

Albanische Forschungen

Begründet von
Georg Stadtmüller

Für das Albanien-Institut
herausgegeben von
Peter Bartl

unter Mitwirkung von
Bardhyl Demiraj, Titos Jochalas und
Oliver Jens Schmitt

Band 44

2020

Harrassowitz Verlag · Wiesbaden

Altalbanische Schriftkultur

- aus der Perspektive
der historischen Lexikographie
und der Philologie der Gegenwart -

Akten der 6. deutsch-albanischen
kulturwissenschaftlichen Tagung
(27. September 2019, Buçinas bei Pogradec, Albanien)

Herausgegeben von
Bardhyl Demiraj

2020

Harrassowitz Verlag · Wiesbaden

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen
Nationalbibliografie; detaillierte bibliografische Daten sind im Internet
über <http://dnb.de> abrufbar.

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

Informationen zum Verlagsprogramm finden Sie unter
<http://www.harrassowitz-verlag.de>

© Otto Harrassowitz GmbH & Co. KG, Wiesbaden 2020

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt.
Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne
Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere
für Vervielfältigungen jeder Art, Übersetzungen, Mikroverfilmungen und
für die Einspeicherung in elektronische Systeme.

Gedruckt auf alterungsbeständigem Papier.

Druck und Verarbeitung: KN Digital Prinforce GmbH, Erfurt

Printed in Germany

ISSN 0568-8957

ISBN 978-3-447-11391-5

Dem Gedenken an
WILFRIED FIEDLER (7.5.1933 – 11.9.2019)
gewidmet

INHALT

| | |
|---|-----|
| Vorwort des Herausgebers | 9 |
| I. HISTORISCHE LEXIKOGRAPHIE UND ETYMOLOGIE | |
| OLAV HACKSTEIN | |
| The System of Negation in Albanian: Synchronic Constraints and Diachronic Explanations | 13 |
| SERGIO NERI | |
| Zur Etymologie von altalbanisch <i>nja/një</i> ‚zet ‚zwanzig‘ | 33 |
| CĂTĂLINA VĂTĂȘESCU | |
| Problèmes soulevés par l'étude de la paire de mots <i>besë</i> « croyance » et <i>fe</i> « foi » | 49 |
| ANILA OMARI | |
| Slawismen im Altalbanischen unter Berücksichtigung der neuen philologischen Editionen: "Dittionario italiano-albanese" von Da Lecce (1702 – Ms.) – eine unerforschte Quelle | 53 |
| II. HISTORISCHE LEXIKOGRAPHIE IM ZEITALTER DER DIGITALISIERUNG | |
| MARIA MOROZOVA & ALEXANDER RUSAKOV | |
| The early Albanian texts in an annotated language corpus: An attempt of processing and analysis | 91 |
| BESIM KABASHI | |
| Building diachronic corpora of the Albanian language | 103 |
| CHRISTIANE BAYER | |
| Wie wird ein Lemma digital? Vorüberlegungen zum Digitalen Philologisch-Etymologischen Wörterbuch des Altalbanischen | 109 |
| III. ALTALBANISCHE LITERATUR IM LINGUISTISCHEN UND KULTURHISTORISCHEN KONTEXT | |
| XHEVAT LLOSHI | |
| Complications in Gjon Buzuku's Lexikon | 127 |
| INA ARAPI | |
| Das Verbsystem in der Grammatik <i>Osservazioni Grammaticali Nella Lingua Albanese</i> von Padre Francesco Maria da Lecce (1716) | 133 |
| LUCIA NADIN | |
| Il "Meshari" di Gjon Buzuku. Nuovi dati, nuovi scenari | 156 |

EVALDA PACI

Osservazioni su alcune caratteristiche lessicali e fraseologiche
nelle opere della letteratura ecclesiastica albanese (1555-1743) 167

MIMOZA PRIKU

On the Traces of an Albanian Lexicon: Carlo Tagliavini
and the Albanian Lexemes of the XVI-XVIII Centuries in His Opus 180

ERZEN KOPERAJ & GËZIM PUKA

The Contribution of the Scholar Kolë Ashta
to the Critical Publication of the Early Albanian Writers 193

IV. NEUE PHILOLOGISCHE ANSÄTZE

IN DER ALTALBANISCHEN SCHRIFTKULTUR

JOACHIM MATZINGER

Das erste albanische Alphabet bei Gjon Buzuku (1555)
– ein paar Beobachtungen 205

