis@sch.gr

Theodoros Kokkinos | ORCID: 0000-0002-2086-9673
Hellenic Open University, Patras, Greece
theokokkinos@yahoo.gr

Christos Tsalidis
Neurolingo Company, Athens, Greece
tsalidis@neurolingo.gr

Abstract

Term-based indexing of documents is conventionally implemented by stemmers or their corpus-based improvements, both of which encode implicit linguistic information. Terms are directly derived from document content such that a unique indexing approach is available at indexing run-time. For highly inflectional languages where term variation is high, such techniques are more error-prone. The main focus of the current study is the extraction and normalization of single terms and phrases and the proposal of authenticated control of indexing. The proposed approach relies on the use of explicit linguistic knowledge, appropriately encoded in large language resources. Such control guarantees the highest possible expansion factor for indexing terms as well as indexing consistency. Moreover, it offers a framework where different and eventually contradicting indexing criteria can be practiced, conventional and Natural Language Processing (NLP)-based Information Retrieval (IR) applications can be served, while adaptations can be made for tuning to a specific domain or corpus.
Keywords

natural language processing for Information Retrieval – stemming/morphological analysis – phrase detection and use

1 Introduction

The conventional approach for indexing term identification for Information Retrieval (IR) applications is based on direct extraction of words as they appear in the input text followed by reduction of variant wordforms to common roots by stemming. Stemmers are suited for morphologically poor languages like English (Tzoukermann et al. 1997) and they are widely used for such languages. However, the improvement in retrieval effectiveness is frequently reported as statistically insignificant (Adam et al. 2010; Harman 1991). Improvements to the stemming baseline are also known, for example, Singh et al. 2019 exploit the use of corpus-based wordform co-occurrence information on top of aggressive stemmers. On a similar track Jacquemin 1997 presents an algorithm for automatic acquisition of morphological links between words, also covering multi-word term conflation. Such techniques use additional but still implicit linguistic information, which is acquired off-line from corpora by statistical analyses and is encoded as add-ons to stemmers. As they derive one indexing term per wordform, a unique indexing approach is available at run-time; in other words, there are no options for alternative types of indexing. Even in advanced Natural Language Processing (NLP) systems for IR, the basics of stemming are used along with validity checks against a dictionary (Kang et al. 2010).

Regardless of the weighting and ranking approach, deficiencies in the indexing term identification process are gradually propagated, and naturally they have a negative impact on system effectiveness. For highly inflectional languages, it is impossible to obtain accurate (or at least fairly consistent) mapping of wordforms by suffix stripping without wide lexicon support. The declinable wordforms produce a huge set of morphologically inflected word forms, since Modern Greek is a highly inflectional language. For instance, from a single verb lemma more than 300 inflected word forms can be produced (including both active and passive voice word forms); from an adjective lemma about 100 word forms can be produced (if we include the comparative and superlative forms) (Gakis et al. 2012.). On the other hand, identification of wordform variants (word normalization) is essential for retrieval of text written in such languages. Otherwise, many different terms are used to represent the salient con-
cept expressed by wordform variants, resulting in disappointingly low recall (Krovetz 1993). The normalization task for words of such languages can only rely on explicit linguistic information encoded in extensive computational lexicons.

Furthermore, an old idea in the IR field is to extend the set of indexing terms with multi-word indexing terms (expressions), toward precision enhancement. For the corresponding problem of identification/selection of the important phrases, different approaches have been proposed including: simple conventions that match specific part-of-speech (POS) sequences e.g. (Zheng et al. 2009), usage of co-occurrence frequencies, coding of predefined important phrases, and exploitation of syntactic analysis of text toward syntax-directed extraction of phrases (Kang et al. 2010). To a great extent, the first approaches avoid theories of language structure, and they are thus likely to be similarly applied to other languages. The issue, however, derives from the need to match phrase variants, in order to identify their equivalents in the document space. The task of matching phrase variants acts similarly to word normalization but on phrases. It is called phrase normalization, aiming at automatically normalizing the identified phrases in a form or template, which represents a set of similar phrases. Coping with phrase normalization is a complex task, as phrase variation in natural language is high, and further high in inflectional languages. The basic advantage of computational lexicons (in contrast to printed ones) is their ability to store arbitrary amounts of information in any of their field. (Gakis et al. 2012).

