A Lexical Approach to Acronyms and their Definitions

Cvetana Krstev, Duško Vitas, Ranka Stanković

A Lexical Approach to Acronyms and their Definitions

Cvetana Krstev*, Duško Vitas*, Ranka Stanković†

University of Belgrade, Belgrade, Serbia,
*(cvetana|vitas)@matf.bg.ac.rs, †ranka@rgf.bg.ac.rs

Abstract

In this paper we present a comprehensive approach to acronyms for Natural-Language Processing (NLP) of Serbian texts. The proposed procedure includes extraction of acronyms and their definitions that are usual Multi-Word Units (MWUs), shallow parsing of MWUs that enables MWU lemmatization and production of entries in morphological electronic dictionaries, both for MWU and acronyms, that are provided with grammatical, syntactic, semantic and domain information. This approach enables representation that reflects complex relations between acronyms and their definitions.

1. Presentation of the problem

Acronyms in Serbian, much as in many other languages, represent abbreviations usually formed from initial letters of a multi-word name or a phrase, they are often written in upper-case letters only, without any space or other separators. As a result, MWUs are squeezed into simple words which have specific orthography. Some of these words eventually become lexicalized, e.g. SIDA (*syndrome d'immunodéficience acquise* – AIDS) and then they are written – *sida* – and inflected as any other word (in this case a feminine gender word) – *sidu*, *sidom*, etc. However, a majority of them is never lexicalized, either because of the limited scope of their use or because they cannot be pronounced or used as other words of the language. These new words can pose severe problems to many NLP applications because they represent the unknown words for them: applications based on machine learning techniques have not encountered them in training corpora, while those based on lexical resources do not have them listed in lexicons. However, their adequate treatment is crucial for many applications, e.g. text-to-speech systems (Taylor, 2009), machine translation (Wolinski et al., 1995), indexing for information retrieval and text classification.

In order to adequately treat acronyms a link between them and a name they were derived from should not be lost. For instance, if an English segment *Tesla filed a petition with NHTSA seeking approval...* is translated into French by a machine-translation system as *Tesla a déposé une requête auprès de la NHTSA d'obtenir l'approbation...* then a French-speaking user will have problems in understanding it because she/he most certainly would not know that *NHTSA* stands for *National Highway Traffic Safety Administration*. Moreover, a link existing between an acronym and a multi-word name or phrase is much more complex than the one represented at the beginning of this section and given in many Serbian orthography textbooks. Namely, the acronym in use may be derived from the name in a foreign language (the original name), while the translation of the name is in use, some functional words need not be used in acronyms (*FHP* –âĂŞ ***F**ond* ~~*za*~~ ***h**umanitarno* ***p**ravo* 'Humanitarian Law Fund'), sometimes more letters than the initial one are used (*RATEL* âĂŞ– ***R**epublička* ***a**gencija* ~~*za*~~ ***tel**ekomunikacije* 'Republic Agency for Telecommunications'), sometimes letters that are not initial are used, e.g. for derived word forms (*VMA* –âĂŞ ***V**ojno**m**edicinska* ***a**kademija* 'Military Medical Academy') etc. Finally, sometimes a name (e.g. of an organization) changes but the old acronym remains in use –âĂŞ that is the case of already mentioned *RATEL*, as the new name of this organization is *Regulatorna agencija za elektronske komunikacije i poštanske usluge* 'Regulatory Agency for Electronic Communication and Postal Services'.

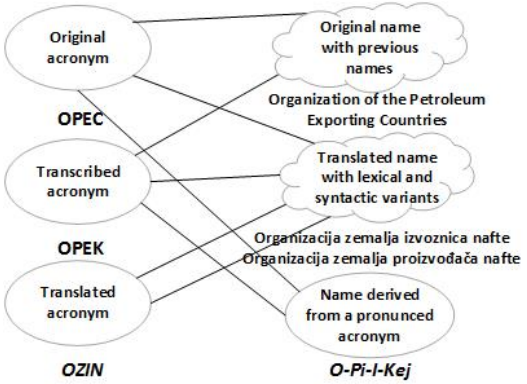

Figure 1: A Many-to-many relation between an entity names and acronyms. Names and acronyms given in italic are possibilities that are not realized for the given example.

