LEXICOGRAPHICA

International Annual for Lexicography
Revue Internationale de Lexicographie
Internationales Jahrbuch für Lexikographie

Edited by
Rufus H. Gouws, Ulrich Heid, Thomas Herbst,
Anja Lobenstein-Reichmann, Stefan J. Schierholz
and Wolfgang Schweickard

2018 · Volume 34

DE GRUYTER

LEXICOGRAPHICA was founded in 1985 by Antonín Kučera, Alain Rey, Herbert Ernst Wiegand, and Ladislav Zgusta.

Notes for contributors to LEXICOGRAPHICA
Information on the submission process and the yearbook's style sheet can be found on the following website: https://www.degruyter.com/view/j/lexi

ISSN 0175-6206 e-ISSN 1865-9403

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: Meta Systems Publishing & Printservices GmbH, Wustermark
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Contents

Non-thematic Part

Reviews

Lexicography in Higher Education

Besim Kabashi

A lexicon of Albanian for natural language processing

Abstract: For many applications in the field of natural language processing, a lexicon is needed. For the Albanian language a lexicon that can be used for these purposes is presented below. The lexicon contains around 75,000 entries, including proper names such as personal, geographical and other names. Each entry includes grammatical information such as parts of speech and other specific information, e.g. inflection classes for nouns, adjectives and verbs. The lexicon is part of a morphological tool, but can also be used as an independent resource for other tasks and applications or can be adapted for them. Sources for the creation and the extension of the presented lexicon include both information from traditional dictionaries, e.g. spelling dictionaries, and a balanced linguistic corpus using corpus-driven methods and tools. The lexicon is still work in progress, but aims to cover basic information for most frequent tasks of natural language processing.

1 Introduction

Lexicons are very important for many tasks in the field of natural language processing/human language technology, where either only part of the information is extracted or the unabridged dictionary is used. For the Albanian language there are

Besim Kabashi, Friedrich-Alexander Universität Erlangen-Nürnberg, Korpus- und Computerlinguistik. Bismarckstr. 6, 91054 Erlangen; Ludwig-Maximilians-Universität München, Vergleichende und Indogermanische Sprachwissenschaft sowie Albanologie, Geschwister-Scholl-Platz 1, 80539 München, e-Mail: besim.kabashi@{fau,lmu}.de

https://doi.org/10.1515/lexi–2018-0012

many types of dictionaries nowadays, cf. Lloshi (1988), for an overview of the time before 1988. In the three decades since Lloshi's report, new dictionaries or new types of dictionaries for Albanian have been compiled, e.g. synonym dictionaries, cf. Thomai et al. (2004), and Dhrimo et al. (2002), antonym dictionaries, cf. Samara (1998), bilingual dictionaries e.g. Newmark (1994), and many specialized dictionaries in the fields of social, natural, technical, and computer sciences.

With the beginning of the digital age and the intensification of natural language processing, there has been an increasing need for more lexical data. They can be used in many areas, either as a final product, or to support the creation of other resources and tools/applications in the field of natural language processing, e.g. spell checkers, morphological analyzers and generators, or part-of-speech taggers.

For Albanian, Murzaku (1994), a kind of orthographical/spelling dictionary, is available (in electronic form), which is a lexicon comprising ca. 32,000 entries, supplied with information about parts of speech and linguistic gender, which can be adapted for natural language processing. However, recent vocabulary from the last two decades, after social and political changes in 1990–1991, is not covered. For a lot of tasks more information is needed. Another dictionary, Snoj (1994), a reverse dictionary of the Albanian language, lists more detailed information than Murzaku (1994), i.e. four forms for nouns (Sg. Indef. Nom., Sg. Def. Nom., Pl. Indef. Nom., and Pl. Def. Nom.), and three forms for verbs (1P. Sg. Ind. Pres. Act. N.Adm., 1P. Sg. Ind. Aor. Act. N.Adm., and Participle). It corresponds with the information given in the traditional dictionaries of the Albanian language like Kostallari et al. (1980) and Kostallari et al. (1984).

Until the year 2010 the maximum number of lexical entries in a dictionary of the Albanian language was 48,000, cf. Thomai et al. (2006). The spelling dictionary by Dhrimo/Memushaj (2010) raises this number to around 75,000 lexical entries, which is more than the double of the spelling dictionary by Kostallari et al. (1976). Dhrimo/Memushaj (first edition, 2010 with around 75,000 lexical entries, second edition 2015 with around 81,000 lexical entries), also has more information, e.g. about syllabification (hyphenation, word division), for the first time for Albanian, and about rarely used word forms, which are given in addition to standard forms. Other dictionaries, e.g. Samara (1998), Dhrimo et al. (2002), and Thomai et al. (2004) extend the lexical information about the Albanian language. Both properties, the higher number of lexical entries as well as the new type of information, offer the possibility to use, combine and organize this information in different forms and ways for the tasks of natural language processing.