This article addresses the term- and phrase-identification and normalization problem in a highly inflectional language (Greek), relying on the usage of extensive linguistic knowledge encoded in computational lexicons. Issues relating to lexicon design and coverage have been recognized as being among the most critical aspects in NLP systems, which are generally only as good as the lexical resources they employ (Boguaev & Pustejovský 1996). With the focus on Greek, the conceptual organization of language resources for term indexing and normalization raises interesting questions, which under similar developments would also appear.

To give the feeling of morphological complexity to the non-native speaker and its implications in IR, we discuss in short a few aspects of the language: The number of possible inflections is very large; however only a few directly indicate a specific tagset. This would be necessary to decide the suffix to be stripped by a hypothetical stemming algorithm. Several inflexional patterns exist for each of the parts-of-speech. The number of different stems in verbs is typically two, but in principle may be up to five. Many forms carried-over from Ancient Greek differ in morphology; these may be exact alternatives for newer
ones, or they lose their morphological and/or semantic transparency (Gakis et al. 2012). Part of the morphological system is the mobile stress. A morphological variant may differ, for instance, only in the position of the stress mark; however there exist semantically irrelevant words that differ in the same manner. The stress is not applied to uppercase wordforms; therefore, simplifications such as transformation to uppercase for reducing term variation cannot be applied without loss of precision.

2 Word normalization

Word normalization is an operation that provides a unique and identical representative for all wordforms representing the same salient concept. It is frequently regarded as ‘lemmatization’ and entirely implemented by stemmers. Sometimes it is assisted by exception lists or lemma lexicons (Kang et al. 2010), and/or improved by statistical analyses of corpora (Jaquemin 1997; Singh et al. 2019). We propose the alternative of morphological normalization, entirely based on full lexicons that appropriately organize wordforms of the same lemma. Besides the difficulties in the effort required to develop the lexicon and ensure high coverage, the main issue concerns its conceptual organization: In practice, the way of lemmatization is neither unique, nor self-evident. From a linguistic perspective, a word is associated with its lemma, but morphology is further discriminated into inflectional and derivational. A solid definition for IR purposes would be only grounded on analytical declaration of wordform variants. Dictionaries however, make inconsistent use of the theoretical discrimination between inflectional and derivational morphology; for example, they list some derivatives of common roots as separate lemmas. For example the word απαντήσεις is nominative, accusative, or vocative plural of the lemma [απάντηση: noun] ‘answer’, with inflectional morphology as well as derivational morphology from the lemma [απαντώ: verb] (Badecker & Caramazza 1998). Usually, this indicates a semantic dependence and/or a different meaning (Krovetz 1993), while for IR it may indicate a different normalization criterion. Additional criteria are derived from the semantic dependencies between words which are not morphologically related. Given the diversity of wordforms, as well as the diversity of possible normalization criteria, it seems impossible to reach a conclusion beforehand on the needed functionality of a global normalization device. In particular for highly inflectional languages, it is nearly impossible to determine and maintain word senses at the morphological level and simultaneously take into account their meaning variation in different domains.
Clearly, we should discriminate normalization criteria to obtain a consistent inclusion or exclusion of wordforms, certain classes of derivatives and/or other semantically related words. Moreover, as the effect of the indexing/normalization strategy on retrieval effectiveness is only known after the experiments, and the needs of conventional and advanced information retrieval systems are different, the ultimate goal is to offer a framework where all the above, different and maybe contradicting normalization criteria can be practiced.

In our design, the normalization functionality is provided by a full lexicon of reasonable linguistic competence, high coverage and accuracy. The basic idea behind the lexicon as an IR system component is that the normalization criteria are all ultimately anchored in linguistic knowledge. Therefore, instead of simple patterns, intuition and statistics, provision of normalization functionality depends on the on-line availability of precise, explicit and appropriately organized linguistic information. Based on the theoretical discrimination between inflectional and derivational morphology as well as on the particular needs for an IR application, we defined a first layer of lexicon organization where generic, globally true and consistent linguistic data are declared (wordforms along with their inherent properties) and their groupings in so-called clusters, which provide the typical (stemming-like) normalization. As complexity increases when relationships between wordforms in the derivational or semantical level are considered, we defined a relational lexicon layer, which is for the declaration of referential links (derivational and semantic links for IR applications). From the implementation point of view, this layer consists of implicit references to data of the first layer, thus asserting the data-link independence criterion.