BARDHYL DEMIRAJ – JUDITH I. HAUG

Ein frühes Zeugnis altalbanischer Schriftkultur und Musik:
Das Kompendium des 'Alī Ufukī (nach ca. 1635) 220

SOKOL ÇUNGA – BARDHYL DEMIRAJ

Ein albanisches Manuskript im Handschriftenbestand
der Klosterbibliothek der Heiligen Trinität auf der Insel Chalki 239

FATOS DIBRA

The Albanian Lexicon of Evliya Çelebi's *Seyahatname*
in the Context of Old Albanian 269

BARDHYL DEMIRAJ

Die albanischen und aromunischen Inschriften der Gravur
von Ardenica (neubearbeitet und ergänzt mit Archivmaterial
aus AQSH Tirana) 315

DORIS K. KYRIAZIS / DHORI Q. QIRJAZI

From V. Meksi's translation to G. Gjirokastriti's redaction:
towards a comparative edition of Albanian texts of NT 351

V. Wir nehmen Abschied von Wilfried Fiedler

JOACHIM MATZINGER – MONICA GENESIN

Wilfried Fiedlers Bedeutung für die Albanologie
und Balkanologie 389

Fiedlers Schriftenverzeichnis 397

BESIM KABASHI

FAU Erlangen-Nürnberg – LMU München

Building diachronic corpora of the Albanian language

Besides serving for the analysis of a language at a specific point in time, linguistic corpora can be used for the analysis of a language during a longer period of time. This approach facilitates the comparison of linguistic properties, e. g. language changes. As we had already compiled two corpora for the modern Albanian language, we proceeded to compile a corpus of the Albanian language of 16th century, i.e. the well-known manuscript of “Meshari” (Missa-
le) written by Gjon Buzuku (ital. Giovanni Buzuku), dated 1555. In this contribution, we present a brief description of an ongoing work to build diachronic corpora for the Albanian language.

1. Language corpora as base for empirical linguistic studies

The role of corpora in linguistic studies is crucial, especially in the last decades. Since the beginning of the 2000s, more and more texts have been available in electronic form. This is an invaluable resource, as corpora allow for a wide empirical base for linguistic studies.

For a modern language, in many cases it is not difficult to collect the empirical data, such as text or recorded speech, and verbal communication as audio respectively video. Moreover, numerous tools and techniques make the processing of this huge amount of collected data possible and convenient.

For diachronic language data many of these advantages are not given.

First, as the writing is a personal variant of each author, the tools must be adapted for each one, e. g. in the case of lemmatization and POS-tagging. In some cases, a manual annotation is accelerated through an adapted tool – it is, however, still more difficult than working with modern language.

2. Corpora of modern Albanian language

Until now there have only been two linguistic corpora of Albanian: on the one hand *The Albanian National Corpus of Sankt Petersburg State University* and on the other hand *The Albanian [text and speech] Corpus (AlCo)*, and the recently completed *The Albanian Corpus of press texts from the year 2017 to the year 2019 (AlCoPress 2017-2019)*, compiled at the University of Erlangen-Nuremberg. The Albanian National Corpus of Sankt Petersburg State University contains around 30 million word forms and is morphologically annotated, but not

disambiguated. It is openly accessible, cf. Morozova – Rusakov (2020, in this volume pp. 91ss.).

2.1 The Albanian text and speech Corpus (AlCo)

The Albanian Corpus (AlCo) contains a hundred million tokens (text words), which makes it the first Albanian corpus of this size. The corpus covers different domains of language and contains different text types – it is a reference corpus. The corpus is annotated with a morpho-syntactic tagset consisting of 77 tags cf. Kabashi & Proisl (2018), and the data is disambiguated. We enable the exploration of this corpus data via a Web interface to the Corpus Query Processor, (CQPweb), cf. Hardie (2012), a web-based corpus analysis system. More details about the AlCo can be found in Kabashi (2017). The corpora from the University of Erlangen-Nuremberg are currently only accessible to a restricted circle of researchers.