Moreover, in many cases a relation between an entity and its name and acronym is not one-to-one. The name can change in time and some shortened variants can be in use, translated names can exhibit serious variations in used lexica and syntactic forms, original acronyms can be translated (*ILO – International Labour Organization* vs. *MOR – Medjunarodna organizacija rada*), sometimes even transcribed (*WBC – World Boxing Council* vs. *VBC – Svetski bokserski savez*), a pronounced acronym can be used together with the original and translated name (*Bi-Bi-Si* for *BBC – British Broadcasting Corporation – Britanska radiodifuzna korporacija*). Thus, this relation can rather be represented as a network of complex object (see Fig. 1).

In dealing with acronyms additional problems arise. Namely, as stated in (Spasic et al., 2005), in biomedical texts new acronyms occur in each fifth to tenth abstract, more than 80% of acronyms are ambiguous, and some of

them have as much as 15 interpretations. Although the ambiguity in general language may not be as high, many acronyms have several interpretations, for instance RAF – *Kraljevsko ratno vazduhoplovstvo* 'Royal Air Forces' and *Frakcija Crvene armije* 'Rote Armee Fraktion'.

In respect to the orthography, acronyms differ from other words in a text. Namely, they can be common names, e.g. *EKG – elektro-kardiogram* 'electrocardiogram', or proper names, e.g. *MOK – Medjunarodni olimpijski komitet* 'International Olympic Committee' but the distinction cannot be made by a simple "initial upper-case" rule.

Acronyms, much as other words in Serbian, are characterized by grammatical categories of number and gender, and they may inflect in case. The inflection is expressed by inflectional endings added after a hyphen. However, according to the Serbian orthography as well as practice the inflection is not obligatory, for instance *Kongres SAD-a* and *Kongres SAD* 'Congress of USA' are both considered correct. It should be noted that some acronyms never inflect, for instance *SFRJ – Socijalistička Federativna Republika Jugoslavija* 'Federal Socialist Republic of Yugoslavia', and it is difficult to deduce from the acronym how it behaves.[1] On the other hand, some other inflect rarely, e.g. those ending in *-a* that have feminine gender; thus, *Naučnici NASA* is more frequent than *Naučnici NASA-e* 'Scientist from NASA'. Namely, in Serbian morphology, inflectional endings are concatenated to a feminine gender lemma after deleting this final *a*. The same does not apply for acronyms, and many find it unnatural.

The grammatical gender of an acronym often stems from the acronym itself – if it ends in *-a* it has the feminine gender, in all other cases it has masculine gender, the neuter gender being extremely rare. The natural gender is inherited from a noun used in a name from which an acronym was derived, while the cases when these two genders differ and both are in use are quite often, e.g. *ISO je objavila* 'ISO announced' (feminine), *ISO je odgovoran* 'ISO is responsible' (masculine). Similar is the case with the number – the grammatical number is by a rule singular, while a natural number can be plural, e.g. *UN je organizovala* 'UN organized' (natural feminine, singular), *UN je odobrio* 'UN sanctioned' (grammatical masculine, singular), *UN su optuživale* 'UN were accusing' (natural feminine, natural plural). It should be noted that the natural number is used only with the natural gender.

2. Description of Tasks and their Goals

Work on acronyms in NLP domain focuses on some similar tasks. The first one is extraction of acronyms from corpora and detection of their variants. The second task is to detect acronym definitions and their variants in corpora and connect them with right acronyms. The third task tackles a problem of disambiguation of ambiguous acronyms in context.[2] For the first task authors use simple patterns or regular expressions (Yeates, 1999; Schwartz and Hearst, 2003; Tsimpouris et al., 2014) or shallow parsing methods (Pustejovsky et al., 2001). The second task is sometimes performed manually (Tsimpouris et al., 2014), by using some heuristics (Yeates, 1999; Schwartz and Hearst, 2003; Wren et al., 2002) or machine-learning methods (Jacobs et al., 2014). A window in which definitions of acronyms are looked for is usually narrow – definitions appear in local context – but authors in (Jacobs et al., 2014) report that they are looking for non-local expansions of acronyms (they need not appear in same documents as acronyms). The third task can be tackled by using supervised machine-learning techniques in order to assign the appropriate sense to ambiguous acronyms and abbreviations (Moon et al., 2012). Authors in (Ranchhod et al., 2004) take a different approach – they present how a list of known acronyms and their definitions is incorporated into existing electronic lexicons.

Although most of the research on acronyms focuses on general texts (e.g. newspapers), significant work was done for specific domains, especially biomedical (Schwartz and Hearst, 2003; Pustejovsky et al., 2001; Spasic et al., 2005; Moon et al., 2012), but also legislative (Tsimpouris et al., 2014). As is the case for other NLP topics, work on English prevails, but endeavors are reported for other languages as well: Greek (Tsimpouris et al., 2014), Hebrew (Jacobs et al., 2014), and Portuguese (Ranchhod et al., 2004).