2.1 **Word normalization at inflectional layer**

For mapping wordform variation at the inflectional level onto a unique indexing term, we organized the lexicon into so-called inflectional morphology clusters, or simply clusters. Each cluster comprises wordforms and their attributes (values of linguistic features and special purpose flags to indicate, for instance, that the wordform is an older form; see the examples in the caption of Table 1). In addition, we exceptionally include specific types of derivatives, since clearly their meaning is most of the time very similar to the needs of typical normalization. When a different grouping is required, this can be performed by attribute-based operations. With respect to wordforms, an inflectional cluster\(^1\) consists

\(^1\) The same definition has been also used for the conceptual organization of the lexicon as a component of a grammar checker.
TABLE 1 Inflectional cluster for a noun

<table>
<thead>
<tr>
<th>Noun</th>
<th>ημέρα <a href="day">iméra</a></th>
<th>Cluster Identifier: 21593</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wordforms</td>
<td>Attributes</td>
<td></td>
</tr>
<tr>
<td>ημέρα</td>
<td>NOUN FEM SING NOM ACC VOC</td>
<td></td>
</tr>
<tr>
<td>ημέρας</td>
<td>NOUN FEM SING GEN</td>
<td></td>
</tr>
<tr>
<td>ημέρες</td>
<td>NOUN FEM PLU NOM ACC VOC</td>
<td></td>
</tr>
<tr>
<td>ημερών</td>
<td>NOUN FEM PLU GEN</td>
<td></td>
</tr>
<tr>
<td>μέρα</td>
<td>NOUN FEM SING NOM ACC VOC</td>
<td></td>
</tr>
<tr>
<td>μέρας</td>
<td>NOUN FEM SING GEN</td>
<td></td>
</tr>
<tr>
<td>μέρες</td>
<td>NOUN FEM PLU NOM ACC VOC</td>
<td></td>
</tr>
<tr>
<td>μερών</td>
<td>NOUN FEM PLU GEN</td>
<td></td>
</tr>
</tbody>
</table>

NOUN, FEMinine, SINGular, PLUral, NOMinative, GENitive, ACCusative, VOCative

of: (i) All inflectional forms of a word. (ii) Degrees for adjectives or adverbs. (iii) Participles for verbs. (iv) Wordforms for both active and passive voice verbs. (v) Diminutives and augmentatives for nouns or adjectives. (vi) Contracted forms. (vii) Abbreviations. Nominalized adjectives and participles are neither declared in separate clusters, nor represented by redundant wordforms in one cluster. Clusters are not necessarily distinct in respect to the wordforms they contain because wordforms may be ambiguous. For all clusters of the lexicon, a distinct numeric identifier (called cluster identifier) uniquely characterizes a cluster and is used as the default indexing term for word normalization at the inflectional layer. Table 1 presents the contents\(^2\) of a noun cluster, where the identifier 21593 has been assigned automatically and represents all eight wordforms along with their attributes. A corresponding definition for, for example, a verb cluster would typically contain about two hundred wordforms.

The entire lexicon space is viewed as a set of inflectional clusters. Table 2 presents statistics of the current version of the lexicon, including the number of clusters per POS and their distribution in wordforms. Coverage of wordforms and morphosyntactic information is extremely high, and currently, it is the largest computational resource of its type available for Greek.

\(^2\) The details of the lexicon development approach e.g. morphological description, lexicon consistency, redundancy vs. ambiguity etc. are outside of the scope of this paper.
### Lexicon statistics

<table>
<thead>
<tr>
<th>Category</th>
<th>#Clusters</th>
<th>#Wordforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>33,045</td>
<td>155,054</td>
</tr>
<tr>
<td>Adjectives</td>
<td>13,921</td>
<td>196,566</td>
</tr>
<tr>
<td>Verbs</td>
<td>7,203</td>
<td>569,461</td>
</tr>
<tr>
<td>Adverbs</td>
<td>3,718</td>
<td>5,951</td>
</tr>
<tr>
<td>Closed classes</td>
<td>152</td>
<td>676</td>
</tr>
<tr>
<td>Rest POS</td>
<td>479</td>
<td>560</td>
</tr>
<tr>
<td>Total</td>
<td>58,518</td>
<td>928,268</td>
</tr>
</tbody>
</table>

