Collecting Collocations for the Albanian Language

Besim Kabashi

Corpus und Computational Linguistics,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Albanology, Ludwig-Maximilians-Universität München, Germany
E-mail: besim.kabashi@{fau,lmu}.de

Abstract

The presented paper describes the collecting of data from different sources to build a collocation data set with the aim of compiling the first contemporary collocation dictionary for the Albanian language. The work is based (1) on the analysis of empirical data, i. e. linguistic corpora, using the computational methods and tools, as well as (2) on traditional dictionaries. As empirical data we use the AlCo (Albanian Text Corpus), the AlCoPress 2017-2019, N-Grams extracted from both, methods like Log-likelihood and Dice coefficient using the IMS Open Corpus Workbench (CWB) and the Corpus Query Processor, Web version (CQPweb). Despite the enormous support, an unsupervised automated compilation of a collocation dictionary of high quality, like those created by lexicographers, seems to be impossible without intervention. In order to complete the collection of the data we additionally use lexical information extracted from traditional dictionaries. The primary goal is to create a language resource that can be used among others also for Natural Language Processing purposes. The presented work is still in progress and, of course, will change until its final version.

Keywords: Albanian, collocations; NLP lexicography; corpus linguistics; language resources

1. Collocations as lexical data

Linguistic data are important or even necessary for numerous applications to make the communication easier, or at least as support data for building further linguistic datasets as resources for natural language processing.

Collocations may serve one of two purposes in dictionaries. On the one hand, they are used as standalone data in collocation dictionaries. On the other hand, they serve as “additional” information in other types of dictionaries, e.g. definition dictionaries. They are not only important for non-native language learners, but also for native speakers, who sometimes need to find established ways of combining lexical items. Typical examples of collocations that can cause problems for foreign language learners are combinations such as *strong tea* (e.g. instead of *powerful tea*) in English. Many researchers distinguish between various related concepts of collocation, sometimes labelled “significance-oriented” and “statistically-oriented”, e.g. Herbst (1996). The former are often semantically restricted and are thus particularly difficult to learn. But even items that just frequently co-occur without being conventionalized may be relevant for dictionary users, since such combinations are often differentiated in usage style or are only common in specific domains. Those data can be used not only while translating from one language into another, but also for writing or speaking in a specific field, working on a desktop computer, or simply searching on a smartphone for a specific word usage.

2. Collecting collocations for a dictionary of Albanian

Currently no collocation dictionary exists for Albanian. The aim of this project is to fill this gap in Albanian lexicography by collecting collocation data for such a dictionary. For the speakers of a language, a collocation dictionary, e.g. Benson et al. (2010), Quasthoff (2010), or Häcki Buhofer et al. (2014), offers the possibility to select fine-grained collocations, to express oneself idiomatically in a conversation or text.

As this work is very data intensive, and a basic data set, e.g. for extending a given resource using NLP tools, is not available, there is a need of elementary work. For this reason we decided to take an approach consisting of three steps.

In order to collect the lexical data, we use a lexicon for NLP, presented in Kabashi (2019), and an automatic morphology for the Albanian language, presented in Kabashi (2015), to lemmatize the word forms, because Albanian is an inflected language and has a rich morphology. We also use tagged texts and the *AlCo* presented in Kabashi (2017), using the tagset presented in Kabashi & Proisl (2018). Since there is no syntactic parser available for Albanian, the extraction process is based on surface-oriented methods, e.g. n-grams and distance-based cooccurrences, see for example Evert (2013) and Proisl (2019).

3. Selecting the data sources

For an empirical data driven approach, selecting data sources also determines the quality of the data. One of the data sources is the *AlCo* (An Albanian Text Corpus), cf. Kabashi (2017). The corpus contains 100 million words and covers different domains of language and contains different text types. Additionally, another recently compiled corpus of press texts serves as a further data source – it is a reference corpus named *AlCoPress* (2017–2019) that contains approximately 32 million text words, taken from seven newspapers and a news agency. Around 70 million words are currently raw data. All in all, the data sources are around 200 million words. The amount of data, from an empirical point of view, compared to similar corpora of other languages, is still too small, but it allows extracting valuable information in most search cases and also profiling the knowledge derived from the data.

4. Methods and tools for exploring the data sources

To explore the linguistic data we use n-grams (2- to 10-), IMS Open Corpus Workbench (CWB)[1], and the Corpus Query Processor, Web version (CQPweb). This information is then complemented by the traditional selection of lexical entries from different dictionaries and lexicons, e.g. Kostallari et al. (1980), Samara (1998), Thomai et al. (2004), Dhrimo et al. (2007), and Thomai et al. (2006).