In the context of the previous research briefly presented, our approach is situated as follows: we are working with general texts written in Serbian, in which we are looking locally for acronyms, their definitions and their variations, with a final goal to incorporate collected information into lexical resources for Serbian. In order to achieve these goals we have to deal with complex inflection of both Serbian MWUs and acronyms. We have followed these steps:

1. Extraction of pairs Acronym – Definition from a large corpus, where a definition represents a MWU name related to an acronym (usually, from which an acronym was derived). We presume that a MWU name is a nominal phrase having some syntactic form common for Serbian.

(1) *Medjunarodne mirovne snage na Kosovu* (KFOR)
'Kosovo Force', literally 'International peaceful forces at Kosovo'

2. Filtering of the list obtained in Step 1 in order to eliminate duplicates and false recognitions.

(2) *Zoran Djindjić* (DS) – such occurrence states that Zoran Djindjić represents Democratic Party, not that DS is an acronym for Zoran Djindjić.

3. Lemmatizing the MWU names from the list obtained in Step 2 in order to obtain names in a dictionary form, normally in the singular, nominative case, sometimes in the plural.

(3) KFOR - *Medjunarodna mirovna snaga na Kosovu* (nominative, singular)
KFOR - *Medjunarodne mirovne snage na Kosovu* (nominative, plural)

4. Checking the list obtained in Step 3 in order to correct obtained results, if necessary, choose right MWU lemma, if several were offered, and add additional in-

[1] Similarly, (Liberman and Church, 1992) state that for English it is difficult to deduce how an acronym is going to be pronounced, as a word or by spelling.

[2] In literature acronyms and abbreviations are sometimes called "short forms" and their definitions "long forms". We will not use these terms because we are dealing only with acronyms.

formation for each MWU lemma (such as is it a common or a proper name, can it be semantically described, does it belong to some specific domain, etc).

(4) KFOR – AANNxNx(plu) – *Medjunarodne mirovne snage na Kosovu* – +NProp+Org+DOM=Mil (the name is in the plural, a proper name, representing an organization from a military domain)

5. Process MWU names in order to obtain all their inflected forms associated with rich information: a lemma and associated acronym, grammatical categories, semantic and domain information, etc.

(5) *Medjunarodnim mirovnim snagama na Kosovu, Medjunarodne mirovne snage na Kosovu*.N+NProp +Org+DOM=Mil+ACR=KFOR+SIN=AANNxNx(plu) *:fp3q:fp6q:fp7q* (a form of a proper name in dative, instrumental and locative case)

6. Performing some tests on a large corpus in order to obtain information about behavior of acronyms obtained as result of Step 4: do they inflect, what is their grammatical and natural gender and number.

(6) KFOR inflects and is in the masculine gender.

7. Process acronyms obtained in Step 4 using information obtained in Step 4 in order to obtain inflected forms of acronyms, where applicable, and associate them with a MWU lemma, grammatical categories, semantic and domain information, etc.

(7) KFOR-*u, Medjunarodne mirovne snage na Kosovu*.N +NProp+Org+DOM=Mil+ACR=KFOR*:ms3:ms7* KFOR-*u*,KFOR.ABB+NProp+Org+DOM=Mil +ACR=KFOR*:ms3:ms7*

3. Used Resources and Tools

Corpus: As a corpus we have used an excerpt from the Corpus of Contemporary Serbian[3] that has more than 22 million simple word forms (more than one million sentences). This corpus contains 70% of newspaper texts (57% daily, 8% weekly and 5% monthly newspapers) and 6% of monographs and textbooks (Krstev and Vitas, 2005), which are types of texts that tend to use acronyms and provide definitions. Besides that we used two more samples of newspaper texts (having 600 thousand and 1.200 thousand simple word forms, respectively) for bootstrapping the extraction and lemmatization graphs. The final results were obtained from all texts.

