Bootstrapping the Albanian Information Retrieval

Nikitas N. Karanikolas
Department of Informatics
Technological Educational Institute (TEI) of Athens
Athens, GREECE
e-mail: nnk@teiath.gr

Abstract — **In this paper we investigate the Albanian language and try to uncover the characteristics of the language that will permit the Information Retrieval (IR) community to develop IR systems adapted for the specific language. As a consequence of our study (investigation) we provide a naive-single-step (rudimentary) stemming algorithm for the Albanian language. A stopword list is also created. Human experts are contacted for the evaluation of the provided stemming algorithm. The evaluation method used and the observation of the method's results uncover more rules, which could improve the capabilities of the rudimentary stemming algorithm. We believe that our approach for this specific language could become a standard way for building Information Retrieval functionalities (tools, functions, etc) for languages less perused, as is the language studied in this paper.**

Keywords – information retrieval; stemming algorithm; stopword list

I. INTRODUCTION

The Albanian language is not a language perused by the Information Retrieval (IR) [1, 2, 5, 6, 9, 10] community. It is based on the Latin alphabet but introduces special markings (diacritical marks) for some of the Latin alphabet letters. As a consequence of these observations, Search Engines (Web Retrieval) engages the rules and methods defined for the lingua franca of IR, English.

In our awareness of the domain, there are neither adapted desktop IR applications nor adapted web search engines for the Albanian language (a fact exist also in other languages [7]). Since Information Retrieval is a considerable instrument for literature research, we appraise that Albanian language awareness of IR, should also be given attention. Moreover, since we are not speakers of the Albanian language, our approach could be a prescription (formula) for adapting desktop IR applications and Web search engines for other languages also.

In the following sections, we are going to locate and discuss the Albanian language's characteristics. Next, we will compose a stopword list, based on lingual characteristics (not statistical approaches). Next section, 4, will be devoted to the localization of the most common suffixes of the Albanian language and the creation of a basic Albanian word-stemming algorithm. Stemming algorithms are used to conflate word variants. Since for evaluation purposes we need an Albanian corpus (document collection), we have addressed this need and we describe our collection in section five. Section six presents our approach and the results of the evaluation of the basic Albanian word stemmer. Last section presents the results of our effort and some future extensions.

II. THE ALBANIAN LANGUAGE

The Albanian alphabet uses twenty-five (25) out of twenty-six (26) letters of the English alphabet (does not use the letter W); it also uses the latin letter E with the diacritical mark of Diaeresis (Ë), the latin letter C with the diacritical mark Cedilla (Ç) and nine couples of letters (used as a single letter). A total of 36 letters constitutes the Albanian alphabet. Table I contains the couples of letters, used as single letters, in the Albanian language.

TABLE I. COUPLES OF LETTERS, USED AS SINGLE LETTERS

dh	gj	ll	nj	rr	sh	th	xh	zh

There is no code page defined specifically for the Albanian language. However, the ISO-8859 part 16 (Latin Alphabet No 10) is intended to cover South-eastern European languages, i.e. Albanian, Croatian, Hungarian, Italian, Polish, Romanian and Slovenian.

Moreover, the current state of system utilities, does not utilize the Latin Alphabet No 10 of ISO-8859. For example, the Windows Notepad offers the selection between the scripts: Greek, Western, Hebrew, Arabic, Turkish, Baltic, Central European, Cyrillic and Vietnamese. Otherwise, this utility can handle only Latin/Greek, Latin-1, Latin/Hebrew, Latin/Arabic, Latin-5, Latin-4, Latin-2, Latin/Cyrillic and CP1258 codepages, respectively.

A better solution for Albanian language text storage is the utilization of the Unicode (ISO 10646-1 [3]) standard. The most thrifty storage method for Unicode texts is the UTF-8 variable length character encoding. In table II, we provide the complete Albanian alphabet, both upper and lower case, with their encoding (decimal and hexadecimal), utilizing UTF-8 (rfc3629).