[1] Cf. http://cwb.sourceforge.net/ .

4.1. N-Grams

With the n-grams technique it is possible to extract the data from raw text without[2] any preparation, e.g. tagging or formatting. Frequency lists of n-grams allow the researcher to find words that often occur together by aggregating common combinations, cf. the 4-grams listed below. Example *10044 për herë të parë*, eng. *for the first time*, shows this effect – the accumulation of frequent word sequences. This very simple method is very useful, but a lot of cases remain, i.e. entries with low frequency, which may still have valuable collocation information, and are not listed on the top of the frequency list, but towards the end. See for example the second part of the list, after the frequency 62, where the frequencies of the word *çaj*, eng. *tea*, are listed. In this case the word-forms (of *çaj*) are not lemmatized, so the word-forms *çaj*, *çaji*, *çajin*, *çajit*, ... (with the properties case, number, gender, and definiteness for nouns), are listed separately as they originally occur in texts.

10044	për herë të parë
6599	. Nga ana tjetër
2999	do të thotë që
2659	një kohë të gjatë
2598	. Për këtë arsye
2304	*pjesën më të madhe*
2241	gjithnjë e më shumë
2083	*pjesa më e madhe*
437	një kohë të shkurtër
433	të nivelit të lartë

62	Si përgatitet çaji:
23	të çajit të gjelbër
22	e çajit të gjelbër
15	i çajit të gjelbër
16	një lugë çaji me
11	Nga çaji i përgatitur
10	që nuk pinë çaj
10	për të bërë çajra
10	një filxhan me çaj
9	filxhan çaj jeshil .
9	bërë çajra kundër sëmundjeve
9	bërë çajra kundër sëmundjeve
8	se çaji i zi
8	një çaj të ngrohtë
8	nga një gotë çaji
8	çaji i malit dhe
7	të çajit të zi
7	një çaj bimor
7	. Çaji i gjetheve
6	rastet çaji zihet derisa
6	ta përzieni me çajin
6	përbërjen e çajrave .
6	me çajin nga kamomili
6	lugë çaji të kanellës
6	lugë çaji me piper
6	Ky çaj përdoret për
5	shumë çaj të ftohtë
5	i pemës së çajit
5	e çajit të koprës
5	1 lugë çaji pluhur
4	vinte çaji i darkës
4	tufë çaji në tregun
4	shumë çaje qetësuese ,
4	të prodhimit të çajit
4	rigoni e çaji i
4	qese të çajit të
4	që çaji i kajsisë
4	përgatisni çajrat me fruta
4	përdorni çaj të tharë
4	ose çaj me sheqer
4	monopolin e çajit kinez
4	pini çaj pa sheqer
[...]	
1	përgatisni një çaj frutash
1	një çaj para gjumit
1	Një çaj pa avull
1	me çaj para buke

List 1: The list of some of the most frequent 4-grams and the 4-grams of çaj.

Not all occurrences of the words are collocations of the word çaj, e.g. *një lugë çaji me*, eng. *teaspoon*, where the word *tea* is a collocation of the word *spoon.* At the same time,

[2] In this case, driven by a script running on a Linux operating system.

not all collocations of the word *çaj* can be found in the n-gram lists. In this case, the types and/or the amount of text do not cover all collocations of the word. Additional texts from certain domains and increasing the overall amount of text would increase the probability of covering them.

4.2. CWB & CQPweb

The IMS Open Corpus Workbench (CWB) is "a collection of open-source tools for managing and querying large text corpora [...] with linguistic annotations".[3] CQPweb (Corpus Query Processor) is a software package, a web-based corpus analysis system, to explore corpus data, cf. Hardie (2012).

In contrast to the n-gram method, CWB and CQPweb, and particularly CQPweb, offer a lot of functions for calculating collocations. As the tools support the processing of linguistic data, also based on linguistic annotations, the data can be explored on more dimensions, e.g. by searching based on POS-tags or within certain domains.

To find the collaboration candidates, CQPweb can use the *Conservative LR*, *Dice coefficient*, *Log-likelihood*, *Log Ratio* (filtered), *MI2*, *Mutual Information*, *T-score*, *Z-score*, and as well as the simple *rank by frequency.* For each lexical entry the different measurement results help to find words which can be added to the respective lexical entry.