In addition to the traditional creation of dictionaries, the enrichment of lexical data and types of data is very important to cover as much lexis and as many language properties as possible. For this purpose, we have started using a 100 million word corpus called AlCo (Albanian Corpus), which is compiled from a variety of sources, cf. Kabashi (2017). This corpus is used to update and revise the lexical data based on linguistic features/attributes, and on data like frequencies, collocations, or n-grams, extracted from the corpus. It is annotated with a fine-grained

tagset designed by Kabashi/Proisl (2016, and 2018). Together with morphological tools based on Kabashi (2015), a full form lexicon can be generated or word-forms can be lemmatized.

2 Some notes on the Albanian language

The Albanian language is used by ca. 5.5 million people in South-Eastern Europe and ca. 1.5 million people in other parts of the world. Albanian is an Indo-European language that constitutes a subgroup of its own. It is on the same level as the Hellenic, Romance, Slavic or Germanic subgroups. The language is characterized by a diverse vocabulary with many loan words due to language contact with Greek, Latin/Italian, Slavic languages and Turkish, and due to the influence of French and especially English, as world languages.

Albanian as a writing system is based on Latin. The Albanian alphabet is an extended one with combinations of basic letters of the Latin alphabet, i.e. digraphs (*dh*, *gj*, *ll*, *nj*, *rr*, *sh*, *th*, *xh*, and *zh*) and two letters with diacritic signs (*ë*, and *ç*). Seven of the thirty six letters of the Albanian alphabet are vowels (*a*, *e*, *ë*, *i*, *o*, *u*, and *y*).

Albanian has a rich morphological system. Nouns, adjectives and numerals have 20 forms each, combined from five cases (Nominative, Genitive, Accusative, Dative and Ablative), two numbers (singular and plural), as well as definiteness (indefinite and definite). Proper names are also declinable.

The use of multi-word units is typical of the Albanian nominal system, i.e. some words have articles or particles as their first part, written as two separate graphical tokens e.g. *mirë adv.*, engl. good, vs. *i mirë*, *masc.* / *e mirë*, *fem. adj.*, engl. good. According to Newmark et al. (1982) the categories of verbs are as follows: person (1st, 2nd, 3rd), number (singular and plural), voice (active and non-active, i.e. passive, middle, reflexive or reciprocal), mood (indicative, subjunctive, optative, admirative, and imperative), tense (present, past and future), aspect (common, perfect, progressive, inchoative, definite, and imperfect), finiteness (finite and non-finite, i.e. infinitive, participle, gerundive, and absolutive). Verbs (counted with infixed pronominal clitics) have up to 90 forms.

3 A standard lexicon

A dictionary, e.g. a spelling dictionary, as one type with minimal information, lists the lexical entries, separated in hyphenation places, and gives additional notes in relevant cases, e.g. a spelling variant of the entry. The lexical entries are ordered alphabetically. Each lexical entry contains at the very least information about spelling, grammatical category (part-of-speech), and other properties like grammatical

gender, or valency (in-/transitivity) of the verb. The lexical entries of verbs and nouns in the spelling dictionary of the Albanian language (Kostallari et al., 1976), which have been used also in the later dictionaries until now e.g. Dhrimo/Memushaj (2010), as standard, resemble the following examples 1 and 2:

(1) *bím/ë, ~a f., sh. ~ë, ~ët* (engl. plant)

(2) *s/jéll fol. kal. ~ólla ~jéllë* (engl. to bring)

The lexical entry (1) has the lemma (*bímë*), an alternation of the definite form in singular (*~a*, i.e. *bíma*), the part-of-speech information (*f.* i.e. feminine and means the gender and so finally noun). Next, the alternations of plural forms are given (i.e. *sh.*), in the indefinite (*~ë*, i.e. *bímë*) and definite (*~ët*, i.e. *bímët*). The lexical entry (2) has the lemma (*sjéll*), the part-of-speech information (*fol.* i.e. verb, *kal.* i.e. transitive), followed by the form alternation of the verb in the aorist (*~ólla*, i.e. *sólla*), and finally the participle of the verb (*~jéllë*, i.e. *sjéllë*).