#### 2.2 Word normalization at relational layer

The relational layer is for defining groups of wordforms that could all be represented by the same indexing term taking as the relevant criterion their semantic similarity. Such groups constitute one of the following: (i) they are words of common roots (derivatives) defined in different clusters but their semantic difference—as IR indexing terms—is unimportant. (ii) They are grammatically irrelevant wordforms, which are related through a general or domain-specific relationship such as synonymy. (iii) They are exceptions of inflectional clustering when e.g. some wordforms pose a different or additional meaning. The formal definition of relationships is made by implicit\(^3\) links to inflectional clusters. Each relationship has a name, a type and references to lexical entries. The type is the basis for interpretation of the referred lexical entries of clusters, in terms of either grouping or separating the referred clusters. The entries are consecutive declarations, each consisting of one or more of the following:

- An implicit link to a cluster, using a backslash and any unambiguous\(^4\) wordform of the cluster, enclosed in angle brackets.
- Positive attributes for selecting particular wordforms of the referred cluster.
- Negative attributes for filtering out particular wordforms of the referred cluster.
- Explicit\(^5\) references to one or more wordforms included in quotes.

---

\(^3\) Explicit references to cluster identifiers are not suitable because the first layer changes all the time.

\(^4\) It should always be possible to locate an unambiguous wordform; otherwise the cluster is redundant.

\(^5\) Rarely used for assigning very exceptional semantics and they always interpreted as implicit.
Table 3 presents some entries of typical relationships. From the perspective of end-users who ask for documents that contain any of the, say, 9 variants of πώληση [pólisi] (the sale), it seems reasonable that they are probably interested in documents containing the verb πουλώ [puló] (to sell) or any of its 201 variants. The organization in clusters assigns different indexing terms for each of these cases, so that there is a need for selectively defining superclusters, for instance set unions of two or more clusters. The relationship DER_ALL connects verbs with the corresponding nouns or adjectives, etc. The criterion for such connections is not widely linguistic (anchored in etymology), as it does not take into account major semantic differences.

Correspondingly, the purpose of NOM_ADJ_PCPL is to discriminate wordforms of a cluster that pose an important semantic difference, due to transcategorization as nouns of certain adjectives and participles. The corresponding entries in Table 3 indicate the meanings of the separated wordforms (original vs. most frequent meaning) and the number of wordforms discriminated by the corresponding declaration, out of the total of wordforms in the cluster. Notice that the separation is made by using properties of the wordforms instead of the wordforms themselves.
Another criterion for word normalization refers to the association of word-forms which are grammatically irrelevant but semantically synonymous. Table 3 presents excerpts from two relationships, which encode variation of this type at a specific domain (financial) and at the general domain. Additional relationships, attaching different semantics to lexicon entries, can be defined accordingly. Dedicated separation or group relationships can be defined for every case where the cluster-based normalization criterion is not optimal for information retrieval.

The lexicon relationships are analyzed in a second pass; the first is the creation of the binary file for the inflectional layer. The purpose of the second pass is to attach additional indexing terms to wordforms referred implicitly at the relational layer by a group or a separation relationship. During the second lexicon pass, dangling cluster references may appear. In this case, warnings are produced, which suggest defining the missing clusters at the inflectional layer. Similar warnings are generated when a wordform used for an implicit link to a cluster is located in more than one cluster.

Another problem is the consistency of linking data, particularly when the system is obliged to consider and interpret contradictory relationships. For example, a row in the NOM_ADJ_PCPL may indicate splitting of some wordforms of a cluster, while the entire cluster is grouped with others via an e.g. DER_ALL or SYN relation. Frequently, this may require correction of the group relationship to exclude the wordforms separated in, for instance, NOM_ADJ_PCPL, but in general, it is a great subject of experimentation as both may be valid when different domains are covered.

Ultimately, for indexing/normalization purposes, we can choose between using the default cluster identifier as an indexing term, and/or one of the additional indexing terms. The lexicon as a resource for IR consists of its binary files and a search library offering normalization capabilities for every wordform it includes, along with search capabilities such as digital searching. The organization in two separate lexicon layers offers the ability to define the semantic similarities between inflectional clusters instead of encoding detailed similarities between wordforms. The relationships declared are application-specific, as it is known that general-purpose thesauruses do not consistently improve IR effectiveness e.g. (Gurevych et al. 2012). Note that corresponding semantic differences or similarities may exist for a few wordforms of large lemmas, and in this case, we would have to declare wordforms in analytic form, in order to attach such exceptional semantics. In addition, it offers flexibility in defining contradictory relationships, to be used for instance in a similar application that covers a different domain (the frequent meaning of wordforms may change when the domain changes). Finally, the data-link independence crite-
rion, which is important for lexicon maintainability, is asserted: relational data are not general-purpose and not globally true, and they are not mixed with the globally true lexicographic data and generic links of the first layer.