Extraction graphs: They are used in Step 1. Since we were looking for acronym definitions locally, we used for the extraction two simple patterns:

beg_1 *definition between*$_1$ *acronym* end_1

beg_2 *acronym between*$_2$ *definition* end_2

where *beg*, *between*, and *end* represent strings that indicate that (probably) in a text an acronym/definition pair is detected. One set of these three strings is:

<**E**> *acronym* **(skraćeno od** *definition* **)**

('abbreviated')

For acronyms, simple patterns were used: `^[A-Z]{n}$`, where $n \in [2-6]$. The patterns for definitions were a bit more complicated and can be expressed as follows:

- A definition consists of a number of words in a sequence that corresponds to the length of the associated acronym;
- The first of these words has to be in the upper-case;
- Words can be interspersed with prepositions and/or conjunctions;
- The number of selected words can be less than a length of the associated acronym if one (or more of them) represent a compound adjective (as explained in Section 1).

It should be noted that we do not impose strict condition for a definition implying that words have to begin with the same letters used in acronyms and in the same order, because we are trying to cover as many relations between acronyms and their definitions as possible, as explained in Section 1. Also, "words" are just potential words of a language – strings of alphabetic characters – and we do not look for them in dictionaries. However, we have to look in dictionaries to confirm, for instance, the occurrence of prepositions and/or conjunctions. These patterns are implemented as Unitex[4] transducers, which produce input for the next step by recognizing modelled patterns.

Graphs for filtering and simple word lemmatization: They are used in Step 2 and the first phase of Step 3. Filtering is done by imposing certain syntactic forms. For instance, two word definitions can have three forms: AN, NprepN or NconjN (A stands for an adjective, N for a noun, prep for a preposition and conj for a conjunction; adjectives and/or nouns in a construction have to agree in gender, number, case and animatness, as appropriate for each particular form). Three word definitions have much more versatile forms, most used being: AAN, AnAN, ANN, ANNx, NprepNconjN, NprepNprepN, etc. (An stands for an adjective in the neuter singular nominative case, Nx a noun that does not agree with A/N, usually in the genitive). Equally versatile are forms of definitions that yield four to six letter acronyms. E-dictionaries are used in order to recognize certain forms and check necessary grammatical agreement. Recognition of one particular construct – AANprepNp – is represented in a graph in Fig. 2 (upper and middle part). The agreement check is performed in the upper part; namely, if a feminine gender noun is in, for instance, genitive singular than so have to be two adjectives that precede it.

Simple word lemmatization is performed by the same graph – information about recognized word form lemma is retrieved from e-dictionaries. For instance, *mirovne* from a sequence *Medjunarodne mirovne snage na Kosovu* is recognized as an adjective in the genitive feminine singular form (A:fs2) or in the nominative feminine plural form (A:fp1). This becomes a value of a variable `$a$` (upper part of the graph in Fig. 2), and its lemma (`$a.LEMMA$`) is retrieved from the following e-dictionary lines (lower part of the same graph):

(8) *mirovne*,mirovan.**A:**ae**fs2**g
mirovne,mirovan.**A:**ae**fp1**g

[3] http://www.korpus.matf.bg.ac.rs/

[4] In Unitex complex grammars can be modelled by using finite-state transducers and e-dictionaries (http://www-igm.univ-mlv.fr/ unitex/)

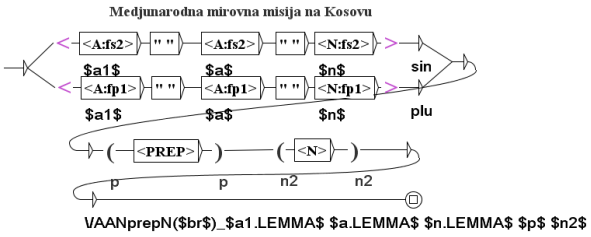

Figure 2: Two paths from a graph that filters AANprepNp constructions and performs single word lemmatization.

(in italic is the recognized form – `$a$` – in bold information used for grammatical constraints, underlined a lemma âĂŞ `$a.LEMMA$`; this is the form of dictionaries used for recognition)

In this way, the following output is performed for the given example:

(9) AANprepNp(sin) KFOR *Medjunarodan mirovan misija na Kosovu*
AANprepNp(plu) KFOR *Medjunarodan mirovan misija na Kosovu*

It should be noted that these graphs produce as output the form of a recognized definition and its (potential) number, to be used in next steps.