The Albanian nouns are inflected by gender (masculine, feminine and neuter), number (singular and plural) and case (nominative, accusative, genitive, dative, ablative and vocative). However the vocative case is rarely used. The cases apply to both definite and indefinite nouns. The definite article is placed after the noun and can be in the form of noun suffixes, which vary with gender and case. It is obvious that the mentioned language rules necessitate a long number of variations for a single noun.

The Albanian verbs are governed by six (6) type of moods, eight (8) tenses (three simple and five complex) and two (2) voices (active and non-active). Obviously, a single verb has a great number of variations to support the rules that govern Albanian verbs.

The previous two conclusions (regarding the variation of nouns and verbs) make clear the need for some method to replace all inflected forms of a noun with a common (for all forms of the noun) theme. Also, an analogous method is needed for verbs. These can be provided either with a lemmatization lexicon or with some stemming algorithm.

TABLE II. THE COMPLETE ALBANIAN ALPHABET

Upper case letter	DEC	HEX	Lower case letter	DEC	HEX	Vowel (V) or Conso-nant (-)
A	65	41	a	97	61	V
B	66	42	b	98	62	-
C	67	43	c	99	63	-
Ç	195:135	C3:87	ç	195:167	C3:A7	-
D	68	44	d	100	64	-
DH			dh			-
E	69	45	e	101	65	V
Ë	195:139	C3:8B	ë	195:171	C3:AB	V
F	70	46	f	102	66	-
G	71	47	g	103	67	-
GJ			gj			-
H	72	48	h	104	68	-
I	73	49	i	105	69	V
J	74	4A	j	106	6A	-
K	75	4B	k	107	6B	-
L	76	4C	l	108	6C	-
LL			ll			-
M	77	4D	m	109	6D	-
N	78	4E	n	110	6E	-
NJ			nj			-
O	79	4F	o	111	6F	V
P	80	50	p	112	70	-
Q	81	51	q	113	71	-
R	82	52	r	114	72	-
RR			rr			-
S	83	53	s	115	73	-
SH			sh			-
T	84	54	t	116	74	-
TH			th			-
U	85	55	u	117	75	V
V	86	56	v	118	76	-
X	88	58	x	120	78	-
XH			xh			-
Y	89	59	y	121	79	V
Z	90	5A	z	122	7A	-
ZH			zh			-

III. CONSTRUCTING A STOPWORD LIST

A commonly used technique in Information Retrieval (IR) Systems is the elimination of stopwords from both queries and index files (usually Inverted files) [10]. Stopwords are terms that appear very frequently in documents, and consequently their existence (in queries and index files) does not help the IR System to discriminate between relevant and irrelevant, for a query, documents. Therefore, the system should eliminate such words, during the indexing and querying phases. There are two main approaches to select stopwords. One uses statistical measures in huge document collections (corpus), while the other uses language resources to select the stopwords. In the former approach the very frequent terms are characterized as stopwords. In the later approach, articles, prepositions, conjunctions, pronouns and some auxiliary verbs are considered as candidates for the construction of a stopword list.

In our approach for creating a stopword list we adopted the language resource approach. The first reason for our decision was that the Albanian language is not a so proliferated language and it is difficult to create a huge corpus. Secondly, there are some words that are not very frequent (e.g. reflexive and reciprocal pronouns) but they are completely unable to discriminate between relevant and irrelevant documents, for any query. Thus, these words cannot be removed by a statistical approach. On the other hand, language resources, at least hard copy books, exist for the considered language. We have considered the following books:

- Gjuha shqipe 8, ISBN 978-99956-625-1-6, by Mimoza Gjokutaj and Abdullah Ballhysa
- Gjuha shqipe 7, ISBN 99943-883-6-3, by Rita Petro, Idris Metani, Shpresa Vreto and Adelina Çerpja
- Gjuha shqipe 6, ISBN 978-99943-945-6-2, by Rita Petro, Natasha Pepivani, Adelina Çerpja and Mimoza Gjokutaj
- Gjuha shqipe 5, ISBN 99943-2-103-x, by Bahri Beci, Loredan Bubani and Zamira Gurabardhi
- Gjuha shqipe për të Huajt dhe Shqiptarët jashtë Atdheut, ISBN 99927 1 454 9, by Gjovalin Shkurtaj and Enver Hysa

From these books (mostly from Gjuha shqipe 6 and Gjuha shqipe 7), we derived conjunctions (lidhëzat), pronouns (Përemrat), Adverbs (Ndajfoljet), Particles (Pjesëza), Prepositions (Parafjala), some adjectives (Mbiemra) and Exclamations (Pasthirrmat). These, formed our stopword list (about four hundreds and seventy words). However, some of the list contents are a subject for disagreement and possibly should be removed (i.e. Adverbs).