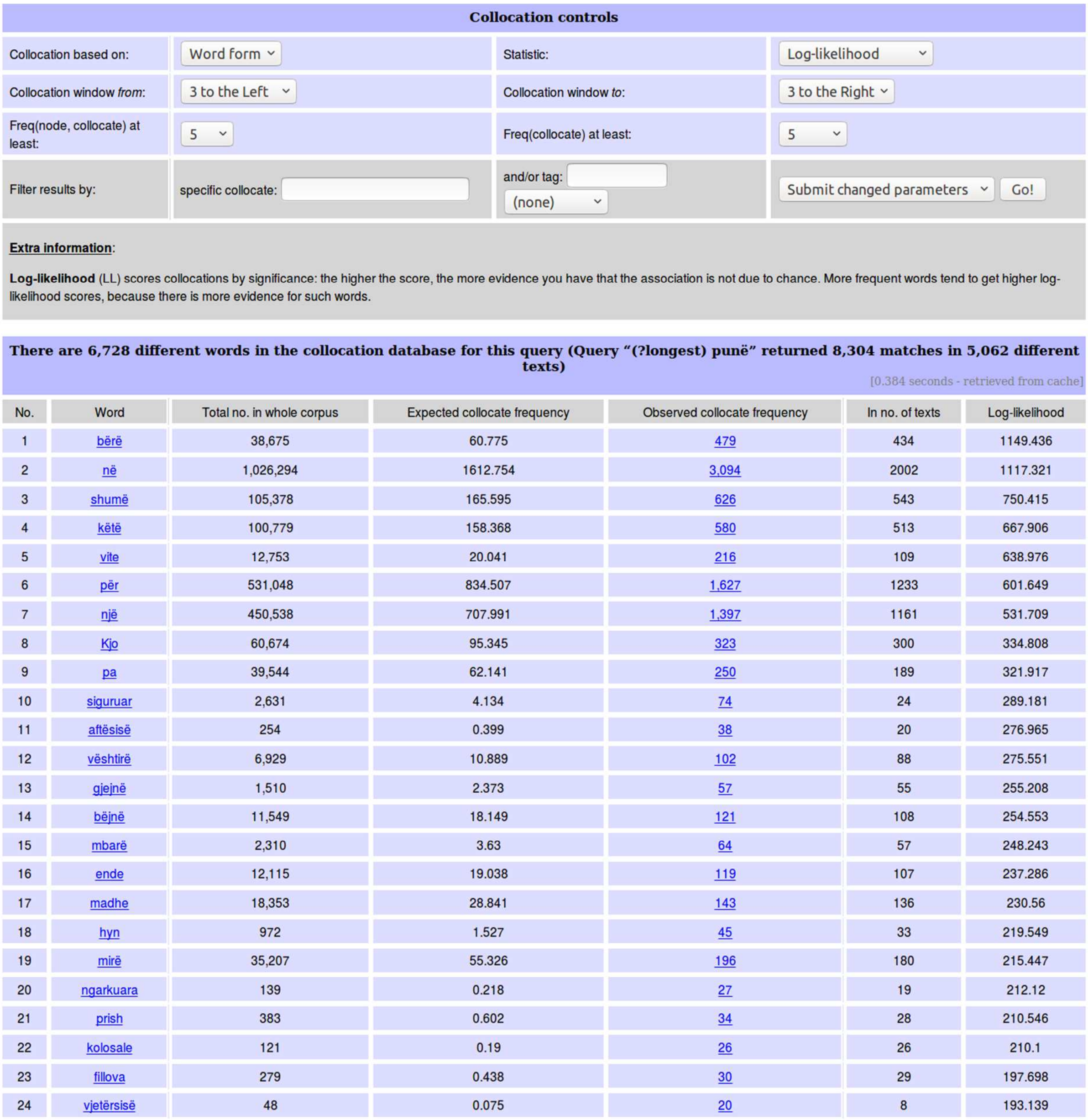
In the example above (cf. Figure 1) a list of collocations of the word *punë*, eng. *work*, is shown, calculated based on Log-likelihood by CQPweb. Below in the section *Example Entries* we present a detailed entry (as a working version) for this word. Each method, depending on different criteria, can offer different collocation candidates, e.g. using the *Log-likelihood* results is different than the *Dice coefficient* results. The statistics offered by CQPweb e.g. observed collocate frequency, the number of texts, and Log-likelihood (in figure 1) make selecting collocate candidates easier. Through using all of them, it is possible to gather more collocation candidates. Some of those candidates cannot be used, e.g. because they are function words in a sentence and not, for example, prepositions that are associated with the collocation candidate. An example is the word *në*, eng. *in*, listed in Figure 1, which depending on the concrete context can be a collocation, or not. In positions 7 and 8 (in Figure 1), the words *një*, eng. *one*, and *kjo*, eng. *this*, are listed very high, but in the sense of collocation they are both irrelevant. The list in this abbreviated version does not show the "typical" collocations, as may be expected from a native speaker. The fact that the data sources are only written texts might explain this.