The information in dictionaries described above can be adapted into a lexicon for natural language processing purposes. Also the information can be combined in order to compile a new type of lexical data. For more details about the different types of lexical entries in the dictionaries of the Albanian language, see Kabashi (2015: 99–123).

4 Compiling an Albanian lexicon for the purposes of natural language processing

We first give some notes on the work and improvements to compile lexicons for the purposes of natural language processing of the Albanian language.

4.1 Improvements and work in the past

Kabashi (2003) compiled an electronic lexicon based on word lists extracted from different texts. The lexicon benefits from Kostallari et al. (1976) as well as from a wordlist by M. Snoj (Ljubljana), dated 1993, with grammatical information like in the spelling dictionary of the Albanian language by Kostallari et al. (1976). The lexicon was primarily designed as component of a morphological tool (Kabashi 2003/2004). The information in the lexicon was similar to a spelling dictionary with additional data about inflection of each lexical entry of nouns, adjectives, and verbs. The number of lexical entries amounted to around 55,000.

Trommer/Kallulli (2004) have presented a morphosyntactic tagger for the Albanian language. Their tagger uses “three source lexica for the operative lexicon: 1) the full-form lexicon 2) the stem lexicon and 3) the regular lexicon” (2004: 1237). The operative lexicon has around 53,000 lexical entries.

Piton et al. (2007) created an electronic dictionary and finite state automata/transducers for automatic processing of the Albanian language in the framework of the NooJ platform. It is not clear whether the lexicon can be used separately from the NooJ platform, or whether there are two parallel lexicons which correspond to each other.

Kadriu (2013) uses a lexicon with around 32,000 entries, together with their corresponding part-of-speech information. She uses the lexicon within the NLTK framework, i.e. a natural language toolkit written in the Python programming language, together with a set of regular expressions rules that correspond to them.

Kabashi (2015), based on the previous work (2003 / 2004), created a lexicon, which is used as a base lexicon for a morphological analyzer and generator for word forms of Albanian. On the one hand, it is integrated in the morphological tool, and, on the other hand, it can be used as an independent resource. For more details about the lexicon, see Kabashi (2015: 99–123).

4.2 The idea

As in all of the aforementioned lexicon projects (in electronic form) the lexicon was either integrated in to a framework or directly in to the program code of the tool. The idea in Kabashi (2003) and Kabashi (2015) was to develop/compile a lexicon also as a parallel and independent resource that can be used for other tools and applications. This means the data are machine-readable and can be used for different tasks in natural language processing. The idea and the work presented here is to extend the information of lexical entries in the lexicon presented in Kabashi (2015), beginning with orthographic/spelling information of difficult forms, syllabification information, updating of the morphological information (classification of words into part-of-speech inflection subclasses that make the application of exact rules possible). A completely new kind of data is the phonetic information on the lexical entries. Those data have already been created and are currently in the process of proofreading. The goal is to also convert the data in to the Sampa[1] format.

In general, the new lexicon presented here aims to follow the CELEX Lexical Database, cf. Baayen et al. (1995), but with state-of-the-art methods and goals, as linked data, as well as data supplied with up-to-date information on statistics and other data derived from corpora. As an independent resource, the lexical data can be revised, extended and updated more easily. Also, eventually more authors will be able to collaborate on the resource.

In the following we present the compilation process of the lexicon.

1 SAMPA, the Speech Assessment Methods Phonetic Alphabet, is based on 7-bit ASCII characters. Like IPA, the International Phonetic Alphabet, SAMPA can be used to represent sounds as graphical characters.

4.3 Parts-of-Speech and their subclassification

As a first step we gave every noun and adjective, including numerals, a numerical declension class, as well as every verb their conjugation class. Thereafter, the stored data are tested and can serve as reliable information. Eventually, new additional lexical entries can be recognized, lemmatized and collected preliminarily using regular expressions, extraction rules and other methods. At this stage, lexical entries, e.g. verbs of the conjugation class 7, appear as shown in example 3.

(3) *... adhuroj 7, afroj 7, aftësoj 7, agjëroj 7, ajkoj 7, ajoj 7, ajroj 7, ...*

This information is needed for the modeling of morphological tools and grammars. An important part of the lexical entries are nouns, which are declinable in Albanian, e.g. the name *Tirana* can occur in the forms *Tiranë*, *Tirana*, *Tiranës*, *Tiranën*, *Tirane*. Most other names also have definite and indefinite plural forms, e.g. family names. They all need to be classified and supplied with these numbers.