Figure 1 is an excerpt from page 222 of the book Gjuha shqipe 6. It is a subset of the pronouns (përemrat) presented in the book.

- **vetor**: *unë, ti, ai, ajo, ne, ju, ata, ato* (***Unë*** *për vete e mora librin.*)
- **vetvetor**: *vete, vetja, vetvetja* (*Unë për* ***vete*** *e mora librin.*)
- **dëftor**: *ky, kjo, ai, ajo, i / e atillë, i / e këtillë ...* (***Këtë*** *librin këtu.*)

Figure 1. A subset of Albanian pronouns.

IV. BUILDING A NAIVE-SINGLE-STEP STEMMING ALGORITHM

Between a lemmatizer (lemmatization lexicon) and a stemming algorithm [4, 8, 11] for the normalization of words, we have concluded in favor of the stemming algorithm. In our approach for building a stemming algorithm for the Albanian language, we have followed a rudimentary approach. This is because our purpose was to provide the base for further development and also a yardstick. Our approach does not follow multiple steps of replacements, like the Porter's stemming algorithm [8]. It simply removes suffixes and prefers to remove the longer matching ones. To do so, we needed a list of the suffixes used in the Albanian language, as a result of the language's inflectional system. Our resources to derive the Albanian language suffixes were the same language learning and grammar books, listed in the previous section.

Figure 2 is an excerpt from page 216 of the book Gjuha shqipe 8. It is the declensions of three verbs (la/wash, hap/open and vën/put) in past tense (E pakryera), in active and passive voice (Forma veprore and Forma joveprore, respectively) with the mood of astonishment (Mënyra habitore).

Forma veprore

Zgjedhimi I	Zgjedhimi II	Zgjedhimi III
lakësha	hapkësha	vënkësha
lakëshe	hapkëshe	vënkëshe
lakësh	hapkësh	vënkësh
lakëshim	hapkëshim	vënkëshim
lakëshit	hapkëshit	vënkëshit
lakëshin	hapkëshin	vënkëshin

Forma Joveprore

Zgjedhimi I	Zgjedhimi II	Zgjedhimi III
u lakësha	u hapkësha	u vënkësha
u lakëshe	u hapkëshe	u vënkëshe
u lakësh	u hapkësh	u vënkësh
u lakëshim	u hapkëshim	u vënkëshim
u lakëshit	u hapkëshit	u vënkëshit
u lakëshin	u hapkëshin	u vënkëshin

Figure 2. Declensions of verbs in past tense

According to the rule presented in Figure 2, the suffixes: kësha, këshe, kësh, këshim, kështit and kështin; are used to form the declensions. These suffixes are the constituents with numbers 51, 52, 118, 15, 17 and 16 of our suffix list.

Figure 3 contains three excerpts from page 249 of the book Gjuha shqipe 8. They provide rules that govern the orthography of noun variations corresponding to number, gender inflections for nouns ending with: si, ri, ar, or, tor and tar.

According to the rules presented in Figure 3, the suffixes: si, ri, ar, or, tor and tar; are included in our suffix list. These suffixes are the constituents with numbers 249, 334, 330, 280, 325, 248, 243 of our suffix list.