2.3 Inflectional clustering experiments

Normalization experiments have been carried out on two corpora. ASE corpus is rather small (741 documents/3.7 MB) but it is a carefully compiled and manually classified corpus from the Athens Stock Exchange. The NF collection consists of 6625 documents from newspapers (13.4 MB). Usage of the lexicon offers the ability to know the ultimate expansion factor for each term, as well as to precisely measure the actual occurrences of wordform variants per indexing term. Table 4 presents such information, when the clustering criterion is used; Rows correspond to a proportion (indicated in the first column) of in-lexicon word variants that actually occur in each corpus. A general observation is that the indexing terms for each corpus can ultimately represent many more wordforms than actually occur in the corpora. It is important to note however that the actual wordforms appearing in each corpus cover a wide spectrum of tagsets of morphosyntactic properties. In other words, there are fewer universal exclusions of inflectional types in the corpora.

3 Phrase normalization

Multi-word terms (phrases) frequently express important concepts of texts, concepts which are not implied by each of the participating single terms. Among others, this is a basis for the intuition that phrases are better indexing terms compared with simple keywords (Kang et al. 2010; Kraaij & Pohlmann 1998; Zheng et al. 2009). Existing techniques for phrase identification extend from matching of tagged sequences and usage of co-occurrence information to the development of dedicated syntactic grammars (Kang et al. 2010; Kraaij & Pohlmann 1998). The first approaches seem to be easily re-applied to other languages; still, they assume availability of tagging information. In any case, the primary problem to the use of phrases is not the identification of the salient ones (those potentially relevant to our information goal), but rather their normalization, that is, the problem of matching all the variant linguistic forms of a concept expressed by a phrase.

For phrase identification, we have developed a basic Greek syntactic grammar consisting of 10 metarules, 35 terminals, and 463 production rules; these describe the frequent structures found in financial documents and cover noun, verb, adverb phrase, secondary clauses, and certain elliptical structures. The
### Table 4: Indexing terms and word variants in two corpora

<table>
<thead>
<tr>
<th>% of lexicon word variants that occur in corpora</th>
<th>#INDEXING TERMS</th>
<th>#unique word variants in lexicon</th>
<th>#unique word variants in corpus</th>
<th>Total #wordforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open class words</td>
<td>ASE  NF</td>
<td>ASE  NF</td>
<td>ASE  NF</td>
<td>ASE  NF</td>
</tr>
<tr>
<td>100%</td>
<td>410  1,868</td>
<td>807</td>
<td>4,093</td>
<td>19,501</td>
</tr>
<tr>
<td>95%–70%</td>
<td>166  951</td>
<td>867</td>
<td>5,852</td>
<td>673</td>
</tr>
<tr>
<td>69%–50%</td>
<td>542  2,226</td>
<td>2,587</td>
<td>12,793</td>
<td>1,388</td>
</tr>
<tr>
<td>49%–30%</td>
<td>408  1,571</td>
<td>3,290</td>
<td>17,658</td>
<td>1,253</td>
</tr>
<tr>
<td>29%–10%</td>
<td>1,500  4,864</td>
<td>21,004</td>
<td>99,443</td>
<td>3,500</td>
</tr>
<tr>
<td>9%–0.5%</td>
<td>1,345  2,963</td>
<td>84,331</td>
<td>178,262</td>
<td>2,898</td>
</tr>
<tr>
<td>Total</td>
<td>4,371  14,443</td>
<td>112,886</td>
<td>318,101</td>
<td>10,519</td>
</tr>
<tr>
<td>Stop words</td>
<td>127  152</td>
<td>740</td>
<td>800</td>
<td>310</td>
</tr>
<tr>
<td>Latin</td>
<td>1,160  7,665</td>
<td>2,689</td>
<td>39,411</td>
<td></td>
</tr>
<tr>
<td>Domain/unknown</td>
<td>2,384  21,835</td>
<td>8,248</td>
<td>95,122</td>
<td></td>
</tr>
<tr>
<td>Numbers</td>
<td>682  9,151</td>
<td>7,459</td>
<td>57,420</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>113,626 318,901</td>
<td>14,011</td>
<td>85,464</td>
<td>162,314</td>
</tr>
</tbody>
</table>