Graphs for MWU lemmatization. They are used in the second phase of Step 3. The results produced by simple word lemmatization are not in most of the cases correct MWU lemmas, because simple word lemmas are always in one particular form (e.g. adjectives are in the indefinite singular, masculine gender, nominative case, etc.) and agreement that exists between components of a MWU is not taken into consideration. Also, some lemmas should be in plural, adjectives are usually in a definite form, etc. Therefore, this output has to be corrected. Actually, correction graphs have basically the same form as graphs we used before, only the input is different and the used e-dictionaries as well. For the same example as before and the form (simple word lemma) *mirovan* the following e-dictionary lines are used:

(10) *mirovan*,mirovne.**A:**ae**fs2**g
mirovan,mirovne.**A:**ae**fp1**g

This form of e-dictionaries is obtained from the previous form by exchanging a form and its lemma. Now, input is a simple word lemma and the output is its desired form, as requested by agreement conditions. This form of dictionaries is used for generation. As for our example, two lines in (9) obtain the corrected forms given in Example (3).

Generation of MWU inflected forms. The list of correct MWU lemmas obtained from previous graphs has to be checked by human experts, for instance to decide whether the lemma should be in the plural or singular (it should be noted that both forms of lemmas are offered only if their initial forms in corpus were such that such possibilities exist). In the case of our example, the plural form is chosen as the correct one, and additional information in form of markers is provided (see Example (4)).

Our task in Step 5 is to generate all inflected forms of MWU lemmas produced in previous steps. For this we are using inflectional transducers for MWU lemmas (described in (Savary, 2009)) and a tool for production of proper MWU e-dictionary lemmas that provide all necessary information for inflection (Krstev et al., 2010). This tool for items produced in the previous step offers one or more MWU lemmas for inflection. These lemmas provide information needed for simple word inflection for each of their components that inflect (see Example 11). Normally, a user would have to check them and choose a correct one. For the present application, since structures of lemmas are known in advance (detected by filtering graphs), a correct e-dictionary lemma is selected automatically. For our example, such lemma is:

(11) Medjunarodne(medjunarodan.A7:aefp1g) *comps*
mirovne(mirovan.A18:aefp1g) *that*
snage(snaga.N610:fp1q) *inflect*
na Kosovu, *do not inflect*
NC_AXAXN4X1 *inflectional transducer*
+NProp+Org+DOM=Mil+ACR=KFOR
+SIN=AANNxNx *sin/sem markers*

Application of inflectional transducer NC_AXAXN4X1 to a lemma presented in Example (11) produces 7 inflected forms, one of which is presented in Example (5).

Inflection, gender and number of acronyms. For the list of acronyms produced in previous steps we checked whether they are used with inflectional endings or not. For that we used simple regular expressions:

(12) <ACRONYM> "-" [a|u|om|em] – masculine
<ACRONYM> "-" [e|i|u|om] – feminine

It can be noted that some endings are used with the masculine gender only (*-a*, *-em*), while others are used for the feminine gender only (*-e*, *-i*). For acronyms that occurred with these discriminative endings the gender was recorded.

For other acronyms additional tests were performed. For instance, the ending *-u* is used for dative and locative masculine forms, while the same ending is used for accusative feminine forms. Therefore, in order to confirm the feminine gender we used two tests (13). The similar tests were performed for the masculine gender.

(13) <PREP+P4><ACRONYM> "-" u
(<A:4sf>|<PRO:4sf>) <ACRONYM> "-" u

(<PREP+P4> is a preposition that agrees with the accusative, <A:4sf> and <PRO:4sf> are an adjective and a pronoun, respectively, in the feminine accusative singular.)

Finally, we performed test to establish the gender and the number of acronyms on the basis of their agreement with verbs. A test for feminine gender acronyms in singular is given in example (14).

(14) <ACRONYM> je <ADV>? <V:Gfs>
<V:Gfs> je <ACRONYM>

(<ADV> is an adverb, <V:Gfs> is a present participle in the feminine gender singular, *je* is 'is'.)

At the end, we collected all information to produce dictionary entries for acronyms (as in Example (7)). Appropriate forms were generated for all acronyms that inflect. Entries with information about the gender and the number were generated for those that do not inflect and this was established by test like (13) and (14). For all remaining the simple rule of thumb was used – all are singular, and only those that end in *-a* are in the feminine gender.

4. Results

In Table 4. we present the results obtained by our procedure, step by step. In Table (acr., infl. def.) stands for acronym-definition pairs where acronym name is in some inflected form (*all* denotes all retrieved pairs, while *diff* denotes all different pairs), (acr., SWLs) stands for acronym-definition pairs where acronym name is a string of simple word lemmas, and (acr., MWL) stands for acronym-definition pairs where acronym name is a multi-word lemma (*prop* denotes proposed pairs, while *corr* denotes all chosen correct pairs). By e-dict_MWL we denote correct MWLs in the form required by morphological e-dictionaries, while e-dict_(MW|acr)_forms represent inflected forms of MW names and acronyms, respectively.