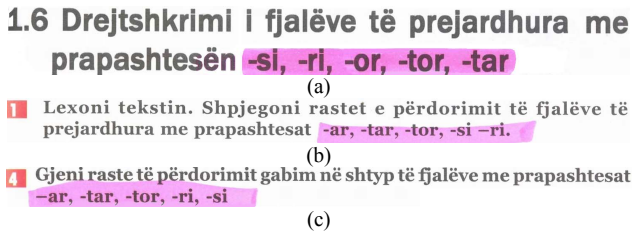

Figure 3. Rules for the orthography of noun variations

V. CORPUS CREATION

In order to evaluate our stemming algorithm, we needed a collection of documents (texts) in the Albanian language. Since we weren't able to find any existing corpus-based computational linguistics resource for the Albanian language, we decided to create our own one. Thus, we collected documents from the Internet. Our resources (URLs) follow:

http://agonek.blogspot.com/
http://www.yllpress.com/594/rracizem-pertej-atlantikut-.html
http://www.zeriyt.com/revista-shkencore/shkenca-e-fshehtesise-t48577.0.html
http://historia.shqiperia.com/rilindja/kreu_14.php
http://news.himara.eu/2009/01/legalizimi-i-dokumenteve-n-ambasadn.html

We have collected a total of fifteen (15) documents from the above locations. This collection of fifteen documents comprises our (small) corpus. The subjects of documents vary between Secondary School texts (two documents), University texts (four documents), News Reports (five documents) and Literature (four documents). The document sizes vary from (about) a hundred and twenty (120) to three thousands (3000) words. The average word size is (about) a thousand and five hundreds (1500).

After the removal of stopwords, the documents of our corpus vary from (about) seventy (70) to a thousand and two hundreds (1200) words. The average non-stopword size is (about) six hundreds and sixty (660) words. In other words, stopword removal reduces the document size to a 44% of the original document (word) size. Removing double word occurrences (in their inflected form) reduced the total number of used words to five thousand (5000).

TABLE III. A SMALL PART OF OUR EVALUATION SET

sn	Inflected word	stem
2487	MËNGJEZIN	MËNGJEZ
2488	MËNJANONTE	MËNJAN
2489	MENJËHERSHËM	MENJËHER
2490	MENTOR	MEN
2491	MËNYRA	MËNY
2492	MËNYRAVE	MËNYR
2493	MËNYRË	MËNYR
2494	MEPARSHEM	MEPARS
2495	MERAKUN	MERAK

Next, the application of our stemming algorithm provided an equivalent number of couples (each couple having an original-inflected word and the suggested stem). This collection of (five thousand) couples is our evaluation set. Table III provides a small part of our evaluation set.

VI. Evaluation of the Stemming Algorithm

From our evaluation set (the collection of five thousand couples) we randomly selected five hundred couples (actually ten groups of fifty continuous couples) and asked Information Retrieval experts to provide their judgments for the applicability of the provided stems. Since the IR experts were not familiar with the Albanian language, we provided the Greek translation of the original Albanian word, as the word's semantics. The experts had to fill a cell (next to the translation's cell) only when they disagreed with (have any argument against) the provided stem. We have defined two types of argumentation: ΚΡ (Κοινή Ρίζα, common stem) when our algorithm provides different stems for a group of (two or more adjacent) words and the IR expert assess that the whole group should have a common stem; ΔΡ (Διαφορετική Ρίζα, different stem) when our algorithm provides the same stems for a group of (two or more adjacent) words and the IR expert assess that some (at least one) of the words of the group should have a different stem.

The alternative way to evaluate our stemming algorithm (and also our stopword list) could be to place the algorithm and the stopword list in an Information Retrieval System and evaluate the System (calculate recall and precision [10]) with and without them. However, in our opinion, this is an evaluation method to compare competitive (and also black box) solutions and not an evaluation method of a solution under development. In the later case (for a stemming algorithm under development), the IR experts' assessments are valuable information needed for the improvement of the stemming algorithm.

TABLE IV. Assessments of IR experts

sn	Inflected word	stem	English translation	The expert's assessment
2487	MËNGJEZIN	MËNGJEZ	The morning	
2488	MËNJANONTE	MËNJAN	The verb discern in the past tense and in the third person	
2489	MENJËHERSHËM	MENJËHER	Directly	
2490	MENTOR	MEN	Announcer	
2491	MËNYRA	MËNY	The manner	KP (MËNYR)
2492	MËNYRAVE	MËNYR	Manner, Plural number, Genitive case	
2493	MËNYRË	MËNYR	Manner	
2494	MEPARSHEM	MEPARS	Previous	
2495	MERAKUN	MERAK	The anguish	

Thus, our IR experts (actually two Assistant Professors, and one Associate Professor) followed our suggestions and assessed our stemming algorithm. Table IV presents a small part of the assessment returned by one of them. (We have replaced the Greek translation with the English equivalent.)