Due to those problems we found it necessary to complement the results from CQPweb with the information contained in traditional dictionaries.

[3] Cf. http://cwb.sourceforge.net/ .

4.3. Dictionaries

Traditional dictionaries like definition dictionaries, i.e. Kostallari (1980) and its newer editions, collect information based on long-term observation of language. They do not offer intentionally typical examples of collocations within their usage examples, but as the goal is different to the collocation dictionaries, they are not listed separately either. As a result, only a small number of collocations can be found within the entries, mostly within the examples. The number of collocations in those cases is higher if the lexical entry has more lexical meanings.



Collocation controls

Collocation based on:	Word form	Statistic:	Log-likelihood
Collocation window *from*:	3 to the Left	Collocation window *to*:	3 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters Go!

Extra information:

Log-likelihood (LL) scores collocations by significance: the higher the score, the more evidence you have that the association is not due to chance. More frequent words tend to get higher log-likelihood scores, because there is more evidence for such words.

There are 6,728 different words in the collocation database for this query (Query “(?longest) punë” returned 8,304 matches in 5,062 different texts)

[0.384 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	bërë	38,675	60.775	479	434	1149.436
2	në	1,026,294	1612.754	3,094	2002	1117.321
3	shumë	105,378	165.595	626	543	750.415
4	këtë	100,779	158.368	580	513	667.906
5	vite	12,753	20.041	216	109	638.976
6	për	531,048	834.507	1,627	1233	601.649
7	një	450,538	707.991	1,397	1161	531.709
8	Kjo	60,674	95.345	323	300	334.808
9	pa	39,544	62.141	250	189	321.917
10	siguruar	2,631	4.134	74	24	289.181
11	aftësisë	254	0.399	38	20	276.965
12	vështirë	6,929	10.889	102	88	275.551
13	gjejnë	1,510	2.373	57	55	255.208
14	bëjnë	11,549	18.149	121	108	254.553
15	mbarë	2,310	3.63	64	57	248.243
16	ende	12,115	19.038	119	107	237.286
17	madhe	18,353	28.841	143	136	230.56
18	hyn	972	1.527	45	33	219.549
19	mirë	35,207	55.326	196	180	215.447
20	ngarkuara	139	0.218	27	19	212.12
21	prish	383	0.602	34	28	210.546
22	kolosale	121	0.19	26	26	210.1
23	fillova	279	0.438	30	29	197.698
24	vjetërsisë	48	0.075	20	8	193.139

Figure 1: The collocation results for the word *punë*, eng. *work*, calculated based on Log-likelihood, by the CQPweb.

Another dictionary type that offers collocation candidates is a dictionary of synonyms, e.g. Thomai et al. (2004) for Albanian. Below – as an example, in Figure 3, marked with the sign ◆and the name of the dictionary – are listed collocations which are taken from the mentioned dictionary. If the data are available in electronic form, then a fast search and extraction of any lexical information is possible, otherwise the only remaining option is to gather it manually.

Combining all three methods, i.e. n-grams, CQPweb and extracting information from traditional dictionaries, makes it possible to collect a lot of lexical information for certain lexical entries which can be used for compiling a lexicon of collocations

5. Selecting the lexicon entries and their types

The collected lexical information needs to be organized in lexical entries. The selection of lexical entries is in some cases very difficult, especially the decision of whether to select an entry at all. Furthermore, the selection of collocation candidates (for the explication part of the entry) is not easy and depends on many criteria. We start with the entries that are semantically related and continue with those that are very frequent. But not every frequent cooccurrence is chosen for a lexical entry, as explained in the above case of *një* and *kjo*, in the section CWB & CQPweb.

Based on other collocation dictionaries, the lexical entries are differentiated according to their part-of-speech. Also, the grammatical relations between collocated words are important to organize the mezzo- and micro-structure of an entry, e.g. Noun–Adjective or Verb–Noun. This is described in the followed sections based on examples.