Analysis does not examine true underlying phenomena; it exploits the morphosyntactic features of the words involved by using the lexicon, and examines agreement criteria and their relative order. It describes the frequent forms of the freer order of the language\(^6\) and avoids generating ambiguous structures. The target of the linguistic description is to extract small phrases which are either NPs along with their modifiers, or VPs along with complements (Ntoulas et al. 2001). There are versions of the syntactic grammar that allow additional transductions; for instance, experiments have been made with all structures proposed by (Kang et al. 2010), which have been also used by others (Kraaij & Pohlmann 1997). Beyond phrase identification, the parser has been used as a phrase normalization engine itself, and to construct and structure phrase databases in order to study the problem of phrase normalization.

---

\(^6\) That is, the individual words or constituents making up a sentence or a clause can be permuted into many orders (‘freely’) without affecting the well formedness of the sentence.
3.1 Parser as phrase normalization engine

The parser is used to minimize the differences of identified phrases due to the different order of their constituents. This includes reordering for free-order structures around their semantic head, and additional changes which can be concluded from the syntactic analysis, such as the elimination of closed class words or other unimportant wordforms including numbers, stock elements and embedded phrases.

More interestingly, the parser is used for identification of phrase variation which is due to declination or conjugation. Directed acyclic word graphs (Sgarbas et al. 2000a; 2000b) and machine learning techniques (Papageorgiou et al. 2000; Petasis et al. 2001), as well as statistical methods (Tambouratzis & Carayiannis 2001), constitute alternative models that have been developed for Modern Greek. Due to morphological richness, phrase variation without structural differences and POS change of the words involved are very high in Greek. For example, the simple NP phrase: η συγχώνευση της τράπεζας ([i sinhónefsi tis trápezas] ‘the merging of the bank’) can ultimately appear in 15 similar constructions7 (8 out of these appear actually in the ASE corpus), in which the number and/or the case of the participating nouns can change. Correspondingly, the VP phrase ενέκρινε την παραίτηση ([enékrine tin parétisi] ‘he accepted the resignation’) can appear in 128 constructions, related only to verb changes that accompany time, modality, person or number. With the use of the lexicon, the parser can directly perform normalization of these phrases. For the representation of the indexing phrase, the idea is to use the cluster identifiers of the wordforms involved. The common template for the phrase consists of a specific tag for the syntactic structure, which is an abbreviation for one or more terminal symbols, along with the cluster identifiers.

3.2 Matching of phrases with structural changes

Variation may involve an adjective premodifier which is turned into a noun postmodifier or participle pre-modifier, or a noun phrase with a participle modifier which is turned to a verb phrase in the passive voice etc. Normalization of such variations is not straightforward, as it cannot be only based on structure mapping and isolated word normalization. It is possible, however to conclude on phrase equivalence of particular structures, when specific relational information of wordforms is considered. When we focus on phrases with a specific prepositional content and their analytic variation is available, it is possible to

---

7 Without counting free-order alternatives or changes of linguistic properties which do not alter the string representation of phrase.
model their similarities by exploiting parse trees and dedicated relational information. What makes the problem huge is the diversity of phrase ‘concepts’ and underlying dependencies.

In order to study this problem, we used the parser to construct a database of phrases covering frequent small syntactic structures, and many instances of them. Then we build concordances by using the clustering criterion for participating words, and their structural characteristics. Study of the sorted phrases lead to the conclusion that each particular phrase does not occur in a great variation, but each specific structure has many alternatives; the actually occurring ones depend strictly on the participating words. Specifically, the theoretical phrase variation when maintaining the same words (the particular wordforms do of course change) is extremely high. Surprisingly, in a real corpus the actual phrase variation for isolated instances is comparatively low. Then we focused on each specific syntactic structure separately, allowing replacement of a word by another one with similar linguistic properties. For all instances, we identified in the phrase database the syntactic structures of all their equivalents in meaning. The superset of all those equivalents is comparable and in some cases very close to that of the theoretical variation.

