A frequency e-dictionary of Albanian language

Dr. Manjola Lumani Zaçellari

Aleksander Moisiu University of Durres, Albania

Relevant UniDive Working groups: WG1, WG2

INTRODUCTION

Albanian lexicography is the oldest and richest field in Albanian linguistics and is enriched with new dictionaries more rapidly than monographic studies in this field. The demand for dictionaries of different types is an ongoing process, due to the economic, social and cultural developments of each nation. Several types of dictionaries have been compiled in the Albanian language, which can be grouped into encyclopaedic dictionaries and philological dictionaries. Nevertheless, lexicography deals mainly with philological dictionaries. In Albania philological dictionaries are diverse. According to the volume, we have small, medium and large dictionaries (such as The Great Dictionary of the Albanian Language- Alb. Fjalori i madh i gjuhës shqipe (FMGJSH)., which is an ongoing project funded by the Albanian government with The Albanian Academy of Sciences of Albania as project leader (the corpus will also be used for the creation of this e-dictionary). we have a wide range of dictionaries in Albania. However, we do not have a frequency dictionary in the Albanian language. Therefore, the drafting of such a dictionary and then its digitization will fill a gap in Albanian lexicography and pave the way for electronic drafting of dictionaries, given that in Albania we do not yet have a good tradition of electronic dictionaries. For its design, the following were taken into account:

Corpus size

For the creation of a corpus, the idea of "the more the merrier" is what should be kept in mind. If a list of words with high density of use, reflect accurately the frequency of use of these words in the language, then a corpus should contain enough text to approximate the overall use of discourse. For the corpus it will be used: http://albanian.web-corpora.net/; wordsmith 5.0 and other extracted corpus to be elaborated.)

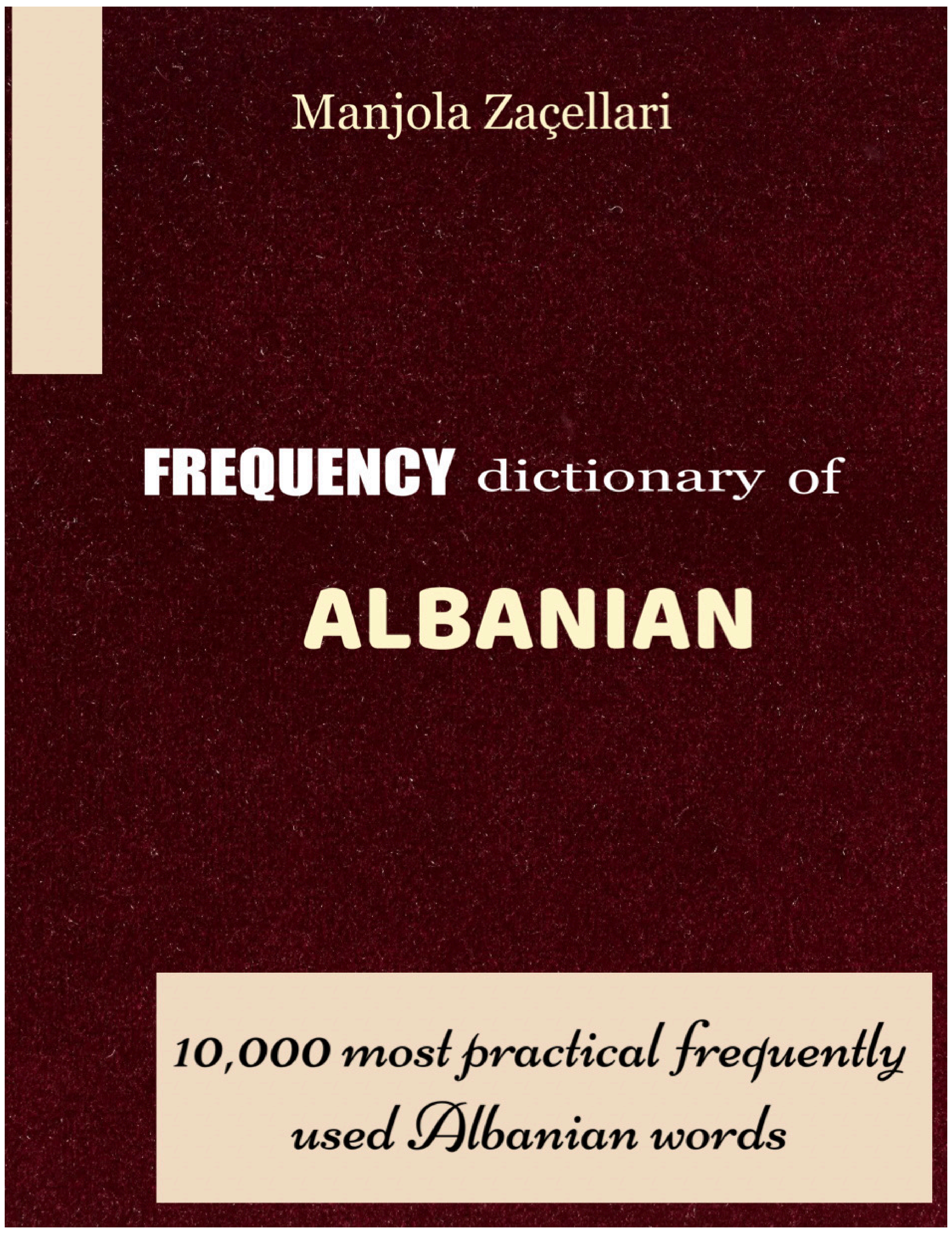
Text types

For the selection of texts, we will be based on the method of Sorell (2013), who has actually reworked the method of Biber (1988) by simplifying some text types into categories suitable for corpus design. Text types suitable for corpus according to Sorell are:

1. Interactive (conversation); 2. General reported exposition (general writing); 3. Imaginative narrative (narrative writing) 4. Academic (Sorrell, 2013, pp. 153–154). To build a frequency dictionary for a language such as Albanian is not that easy. Because corpus of spoken Albanian language is particularly difficult to obtain.