2.2 The AlCoPress 2017-2019

The Albanian Corpus of press texts from the years 2017–2019 (AlCoPress 2017-2019) contains around 32 million tokens (text words). AlCoPress 2017–2019 is built in the same way as the AlCo and contains some additional metadata like information on the newspaper and the topical domains (headings/categories of the newspaper texts, e. g. economy, art and sport). The corpus is annotated with the same tagset as AlCo and is also disambiguated. Analogous to AlCo, we use CQPweb to explore the corpus data. Both corpora can be merged very simply. The only reason to keep them separately was that the fusion of both corpora would disbalance the number of press texts in contrast to the other text types. However, merging the two corpora into one is an option for the future, because the collection of texts has now arrived a number of 200 million tokens.

3. Albanian diachronic corpora

Although building diachronic corpora, especially in terms of annotation and processing, is more difficult than for corpora of modern language, it is necessary to build them. The first step consists of digitizing the manuscripts/texts, which can be achieved more or less satisfactorily without being a diachronic linguist. While the digitization and the technical preparation of raw texts is can be completed in a relatively short time, the linguistic preparation of the texts, on the other hand, takes a long time, as it requires highly specialized linguistic knowledge and resources.

The *Thesaurus Indogermanischer Text- und Sprachmaterialien* (TITUS), Text Database, provided by the University of Frankfurt, Germany, is an example of collecting texts of older stages of a language, and which also includes some

Albanian texts from the 16th to the 18th century. This text collection, stored as a text database, offers search options for word forms and shows the context of occurrence of a word. The database contains *Missale (Meshari)* from Buzuku, data entry by W. Hock (Berlin), *Dictionarium Latino-Epiroticum* from Blanchus (alban. Bardhi), data entry by M. de Vaan (Leiden), *Cuneus Prophetarum* from Bogdani, data entry by M. de Vaan (Leiden), *Breve compendio della dottrina christiana* from Casasi (alban. Kazazi), data entry by B. Demiraj (Munich), *Dottrina christiana* from Matranga, data entry by M. de Vaan (Leiden). Presently, this database – in case of the Albanian language – contains only raw texts, lacking linguistic annotation. As the research group in this field, i. e. diachronic Albanian, is very small, these texts will likely not be annotated in the foreseeable future. Also, TITUS does not cover all known texts of the Albanian language of the 16th–18th century.

For further research, the word forms need to be lemmatized, linked to the respective word forms in present standard Albanian. Tagging the word forms according to their parts-of-speech or morphology would be a next step. This would make diachronic linguistic information available to a wide circle of researchers, not limited to diachronic/mediaeval linguists.

Morozova & Rusakov (2020) report that at *Sankt Petersburg State University* they started building an annotated corpus of the *Dottrina christiana* (alban. *Embsuame e krështerë*) (1592) from Matranga, which would be openly accessible for researchers. From our point of view, this is an important step in the right direction.

As after *Meshari* (1555) up to the end of 18th century followed a lot of texts, written by different authors in different times and places, not all texts can be put together. There is a need for more than one corpus. Building a diachronic corpus of old Albanian texts needs for the categorization of texts. After comparison of word form lists and consulting with Professor Bardhyl Demiraj (University of Munich), the text of *Meshari* will be processed separately as an independent corpus. The texts from the Gheg areal, the Northern dialect of Albanian, will be treated together as one corpus, as well as the texts from the Tosk areal, the Southern dialect of Albanian. Therefore, we have started collecting texts in three groups. An example for the *Old Gheg* texts is *Cuneus Prophetarum* from Bogdani, and one of the *Old Tosk* texts is *Codex Beratinus*.

3.1. The Buzuku (1555) Corpus

The Buzuku Corpus contains the raw text of *Meshari*. Each (hard) page is treated as a separate text in the corpus. This allows us to locate information, e. g. word forms, word positions etc. by the page as in the physical manuscript. The corpus is not yet linguistically annotated in any form yet. For the current first version, we replaced the four Non-Latin letters used by Buzuku

with the corresponding letters from the standard Albanian alphabet and spelling. The text itself is encoded as Unicode/UTF-8 and can be converted into other systems and formats.

As the first large document, without having a prototype orthography as an orientation, *Meshari* does not have a consistent writing of word forms. This makes even searching for a single linguistic word form in the corpus difficult, e. g. searching for *nukë/nuk* (engl. *not*).

3.2. Annotations of the texts

Which text annotations are needed and which can be realized in the next years?