4.4 Morphological information as a full-form lexicon

As a next step, we generated a full-form lexicon with the corresponding morphological information for each word-form. This data can be used for lemmatization of word-forms, generation of a word-form using lemma and the morphological information or for tagging any word-form with the morphologic information. Examples 4 and 5 show this data for a noun and a verb respectively.

(4) Sample of the full-forms of nouns:

...
bimë/bimë/S-020_NS-;S-020_AcS-;S-020_NP-;S-020_AcP-
bima/bimë/S-020_NS+
bimën/bimë/S-020_AcS+
bimës/bimë/S-020_GS+;S-020_DS+
bimët/bimë/S-020_NP+;S-020_AcP+
bimëve/bimë/S-020_GP-;S-020_DP-;S-020_AbP-;S-020_GP+;S-020_DP+;S-020_AbP+
...

(5) Sample of the full-forms of verbs:

...
sjellim/sjell/V-036_1P.Pl.Ind.Prs.Act.Adm-;V-036_1P.Pl.Sbj.Prs.Act.Adm-
sjellin/sjell/V-036_3P.Pl.Ind.Prs.Act.Adm-
sjellka/sjell/V-036_3P.Sg.Ind.Prs.Act.Adm+
sjellkam/sjell/V-036_1P.Sg.Ind.Prs.Act.Adm+

sjellkan/sjell/V-036_3P.Pl.Ind.Prs.Act.Adm+
sjellke/sjell/V-036_2P.Sg.Ind.Prs.Act.Adm+
sjellkemi/sjell/V-036_1P.Pl.Ind.Prs.Act.Adm+
sjellkeni/sjell/V-036_2P.Pl.Ind.Prs.Act.Adm+
sjellkësh/sjell/V-036_3P.Sg.Ind.Ipf.Act.Adm+
sjellkësha/sjell/V-036_1P.Sg.Ind.Ipf.Act.Adm+
...

These data can be generated based on the inflection classes, i.e. conjugation and declension classes, and their corresponding paradigms. Also new lexical entries can be easily integrated if they have been classified before.

4.5 Lexicon size, structure, and technical aspects

The presented lexicon covers, on the one hand, the vocabulary which is given by traditional dictionaries and, on the other hand, it has additional lexical entries which are not covered by them. The lexicon has around 75,000 lexical entries, including 45,500 nouns, 18,500 adjectives, 5,800 verbs, 3,200 adverbs and other parts-of-speech and abbreviations.

The lexicon is organized in alphabetical order as one file, which has a clear and strict data structure (as tables) and as such they can be exported, converted and transformed in other structures or in any database. Each lexical entry, firstly organized as a line separated in to fields, has information on the part-of-speech which it belongs to, i.e. the structure of a noun is different to the one of adjectives, to the one of verbs, to the one of adverbs and to other parts-of-speech, cf. the examples given below.

(6) Sample lexical entry of one noun and verb entry:
...
06241\bimë\bi·m/ë\bIm/ë\bimə\[cv][cv]\cvcv\4\2\3\4\bím~ë\~a\~ë\~ët\ʃ\S\020
...
57195\sjell\sjell\sjell\sjɛ.ł.\[ccv.cc.]\ccvcc\5\2\1\4\s~jèll\s~ó·lla\s~jé·llë\t\V\036
...

The data in example 6 are as follows: The first field is the ID of the lemma, followed by the lemma itself, the syllabification of the lemma with the marking of the alternation segment. Next, the information from the third field is converted into the SAMPA-format in the fourth field. Then the IPA representation of the lemma follows. The syllabification segments are shown in the next field. Next is the queue of the consonants and vowels, followed by the number of the letters of the lemma, the position of the accent, position of the alternation of the possible word-form(s), and the number of alphabet letters, where the digraphs count as one. The next

four fields contain the word-forms Sg. Indef. Nom., Sg. Def. Nom., Pl. Indef. Nom., and Pl. Def. Nom. The last three fields show the gender, part-of-speech and the declension class of the noun. In a similar way the data for a verb lexical entry given in example 6 can be interpreted. The *.ł.* is an IPA representation of the digraph "*ll*", in the following field marked with *.cc.* because the two letters belong together. The number 4 means that "*sjell*" consists of 4 letters of the Albanian alphabet.

The data are encoded in ISO/IEC-8859–1 (latin-1), ISO/IEC-8859–16 (latin-16) and Universal Coded Character Set (UCS) respectively UNICODE and saved in different formats, also as UTF-8 parallel. For more detailed information on coding of the Albanian alphabet see Kabashi (2009).