6. Example Entries

6.1. Nouns

Noun entries are organized as in Figure 2. The entry begins with its head, followed by its part-of-speech. The collocations, without an example of where they occur in texts, are listed as a sequence, separated by commas. This sequence can be separated by a pipe (|) and <*jo*, *pa*, *i*, … [i.e. negations]> for words of opposite meaning and bullets (• i.e. very strong, · i.e. less strong), if the collocations can be grouped/assorted/-separated in sense of meaning. The letter *i*, e.g. in *i shkurtër* (eng. *short*) is the article (determiner) of the masculine gender of the adjective *shkurtër*. The feminine gender is *e*, which is not written. The explication part contains also the collocations with its prepositions, which are listed after the mark □■ , e.g. *para ~i* (i.e. *para afati*), eng. *before the deadline.* The next sign ♦◊ marks the word compounds, e.g. *afatgjatë* (from *afat* + *~gjatë*), eng. *long term.*

Some dictionaries, e.g. Häcki Buhofer et al (2014), also list examples of authentic use for each word. For the work presented here, currently no examples are taken into account, i.e. no such examples are contained in the draft version of the dictionary. Example sentences may still be included in a future on-line version of the dictionary.

afat E (eng. *deadline*, Noun)

+MBE (eng. *Adjective*)
~ i shkurtër | *~ i gjatë* · *~ mesatar* • *~ i kaluar* · *~ i tejkaluar*, *~ i skaduar* · *~ i <pa>mbaruar* • *~ i shtyrë* · *~ i vazhduar* · *~ i zgjatur* | *~ i shkurtuar* •*~ i <pa>përshtatshëm* • *~ i <pa>caktuar* • *~ i <pa>detyrueshëm* • *~ i shlyer* [...]

+F (eng. *Verb*)
shtyej ~ • *respektoj ~* • *(tej)kaloj ~* • *mbaron ~* • *vjen ~* · *afrohet ~* • *caktoj ~* · *vazhdoj ~* [...]

□■ (i. e. used with prepositions, e.g. *me*, *pa*, ...)
me ~, *pa ~*, *para ~it/~ës*, *pas ~it/~ës*, *përtej ~it/~ës* [...]

♦ ◊ (i. e. word-formation)
~caktim, *~vënie* • *~shtyrje* • *~kalim* [...]

>MBE
~shkurtër, *~mesëm*, *~gjatë*, *~caktuar*, *~shtyrë* [...]

Figure 2: The working version of the lexical entry for the collocation afat, eng. deadline, as a result of evaluating the n-grams, using the CQPweb, calculated based on Log-likelihood, and informed by the traditional dictionaries.

Polysemic collocations, those with different senses, are listed separately, as shown in the following entries:

bar, ~i 1 E (eng. *grass*)

+MBE
i njomë | *i tharë* · *i thatë* • *i rritur* • *i gjelbërt* • *i mbirë* · *i mbjellur* • *i prerë*, *i kositur* • *i mbledhur* • *i rrëzuar* • *i mirë* • *i keq* [...]

bar, ~i 2 E (eng. *medicament*)

+MBE
shërues · *qetësues* • *i mire* [...]

□■
kundër dhimbjes • *kundër kollës* · *kundër ftohjes* [...]

bar, ~i 3 E (eng. *bar*)

+MBE
i <pa>njohur • *i frekuentuar (shumë* | *pak)* [...]

□■
~i më i afërt • *i të rinjve* [...]

Figure 3: The working version of the lexical entry for the collocation *bar*, eng. *grass*, *medicament, bar,* as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

Putting together all lexical data gathered from n-grams, cf. figure 1, through CQPweb and extracted from traditional dictionaries, the following entry can be created:

çaj, ~i E (eng. *tea*)

+MbE
i nxehtë | *i ftohtë* • *i ëmbël* | *i hidhur* • *i fortë*, *i rëndë* | *i lehtë*, *i lig* • *i zi* • *i gjelbër* • *mjekësor* [...]

+ E ABL
mali · *bjeshke* · *frutash* · *trëndafili* · *kaçeje* · *bliri* · *dafine* · *murrizi* · *rozmarine* · *borzilloku* · *eukalipti* · *kanelle* · *alku* [...]

□■
me sheqer · *me mjaltë* • *me limon* [...]

Figure 4: The working version of the lexical entry for the collocation *çaj*, eng. *tea,* as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

6.2. Verbs

Verb entries have the same structure as the noun entries. The head of the entry has – like most Albanian dictionaries – the grammatical information on aorist and participle, in addition to the part-of-speech information. The following example shows an entry of a verb.

kry|ej ~eva, ~yer F (eng. *end*, *complete, finish*)

+E
një punë · *një punim* • *një detyrë* · *një detyrim* • *një porosi* • *një vepër* · *një veprim* • *një aksion* • *një shkollë* · *një studim* • *një udhëtim* • *pushimin* DET [...]