List design

It can be stated that the drafting of a frequency dictionary or the list (vocabulary) that the dictionary will contain is more difficult than determining the dictionary corpus from which the words will then be extracted. This is due to the fact that the dictionary compiler will have to work on drafting a general service list and a specialized list. The majority of frequency dictionaries aim to describe the vocabulary of the language as a whole (is often termed the core vocabulary.) Fewer researchers have created frequency dictionaries for a more specific purpose or target audience. Specialized-use lists can be designed to only include words that belong to a specific domain, such as a discipline or trade. Frequency e-dictionary of Albanian language will aim at describing the language as a whole.



Manjola Zaçellari

FREQUENCY dictionary of

ALBANIAN

10,000 most practical frequently used Albanian words

Identifying words

When designing a frequency dictionary, it must also be taken into account how to measure a word. For example, such words like: Alb: mësoj (learn), Alb: mësova (learnt) should be taken as two different words, or as one? Or the irregular verbs jap (give), dhashë (gave), dhënë (given)?

Word family levels

The word family level that suits the FEFGJSH is the lemma, consisting of a word and all of its inflected forms, but counting derived forms as separate words.

The measurement will begin from basic and will progress to the most complex: we will count *tokens, types, lemmas*, *or word families*. Example*:* Alb*:Nuk dua ta mërzit me sjelljen time, por ndonjëherë edhe sjellja e saj është e mërzitshme. I don't want to annoy her with my behaviour, but sometimes her behaviour is annoying too.*

Tokens are the total number of words. Measuring them means measuring the total number of words. The example sentence contains fourteen tokens, which means fourteen words in total.

Counting ***types*** refers to the number of separate and distinct words. That is, *sjelljen (behaviour)* and *sjellja (behaviour) are* the same type, but *sjelljen* is a different type—even a single difference makes them different types. The sentence in the example is composed of thirteen types.

The ***lemma*** consists of the stem of the word and its inflected forms, but not any derived forms of the word. So the words *mërzit, mërzitem, mërzitesh, mërzitet* (annoy) and *e mërzitshme (annoying)* are all the same lemma, but *e mërzitshme (anooying)* is not. This is because *e mërzitshme (annoying)* has the derivational affix *-shme*, which turns it into an adjective.

METHODS

This frequency dictionary uses exclusively objective criteria, such as: frequency, range, and dispersion. Frequency can refer to either raw frequency or normalized frequency. Raw frequency is simply the total number of times that a specific word is attested in the corpus. Normalized frequency is a measure of how many times the item appears for every x tokens in the corpus. Using normalized frequency is more meaningful since it is easier to compare with frequencies found in other corpora. Range is a measure of the number of sub-corpora—or sections of a corpus—in which the word can be found. Range is also sometimes referred to as contextual diversity. To measure this, a corpus must first be divided into a series of sub-corpora. Dispersion is a combination of both frequency and range. It serves as a single number—a distributional statistic—that incorporates the benefits of both of these measures, while also allowing a list to be ranked in a methodical, objective manner. Whereas frequency and range are found simply by counting, dispersion requires a calculation that incorporates multiple variables.

The Dictionary will have 10000-12000 words. The purpose of the Frequency Dictionary of Albanian Language is to provide a list of the most commonly-used lemmas in conversational Albanian.

REFERENCES

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. Behavior Research Methods, 39(3), 445–459.

Bauer, L., & Nation, P. (1993). Word families. International Journal of Lexicography, 6(4), 253–279. Biber, D. (1995). Dimensions of register variation: A cross-linguistic comparison. Cambridge University Press.

Collins Cobuild English grammar. (2005). Glasgow: HarperCollins.

Cowie, A. P. (2009). The Oxford history of English lexicography. Oxford University

Francis, W. N., Kučera, H., & Mackie, A. W. (1982). Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. Corpus Linguistics and Linguistic Theory, 5(1), 1–26. https://doi.org/10.1515/CLLT.2009.001

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 13(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: Further explorations. In S. T. Gries, S. Wulff, & M. Davies (Eds.), Corpus linguistic applications: Current studies, new directions (pp. 197–212). Amsterdam: Rodopi.

Gries, S. T. (2017). Quantitative corpus linguistics with R (2nd ed.). New York: Routledge.

Nation, I. S. P. (2013). Learning vocabulary in another language (2nd ed.). Cambridge: Cambridge University Press.

Nation, P. (2000). Learning vocabulary in lexical sets: Dangers and guidelines. TESOL Journal, 9(2), 6–10. https://doi.org/10.1002/j.1949-3533.2000.tb00239.x

Nation, P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), Language Learning & Language Teaching (Vol. 10, pp. 3–13). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.10.03nat

Schmitt, N. (2010). Researching vocabulary: A vocabulary research manual. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.

Akademia e Shkencave e Shqipërisë: Instituti i Gjuhësisë dhe i Letërsisë, Fjalor i gjuhës shqipe , Tiranë, 2006

Atlas themelor i politikës (Atlas Basico de Politica), Albas 2009

Dashnor Kokonozi: Fjalor Enciklopedik i Politikës, Tiranë 2005

http://albanian.web-corpora.net/

https://gjuhashqipe.com/apps/fjalori-i-madh-i-shqipes

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com