We have been based on the above observations for selecting frequent and important prepositional contexts of the particular domain, and try to model ‘difficult’ normalizations by combined criteria. For instance, matching of the phrase variants such as τα συναλλαγματικά αποθέματα (τα σιναλαγματικά αποθέματα) ‘the exchange deposits’ and τα αποθέματα συναλλάγματος (τα αποθέματα σιναλάγματος) ‘the deposits in exchange’ can be decided upon by reference to the rules NP(HDNOUN1, ADJ) ↔ NP(HDNOUN2, NOUN3) iff HDNOUN1 and HDNOUN2 have the same cluster identifier, and ADJ and NOUN3 have a common derivational indexing term. Similar normalizations have been obtained for simple phrases which differ in one or more noun, adjective, or participle elements, provided that the different elements are grouped by a DER_ALL or SYN relationship.

Some normalizations, however, were obvious for a native speaker but could not be decided by such combined criteria (syntactic structure plus relationships). Many of these were domain-specific, and critical in that, for instance, they involve alternative names for companies, place and person names, and domain terminology. The variation involves transliterations to Latin, different

---

8 Such observations where in fact very practical aids in acquiring data for the relational layer.
formatting of corresponding abbreviations, usage of a profession or a company position instead of a personal name, while many were due to misspellings. The number of unique such occurrences in the ASE corpus is 3036, which have been manually grouped into 1446 classes.

Clearly, the problem of phrase normalization is not (or cannot be) exhausted. Real texts contain numerous critical phrase categories; it is likely that less NLP and more heuristics can effectively address the mapping of their variation. Unfortunately, few heuristics apply on general unrestricted text.

4 Discussion

In modern Natural Language Processing, words are represented with vectors of numbers called word embeddings. These vectors can be considered as points in a multidimensional space and used in Machine Learning and Deep Learning classification algorithms. The vectors are constructed from large corpora using Deep Learning unsupervised training techniques (Mikolov et al. 2013a) so that they will incorporate the distributional semantics of the words (Harris 1954). This means that these vectors encode the meaning of the words and phrases so that the words and phrases that are closer in the vector space are expected to be similar in meaning (Vajjala et al. 2020). This property gives corpus-based synonyms for words and phrases that can be used in semantic indexing and searching as well as IR tasks. In this innovative undertaking for Greek research, we present examples of words and phrases that utilize these technologies and are part of Machine Learning and Statistics in Natural Language Processing. Words are represented by a number that is the position of the word unit in the index of a polymorphic lexicon. We use neighborhood tables that capture the words that have appeared in a neighborhood in texts in the collection. The Skip Grams method tries to guess the syntax from a word in the input. The CBOW (Continuous Bag of Words) method tries to predict one word from its context. The distance division follows the cosine similarity. We create a neighborhood table in which we consider a window of words with a distance of 1, for each word we take its neighborhood with one word from the left and one word from the right if there is and so on (Mikolov et al. 2013a). The neighborhood table 5 (for words) & 6 (for phrases) is as follows:
### Table 5  Neighborhood table (words)

Using word dict: "dicts/embeddings/grcorpus_cbow.dic" with size=1952966834

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Method</th>
<th>Nearest(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΑΓΟΡΑ</td>
<td>469289</td>
<td>Cosine</td>
<td>ΚΕΦΑΛΑΙΑΓΟΡΑ, ΒΙΟΜΗΧΑΝΙΑ, ΚΤΗΜΑΤΑΓΟΡΑ, ΠΑΡΑΓΩΓΗ,</td>
</tr>
<tr>
<td>ΑΓΟΡΑΣ</td>
<td>248031</td>
<td>Cosine</td>
<td>ΑΓΟΡΩΝ, ΖΗΤΗΣΗΣ, ΟΙΚΟΝΟΜΙΑΣ, ΒΙΟΜΗΧΑΝΙΑΣ, ΚΤΗΜΑΤΑΓΟΡΑΣ,</td>
</tr>
<tr>
<td>ΑΓΟΡΕΣ</td>
<td>1913</td>
<td>Cosine</td>
<td>ΑΓΟΡΩΝ, ΑΓΟΡΑ, ΧΟΝΔΡΕΜΠΟΡΙΚΗΣ, ΕΓΧΩΡΙΑΣ, ΚΤΗΜΑΤΑΓΟΡΑΣ</td>
</tr>
<tr>
<td>ΜΕΤΟΧΗ</td>
<td>36596</td>
<td>Cosine</td>
<td>ΜΕΤΟΧΗΣ, FITCO, UNIBRAIN, ΒΙΟΧΑΛΚΟ, ΜΕΤΟΧΗ, ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΗ, EUROLINE</td>
</tr>
<tr>
<td>ΜΕΤΟΧΗΣ</td>
<td>14751</td>
<td>Cosine</td>
<td>ΜΕΤΟΧΗ, ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΗ, ΜΕΤΟΧΩΝ, ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΗ, ΕΙΣΗΓΜΕΝΗΣ</td>
</tr>
</tbody>
</table>