+NdF
mirë | *keq* • *shpejt* | *ngadalë* [...]

Figure 5: The working version of the lexical entry for the collocation *kryej*, eng. *complete successfully,* as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

6.3. Adjectives

The number of these entries is smaller than the numbers for nouns and verbs. Adjectives are more often listed as collocates of the nouns. Lexical entries of adjectives can be created from a reverse index of them. A lexical entry of an adjective looks as follows:

privat ~e MbE (eng. *private*)

+E
punë • *çështje* • *interes* • *lidhje* • *jetë* • *shtëpi* · *banesë* • *makinë* · *pajisje* • *udhëtim* [...]

Figure 6: The working version of the lexical entry for the collocation *privat*, eng. *private,* as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

6.4. Adverbs

The number of adverb entries is currently very small. Similar to the adjective entries, the adverb entries can be gathered from the verb entries, in addition to the mentioned methods used for the extraction of information to create the noun and verb entries. Most adverbs, such as *good*, can occur with a large number of verbs. A lexical entry for an adverb looks as follows:

mirë NdF (eng. *good*)

+ F
jam · *jetoj* · *kaloj* · *ndihem* • *bëj* • *di* • *kuptoj* • *kujtoj* • *shikoj* • *rri* · *pushoj* • *ha* • *flas* • *vishem* • *dukem* • *njoh* • *shkoj* | *vij* • *mendoj* • *luaj* • *mësoj* • *veproj* • *informoj* • *përshtat* • *zgjohem* · *gdhihem* • *dalloj* • *mbroj* · *ruaj* · *kujdesem* • *paguaj* • *pres* • *shfrytëzoj* • *funksionon* • *arsyetoj* • *përmbledh* • *laj* · *pastroj* • *pjek* • *gatuaj* • *siguroj* • *këqyr* · *mbikëqyr* • *sillem* • *eci* • *udhëzoj* • *shkruaj* • *them* • *filloj*, *nis* | *mbaroj*, *përfundoj* • *punoj* • *hap* | *mbyll* • *shpjegoj* • *këshilloj* • *ndaj* • *bashkoj* | *largoj* • *jap* • *përgatit* • *drejtoj* • *realizoj* [...]

Figure 7: The working version of the lexical entry for the collocation *mirë*, eng. *good,* as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

6.5. The detailed form of an entry

The following entry shows in detailed form the entry of the noun *punë*, eng. *work* and the verb *punoj*, eng. *to work.* The possible combinations are listed: N-V, e.g. *nis punën*, eng. *to begin with the work* and N-ADJ/ADV, e.g. *punë e mirë*, eng. *good work*; One important extension would be to add the prepositions and the case information as given with *punë* $_{PREP+DAT}$ *sipas ligjit*, or *filloj/nis* $_{ACC}$ (=*punën*) | *një* $_{NOM}$ (=*punë*).

pun|ë ~a ~ë ~ët E

+ F
(*e*) *nis* ~ <+ACC >, *filloj* • *bëj* • *ndërpres* • *vazhdoj*, *rifilloj* • *kryej*, *mbaroj* • *harroj* • *kërkoj* • *siguroj* • *pëlqej*, (*ia*) *pëlqej* ~ <+OBJ+DAT > • *dua* | *urrej* • *pengoj* | *nxit* • *lavdëroj* | *përqesh* • *kujtoj* • *ngadalësoj*, *zhagit*, *prolongoj* • *pezulloj* • *udhëheq*, *drejtoj* • (*ia*) *mbështet* ~ <e dikujt> • (*ia*) *këshilloj* ~ <dikujt> • (*ia*) *mohoj* ~ <dikujt> • (*ia*) *ndaloj* ~ <dikujt> [...] *shtyj* ~ *përpara*, • (*i*) *fle* ~ <dikujt> [...]

+ MbE
e mirë | *e keqe* • *e vështirë* · *e rëndë* | *e lehtë* • *e shpejtë* | *e ngadaltë* • *e shumtë* | *e paktë* • *e ndryshme* | *e njëjtë* • *e gjatë* • *kujdestare* · *kujdestarie* • *nate* | *dite* · *mbrëmjeje* | *mëngjesi* • *legale* | *ilegale* • <*jo*>*serioze* • *tinëzare* • *publike* | *private* • *e madhe* | *e vogël* • *e lirë* | *e shtrenjtë* • *e* <*pa*>*ndershme* • *e* <*pa*>*ngutshme*, *e*