4.6 Relation to other Albanian resources and lexicons

The main part of the data is taken from the lexicon compiled by Kabashi (2015). Other data are taken from the AlCo-Corpus, cf. Kabashi (2017). Some data, e.g. about syllabification are compared with the corresponding data in Dhrimo/Memushaj (2015). Some data about syllabification and about word-forms that are not used so often, classified as difficult, as well the information about accent/stress in some compound words have been discussed with R. Memushaj (Tirana). The lexicon also benefits from some other data obtained in electronic form directly from him from time to time. New word-forms found respectively extracted from the AlCo-Corpus can be lemmatized and from the lemmas the full form paradigms can be generated, i.e. the new full-form lexicon with neologisms.

As mentioned and shortly introduced in section 4.1, there are only a few resources for Albanian language that were created and compiled for natural language processing purposes. The availability of the lexicon offered online by Murzaku (2003) is the first step to start with a lexicon with more than the basic vocabulary. Other resources and tools are not freely available.

4.7 Status of the project

The current state of the project is work in progress. New entries are added from time to time. In this context linking of the data still creates some difficulties and needs to be revised. Linking data in the lexicon is currently being defined and can be changed as the project progresses.

The phonetic data for the word-forms are currently in the compiling process. Some remaining issues are, on the one hand, the definition and marking of the syllabification and the accent and, on other hand, the IPA-transcription of some of the lexical entries. At the moment, this issue takes up the most time when working on the lexicon. Morphological data only needs to be changed in rare cases, when errors are detected.

As usual during the electronic lexicographic work some corrections are possible at any time. The work shown in detail in example 6 has already been completed.

5 Conclusion

The Albanian lexicon for the purposes of natural language processing presented here is work in progress. The aim is to have an up-to-date, state-of-the-art, and contemporary lexicon, that can be used directly or with small adaptions, or can be easily converted into other formats or structures. As the project is a one-man project, the work is proceeding slowly, as the acquisition and preparation of new data is resource intensive.

6 References

Baayen, Harald R./Piepenbrock, Richard/Gulikers, Leon (1995). *The CELEX Lexical Database*. Linguistic Data Consortium. University of Pennsylvania, Philadelphia, USA. http://celex.mpi.nl (last accessed 28 July 2014).

Dhrimo, Ali/Tupja, Edmond/Ymeri, Eshref (2002). *Fjalor sinonimik i gjuhës shqipe* (= *Dictionary of Synonyms of the Albanian Language*). Tiranë: Toena.

Dhrimo, Ali/Memushaj, Rami (2010). *Fjalor drejtshkrimor i gjuhës shqipe* (= *Spelling Dictionary of the Albanian Language*). Tiranë: Infbotues.

Dhrimo, Ali/Memushaj, Rami (2015). *Fjalor drejtshkrimor i gjuhës shqipe* (= *Spelling Dictionary of the Albanian Language*). Botimi i dytë (= Second Edition). Tiranë: Infbotues.

Kabashi 2003 = Kabashi, Besim: *Automatische Wortformerkennung für das Albanische*. Master's thesis. Universität Erlangen-Nürnberg, Germany, 2003.

Kabashi, Besim (2004). Analiza automatike e fjalëformave të gjuhës shqipe (= Automatic word form analysis for Albanian). In: *Seminari XXIII Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare* (= *The XXIII International Seminar for Albanian Language, Literature and Culture*). Universiteti i Prishtinës, Kosovo/Universiteti i Tiranës, Albania. Libri 23/1. 129–135.

Kabashi, Besim (2005). Disa propozime për modelimin e informacionit në leksikografinë kompjuterike (= Some proposals for modeling information in computer lexicography). In: *Seminari XXIV Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare* (= *The XXIV International Seminar for Albanian Language, Literature and Culture*). Universiteti i Prishtinës, Kosovo/Universiteti i Tiranës, Albania. Libri 24/1. 179–184.

Kabashi, Besim (2009). Das albanische Alphabet aus sprachtechnologischer Sicht. In: Demiraj, B. (Hrsg.): Der Kongress von Manastir. Herausforderung zwischen Tradition und Neuerung in der albanischen Schriftkultur. Hamburg: Verlag Dr. Kovač. 175–208.

Kabashi, Besim (2015). *Automatische Verarbeitung der Morphologie des Albanischen*. Erlangen: FAU University Press.