	Input	Size	Output	Size
1	corpus	23MW	(acr., infl. def.)$_{all}$	3,942
2	(acr., infl. def.)$_{diff}$	2,163	foreign	96
			manually	57
			rejected	233
			(acr., SWLs)$_{diff}$	2,812
3	(acr., SWL)$_{diff}$	2,812	(acr., MWL)$_{prop}$	2,836
4	(acr., MLW)$_{prop}$	2,836	(acr., MWL)$_{corr}$	1,190
			corrected	54
			diff. acronyms	987
			diff. MWLs	1,160
5	MWLs	1,038	e-dict_MWLs	1,017
			manually	21
			multiple entries	19
			e-dict_MW_forms	9,956
6	corpus	23MW	acr. inflects	333
	acronyms	987	acr. in masc.	323
			acr. in fem.	64
			acr. both m&f	35
			acr. in plural	2
7	acronyms	987	e-dict_acr_forms	6,946

Table 1: Results of the procedure on our corpus

The majority of (acronym, definition) pairs represent proper names – 996, and the majority of these proper names are organizations 912, followed by toponyms 35. The majority of organizations are political (213), business and financial (113), government (94) and sport (85).

5. Future Work

In the future, we plan to apply our procedure to different corpora in order to obtain more acronym-definitions pairs and to supply missing information concerning grammatical and inflectional properties of acronyms. Moreover, we intend to improve our extraction procedure in order to improve the recall and lemmatization procedure as to lemmatize successfully less frequent MWU structures. We plan to develop procedures for the detection of semantic properties of acronym-definitions by looking into results obtained by this research.

On the basis of the presented approach we plan to enhance our approach to domain specific acronyms, where an "upper-case" principle may not be applicable.

Acknowledgments

This research was supported by the Serbian Ministry of Education and Science under grants 47003 and 178003.

6. References

Jacobs, K., A. Itai, and S. Wintner, 2014. Acronym Dictionary Construction and Disambiguation (abstract). 3rd Parseme General Meeting – Poster Session.

Krstev, C., R. Stanković, I. Obradović, D. Vitas, and M. Utvić, 2010. Automatic construction of a morphological dictionary of multi-word units. In *IceTAL*, volume 6233 of *LNCS*. Springer.

Krstev, C. and D. Vitas, 2005. Corpus and Lexicon – Mutual Incompletness. In *Proc. of the Corpus Linguistics Conference, Birmingham*.

Liberman, Mark Y and Kenneth W Church, 1992. Text analysis and word pronunciation in text-to-speech synthesis. *Advances in speech signal processing*:791–831.

Moon, S., S. Pakhomov, and G. B. Melton, 2012. Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. In *AMIA Annual Symposium Proceedings*, volume 2012. American Medical Informatics Association.

Pustejovsky, J., J. Castano, B. Cochran, M. Kotecki, and M. Morrell, 2001. Automatic extraction of acronym-meaning pairs from MEDLINE databases. In *Medinfo 2001: Proc. of the 10th World Congress on Medical Informatics*, volume 84.

Ranchhod, E., P. Carvalho, C. Mota, and A. Barreiro, 2004. Portuguese Large-scale Language Resources for NLP Applications. In *4st LREC*.

Savary, A., 2009. Multiflex: a multilingual finite-state tool for multi-word units. In *Implementation and Application of Automata*. Springer, pages 237–240.

Schwartz, A. S. and M. A. Hearst, 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, volume 8. World Scientific.

Spasic, I., S. Ananiadou, J. McNaught, and A. Kumar, 2005. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in bioinformatics*, 6(3):239–251.

Taylor, Paul, 2009. *Text-to-speech synthesis*. Cambridge University Press.

Tsimpouris, C., K. Sgarbas, and S. Panagiotopoulou, 2014. Acronym identification in Greek legal texts. *Literary and Linguistic Computing*, 30(3):440–541.

Wolinski, F., F. Vichot, and B. Dillet, 1995. Automatic processing of proper names in texts. In *Proceedings of the 7th conference on European chapter of the ACL*. Morgan Kaufmann Publishers Inc.

Wren, J. D., H. R. Garner, et al., 2002. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of information in medicine*, 41(5):426–434.

Yeates, S., 1999. Automatic extraction of acronyms from text. In *New Zealand Computer Science Research Students' Conference*.