<pa>nxitueshme, e <pa>nxituar • vullnetare • e paparë | e mirënjohur • <jo>profesionale • amatore • kuptimplote • e <pa>këndshme • e pistë · e ndyrë | e pastër • e <pa>ditur · e <pa>njohur • e <pa>vlefshme • e mirëfilltë, e kënaqshme • e <pa>vëmendshme • e <pa>kuptueshme • <jo>detyruese • e <pa>përshtashme • <e parë, dytë, e tretë, ...> • e qetë • e <pa>ndërprerë • e <pa>kryer • e <pa>rregullt • e ligë • e <pa>shëndetshme • e mundimshme • e <pa>rrezikshme • e frikshme • e <pa>parëndësishme • e <pa>drejtë • e gatshme • e dështuar • e <pa>zakonshme | e jashtëzakonshme • <jo>precize • <jo>sistemore • e kotë • <in>formale • <jo>normale • e <pa>përfunduar, e posapërfunduar • e përkryer • <jo>humane • inxhinierike • edukuese • sezonale • e nisur, e posanisur • e lënë përgjysmë • marramendëse • e marrëzishme • e lodhshme, lodhëse • bujqësore · blegtorale • ndihmëse · plotësuese, mbështetëse • kryesore, kyçe • dytësore • banale • fisnike • tregtie • zejtare • aktive | pasive • madhështore, e mrekullueshme • poshtëruese, e poshtër, • patriotike, atdhetare, atdhedashëse, frikësuese • rraskapitëse • eksploatuese • e zezë • e mbarë • e prapë • diletante • minimale | maksimale • e dënueshme • përfituese • përgatitore • përmbyllëse • intensive, e sistemuar • artistike • sociale • mendore • motivuese [...]

+ Prep

+NOM

<me | pa> ligj • <me | pa> normë • <me | pa> rregull • <me | pa> vullnet • <me | pa> hamendje • <me | pa> dyshim • <me | pa> marrëveshje • me orë të shumta • pa u lodhur • <me | pa> përtesë • <me | pa> shije • <me | pa> dinjitet • <me | pa> kuptim • në të zezë · <me | pa> letra · <me | pa> dokumente • <me | pa> detyrim • <me | pa> leverdi • <me | pa> plan • pa fund • <me | pa> fat • <me | pa> rëndësi • <me | pa> pagesë • <me | pa> para • <me | pa> kujdes, (GgK. *pa lidhje)* · *pa ide • <me | pa> nxitim · <me | pa> ngut • <me | pa> nder* [...]

+DAT

sipas ligjit · sipas rregullave · sipas normës · sipas planit [...]

+E

fillestari/eje · amatori/eje • profesionisti/eje • diletanti/eje • dreqi • gomari • fëmijësh · të rinjsh · djemsh | vajzash · burrash | grash • dimri · pranvere · vere · vjeshte · fshati • ndërtimtarie • hajduti, hajni · rrugaçi • dembeli, përtaci • pasioni · qejfi • trimash, trimërie [...]

+ F

filloj, nis +ACC+DET, *një* +NOM+INDET • *mbaroj* +ACC+DET • *humb* ACC+DET • *ndërpres* ACC+DET • *kujdesem për* NOM+INDET • *kërkoj* (*një*) NOM+INDET • *dua* (*një*) NOM+INDET, *nuk dua* NOM+INDET • *mendoj për* (*një*) NOM+INDET [...]

-dhënës/-je • -marrës/-je • -kërkues/-im • -prishës/-je • -ndreqës/-je • -gjetës/-je • -kryes/-erje [...]

pun|oj ~ova ~uar F

+ NdF

mirë | keq • shpejt | ngadalë • shumë | pak • ndryshe · kështu · ashtu • lirë | shtrenjtë • gjatë • kot • si i çmendur • vetëm • <i>legalisht • <jo>seriozisht • tinëzisht • privatisht | publikisht • falas • natën | ditën • mëngjeseve | mbrëmjeve • të dielave • së mbari, së prapthi <jo>sistematikisht • intensivisht • rëndë • pastër • qetësisht • i <pa>stresuar • i <sh>qetësuar • i <pa>pakoncentruar • i vetmuar • fizikisht • vullnetarisht [...]