Kabashi, Besim (2017). AlCo – një korpus tekstesh i gjuhës shqipe me njëqind milionë fjalë (= AlCo – a hundred million word corpus of the Albanian language). In: *Seminari XXXVI Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare* (= *The XXIV International Seminar for Albanian Language, Literature and Culture*). Universiteti i Prishtinës, Kosovo/Universiteti i Tiranës, Albania. Nr. 36/2017. 123–132.

Kabashi, Besim/Proisl, Thomas (2016). A Proposal for a Part-of-Speech Tagset for the Albanian Language. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, 2016. Ed. by Nicoletta Calzolari etc. European Language Resources Association (ELRA) Paris. 4305–4310.

Kabashi, Besim/Proisl, Thomas (2018). Albanian Part-of-Speech Tagging: Gold Standard and Evaluation. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan. Ed. by Nicoletta Calzolari etc. European Language Resources Association (ELRA) Paris. 2593–2599.

Kadriu, Arbana (2013). NLTK Tagger for Albanian using Iterative Approach. In: *Proceedings of the 35th International Conference on Information Technology Interfaces (ITI 2013)*. Cavtat, Croatia.

Kostallari, Androkli/Domi, Mahir/Lafe, Emil/Cikuli, Nikoleta (1976). *Fjalori drejtshkrimor i gjuhës shqipe* (= *The Spelling Dictionary of the Albanian Language*). Tiranë: Akademia e Shkencave e RPS të Shqiperisë.

Kostallari, Androkli (Kryeredaktor)/Thomaj, Jani/Lloshi, Xhevat/Samara, Miço (1980). *Fjalor i gjuhës së sotme shqipe* (= *Dictionary of Contemporary Albanian Language*). Tiranë: Akademia e Shkencave e RPS të Shqiperisë.

Kostallari, Androkli (Kryeredaktor)/Thomaj, Jani/Samara, Miço/Kole, Josif/Daka, Palok/Haxhillazi, Pavli/Shehu, Hajri/Sima, Kornelja/Feka, Thanas/Keta, Beatriçe/Hidi, Agim (1984). *Fjalor i gjuhës së sotme shqipe* (= *Dictionary of Contemporary Albanian Language*). Tiranë: Akademia e Shkencave e RPS të Shqiperisë.

Lloshi, Xhevat (1988). Compiling and Editing Bilingual Dictionaries in Albania. In: *The Third EURALEX International Congress on Lexicography (EURALEX 1988)*. Budapest, Hungary.

Murzaku, Aleksandër (1994). Albanian. In: *European Corpus Initiative Multilingual Corpus I (ECI/MCI)* CD-ROM. Utrecht: ELSNET.

Murzaku, Aleksandër (2003). *Inverse Dictionary of Albanian*. Lissus Language, Literature, Computing. Albanian Linguistics, 2003. Accessed at: http://www.lissus.com/albanian (last accessed 18 February 2018).

Newmark, Leonard (1994). *Albanian–English Dictionary*. London etc.: Oxford University Press.

Newmark, Leonard/Hubbard, Philipp/Prifti, Peter (1982). *Standard Albanian – A Reference Grammar for Students*. Stanford: Stanford University Press.

Piton, Odile/Lagji, Klara/Përnaska, Remzi (2007). Electronic dictionaries and transducers for automatic processing of the Albanian language. In: *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007)*. Paris, France. 407–413.

Samara, Miço (1998). *Fjalor i antonimeve në gjuhën shqipe* (= *Dictionary of Antonyms in the Albanian Language*). Shkup: Shkupi.

Snoj, Marko (1994). *Rückläufiges Wörterbuch der albanischen Sprache*. Hamburg: Buske.

Thomai, Jani/Samara, Miço/Shehu, Hajri/Feka, Thanas (2004). *Fjalori sinonimik i gjuhës shqipe* (= *The Dictionary of Synonyms of the Albanian Language*). Tiranë: Akademia e Shkencave e Republikës së Shqipërisë.

Thomai, Jani/Samara, Miço/Haxhillazi, Pavli/Shehu, Hajri/Feka, Thanas/Memisha, Valter/Goga Artan (2006). *Fjalor i gjuhës shqipe* (= *Dictionary of the Albanian Language*). Tiranë: Akademia e Shkencave e Republikës së Shqipërisë.

Trommer, Jochen/Kallulli, Dalina (2004). A Morphological Analyzer for Standard Albanian. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal. 1271–1274. Ed. by Maria Teresa Lino etc. European Language Resources Association (ELRA) Paris, 2004.