According to table IV, the IR experts suggest that the stems provided by the algorithm for words with serial numbers 2487, 2488, 2489, 2490, 2494 and 2495 are satisfactory; the words with serial numbers 2491, 2492 and 2493 should have the common stem MËNYR. Thus in this small part the algorithm has eight (8) satisfactory, out of nine (9) cases (about 89% success). The success rate for the whole set (of the translated five hundred couples) is 80%.

VII. Conclusions and Future Work

Obviously, a percentage of 80% of correctly assessed stems is an encouraging factor for our research. However, we do not claim that we have built a perfect stemming algorithm for the Albanian language. On the contrary, we believe that there is a long way to approach the perfect algorithm. However, our contribution is the bootstrapping step for the development of Information Retrieval Systems adapted for the Albanian language.

The way we have approached the evaluation of our stemming algorithm gave us considerable information for its improvement. For example, one of the drawn deductions is that: whenever a stem ends with a couple of two consecutive consonants and these consonants do not belong to the nine couples of the Albanian alphabet used as single letters, we can apply a second step and remove the final consonant of stem. This deduction applies to and resolves five cases that the IR experts provided argumentation, without creating any problem to the rest assessed couples (original word and stem) of our evaluation set.

Considering further the IR experts assessment gave rise to some conjectures. We believe that some of the endings included in our list, if removed, will further improve the stemming algorithm. However, this is not a clear deduction, as is the case for the stems ending with a couple of two consecutive consonants, and should be tested. We are planning to implement an intelligent method that will be fed with the experts' judgments regarding the results of the naive single-stem-suffix-removal algorithm, the deductions that humans draw by studying the judgments and the conjectures and will try to compose an improved stemming algorithm. In other words we are planning to create a trial-and-test algorithm that will mine a better algorithm than the naive single-stem-suffix-removal stemming one.

Acknowledgment

I would like to thank my colleagues Eleni Galiotou, Sofia Stamou and Ioannis Voyiatzis. I also thank my students Ellie Chavari and Olga Ntavarinou for their great patience to follow my instructions for one-year (and maybe longer) period. Without their help, this work would still be in my mind.

References

[1] G. Canfora, L. Cerulo, "A Taxonomy of Information Retrieval Models and Tools", Journal of Computing and Information Technology, vol. 12, pp. 175-194, 2004.

[2] D.C. Deal, “Techniques of Document Management: A review of Text Retrieval and related technologies”, Journal of Documentation, vol. 57, pp. 192-217, 2001.

[3] ISO 10646-1, Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane, 1993.

[4] T.Z. Kalamboukis, “Suffix stripping in Greek”, Program, vol. 29, pp.313-321, 1995.

[5] N.N. Karanikolas, “Low cost, cross language and cross-platform Information Retrieval and Documentation tools”, Journal of Computing and Information Technology, vol. 15, pp. 71-84, 2007.

[6] G. Kowalski, Information Retrieval Systems. Theory and Implementation. Kluwer Academic Publishers, USA, 1997.

[7] F. Lazarinis, “Web retrieval systems and the Greek language: do they have an understanding?”, Journal of Information Science, vol. 33, pp. 622-636, 2007.

[8] M. Porter, “An Algorithm for suffix stripping”, Program, vol. 14, pp. 130-137, 1980.

[9] C.J. Van Rijsbergen, Information Retrieval. Butterworths, London, UK, 1979.

[10] G. Salton, Introduction to Modern Information Retrieval. McGraw-Hill, USA, 1983.

[11] G. Tambouratzis, C. Carayannis, “Automatic Corpora-based Stemming in Greek”, Literacy and Linguistic Computing, vol. 16, pp. 445-466, 2001.