+ Prep

+NOM

deri vonë · gjatë festave • <me | pa> kujdes, (GgK. pa lidhje) · pa ide • <me | pa> nxitim · <me | pa> ngut • <me | pa> ligj • <me | pa> normë • me plan • <me | pa> rregull • <me | pa> para, <me | pa> pagesë • <me | pa> vullnet • <me | pa> hamendje • <me | pa> dyshim • me orë • me ditë • <me | pa>

marrëveshje • *me orë të shumta* • *<me | pa> përtesë* • <me | pa> *vëmendje* • *nën tarifë* • *me qetësi, në qetësi* • *<me | pa> kokë/krye* • *<me | pa> mend* • *<nën | pa> presion* • *tërë ditën | tërë natën* • *<me | pa> dëshirë* • *në <ndërtimtari ...>* • *për* <dikë+ACC> • *si <inxhinier , mjek ...>* • *si i pavarur* • *si udhëheqës* • *si ndihmës* [...]

+DAT
sipas ligjit · *sipas rregullave* · *sipas normës* · *sipas planit* [...]

+ FInf
<pa / duke> u ngutur · *<pa / duke> u nxituar* • *pa u lodhur* [...]

IDM:
si gomar · *si kalë* • *sa* (*për*) *<dy, ...> vetë* [...]

FRZ:
ia $_{\text{DAT+ACC}}$ *punoj* (*keq*/\$*mirë*) <dikujt $_{\text{DAT}}$ > [...]

Figure 8: The working version of the lexical entry for the collocation *punë*, eng. *work*, in the detailed, more extensive version, as a result of evaluating the n-grams, using the CQPweb statistics, and informed by the traditional dictionaries.

The example above contains all data collected for the entries and the current state of the work on lexical entries. In general, the aim is to keep the entries shorter, but they can "grow" to be very detailed.

7. Conclusions

Currently, the number of lexical entries is around 2000, with 40 to 110 entries for each letter of the Albanian alphabet. A number of entries are not yet complete with all their possible information, i.e. the work on these entries is not finished yet. A few of them will be deleted, while some new entries will presumably be added in the ongoing process of reviewing the lexical entries during the work with data. The first results have been encouraging.

8. References

Benson, M., Benson, E. & Ilson, R. F. (2010). *The BBI Combinatory Dictionary of English. Your guide to collocations and grammar.* Third edition revised by Robert Ilson. Amsterdam, Philadelphia, John Benjamins.

Evert, S. (2013). Tools for the acquisition of lexical combinatorics. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography (HSK 5.4)*, 104. Berlin & New York: Mouton de Gruyter, pp. 1415–1432.

Dhrimo, A., Tupja, E. & Ymeri, E. (2007). *Fjalor sinonimik i gjuhës shqipe (= Dictionary of Synonyms of the Albanian Language).* Tiranë: Toena.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. In: *International Journal of Corpus Linguistics* 17 (3), pp. 380–409.

Häcki Buhofer, A., Dräger, M., Meier, S. & Roth, T. (2014): *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag.* Tübingen: Francke. Online: https://kollokationenwoerterbuch.ch/.

Herbst, T. (1996). What are collocations: sandy beaches or false teeth. *English Studies*, 1996, pp. 379–393.

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

Kabashi, B. (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, pp. 123–132.

Kabashi, B. & Proisl, T. (2018). "Albanian Part-of-Speech Tagging: Gold Standard and Evaluation". In N. Calzolari et al. (eds.) *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan.* European Language Resources Association (ELRA) Paris, pp. 2593–2599.

Kabashi, B. (2019). A lexicon of Albanian for Natural Language Processing. In R. H. Gouws, U. Heid, T. Herbst, S. Schierholz & W. Schweickard (eds.) *Lexicographica, International Annual for Lexicography, Vol. 34*, pp. 233–242.

Kostallari, A. (kryeredaktor), Thomaj, J., Lloshi, X., & Samara, M. (1980). *Fjalor i gjuhës së sotme shqipe* (= *Dictionary of Contemporary Albanian Language*). Tiranë: Akademia e Shkencave e RPS të Shqiperisë.

Proisl, T. (2019). *The Cooccurrence of Linguistic Structures.* Erlangen: FAU University Press.

Quasthoff, U. (2010). *Wörterbuch der Kollokationen im Deutschen.* Berlin, etc.: de Gruyter.

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

Sinclair, J. M. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Thomai, J., Samara, M., Shehu, H. & Feka, T. (2004). *Fjalori sinonimik i gjuhës shqipe* (= *The Dictionary of Synonyms of the Albanian Language*). Tiranë: Akademia e Shkencave e Republikës së Shqipërisë.

Thomai, J., Samara, M., Haxhillazi, P., Shehu, H., Feka, T., Memisha, V. & Goga, A. (2006). *Fjalor i gjuhës shqipe* (= *Dictionary of the Albanian Language*). Tiranë: Akademia e Shkencave e Republikës së Shqipërisë.