### Table 6  Neighborhood table (phrase)

Using phrase dict: "dicts/embeddings/grphrase_cbow.dic"

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Frequency</th>
<th>Method</th>
<th>Nearest(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΜΕΡΙΚΗ_ΑΠΑΣΧΟΛΗΣΗ</td>
<td>3957</td>
<td>Cosine</td>
<td>ΜΕΡΙΚΗΣΑΠΑΣΧΟΛΗΣ, ΠΡΟΣΩΡΙΝΗ_ΑΠΑΣΧΟΛΗΣ, ΑΝΑΣΦΑΛΙΣΤΗ ΕΡΓΑΣΙΑ, ΜΑΥΡΗ_ΕΡΓΑΣΙΑ</td>
</tr>
<tr>
<td>ΠΛΗΡΗΣ_ΑΠΑΣΧΟΛΗΣΗ</td>
<td>529</td>
<td>Cosine</td>
<td>ΑΞΙΟΠΡΕΠΗΣ_ΕΡΓΑΣΙΑ, ΠΛΗΡΟΥΣ_ΑΠΑΣΧΟΛΗΣΗ, ΑΠΑΣΧΟΛΗΣΙΜΟΤΗΤΑ, ΚΟΙΝΩΝΙΚΗ_ΠΡΟΣΤΑΣΙΑ</td>
</tr>
<tr>
<td>ΙΔΙΩΤΙΚΟ_ΧΡΕΟΣ</td>
<td>418</td>
<td>Euclidean</td>
<td>ΒΡΑΧΥΠΡΟΘΕΣΜΟ ΔΑΝΕΙΣΜΟ, ΚΡΑΤΙΚΟ ΧΡΕΟΣ, ΠΟΣΟΣΤΟ ΑΙΤΟΜΑΤΙΕΥΣΗ, ΣΥΝΟΛΙΚΟ ΧΡΕΟΣ, ΦΟΡΟΛΟΓΙΚΟ ΒΑΡΟΣ, ΑΝΑΛΟΓΙΣΤΙΚΟ ΕΛΑΕΙΜΜΑ, ΥΠΕΡΟΓΚΟ ΧΡΕΟΣ</td>
</tr>
</tbody>
</table>
5 Conclusions

We examined the problem of single and multi-term variation in a “demanding” language, and proposed authenticated control of indexing, by using explicit linguistic knowledge, appropriately encoded in large language resources. Cluster-based normalization for single terms offers the typical and domain-independent functionality, which is necessary for the realization of the well-known statistical approaches. We strongly believe that this will consistently improve effectiveness in the particular language and produce consistent rankings. In addition, we considered semantic relationships for terms and formalized their definition explicitly for the needs of IR. Regarding phrases, we have been primarily concerned with the extension of indexing coverage and representation, without avoiding linguistic theories. For developing a working solution on phrases, we recognized that efficiency is an important factor too, and applied reasonable restrictions e.g. adoption of shallow parsing representation. The approach for phrase variation which is due to different word order, or due to declination or conjugation is general-purpose and not domain specific.

Still, this is just a half part of the IR process: weighting and ranking were not addressed. The proposed indexing approaches should be extensively evaluated for retrieval effectiveness. Unlike new approaches for a language like English, for example, they require a maximum of experimental work. Reliable evaluation of the typical word normalization approach is difficult, because of lack of good collections that have been previously examined by others. For advanced normalization criteria, the weighting system is an open and language-independent issue in itself, first and foremost, because it should con-
sider combined indexing/normalization criteria for words that are represented by single terms but also occur in identified/normalized phrases. Moreover, dedicated similarity measures should be devised to consider the case of partial matching of phrasal indexing terms.

References