Normalization of the word forms/text: Raw corpus text offers operations, e. g. searching, only based on the word form surface (i. e. string matching). For more elaborate searches, word forms that are “variants”, or realisations of the same word form, need to be linked e. g. *nukë* and *nuk*. First of all, the “normalization” of these word forms would make it possible to identify word forms – in linguistic sense – and to search for them. This is also a necessary prerequisite for further analysis.

Parts of speech (POS) annotation: Usually the first step in linguistic annotation is the POS annotation. Each word form is labeled with a POS, usually considering its context of occurrence. This type of annotation allows e. g. searching of all word forms that have the same POS, or to disambiguate while searching for a word form and the one exact POS-tag in case of ambiguity. Also, the sequences of POS-tags allows for selecting sentence or phrase types, or to simply extract different N-grams or collocations based on POS-tags.

Other possible meta text annotations can be created and used in parallel with the mentioned annotations and allow for searching relevant features respectively restrictions. What usually follows is syntactic/dependency parsing.

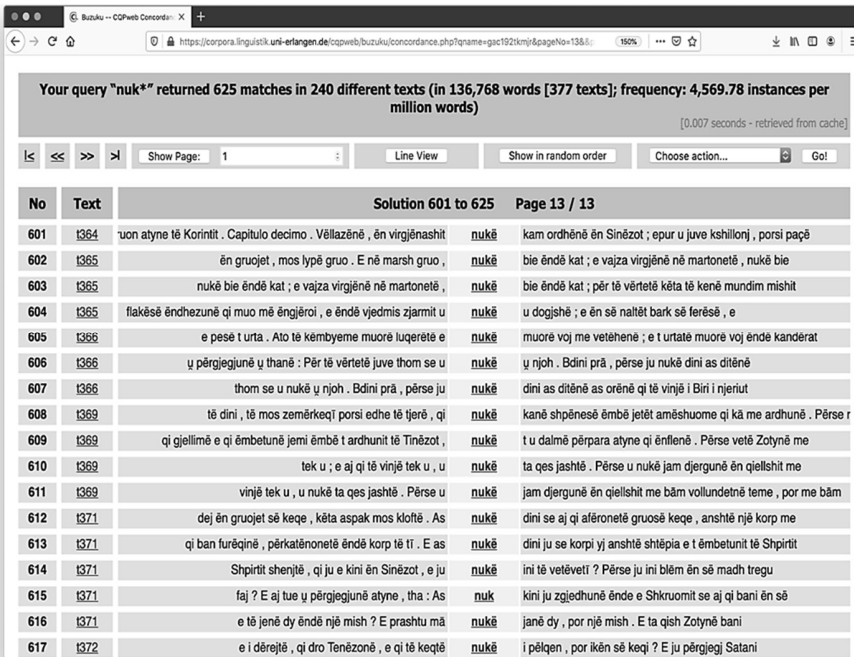
In case of the Buzuku Corpus we began the annotation of word forms on closed classes: conjunctions, prepositions, particles, interjections, as they are simpler to annotate than the other parts of speech. We have begun annotating verbs based on the material collected by W. Fiedler (2004). Since the DFG-Project *Digitales philologisch-etymologisches Wörterbuch des Altalbanischen (15.–18. Jh.)* at Ludwig-Maximilians University of Munich 2018 has started, we are looking to benefit from some of the linguistic data collected, analyzed and prepared there, primary the parts of speech, but also the other information e. g. morphological properties of word forms.

A morphological analyzer like Kabashi (2015) can be adapted to annotate the texts, whereby it is helpful for long texts like e. g. from Buzuku and from Bogdani. Word forms from short texts can be also integrated into the system,

i. e. the word forms can be listed as paradigms and implemented as a morphological analyzer. In contrast to partial analyzers, a system covering all possible forms has the advantage that every possible word form that is a syncretism can be disambiguated respectively used to compare word forms with regards to which text, they occur in, and by which author they are used. A statistical tagger can be used to preprocess the data so that they only need to be verified manually. This step is necessary to have a high qualitative annotated data. The manual testing is possible because the diachronic data are limited in number. Also, a cross validation can help after the first step to test the statistically annotated data.

4. CQPweb

The CQPweb (Corpus Query Processor) is a software package, a web-based corpus analysis system, to explore corpus data, cf. Hardie (2012). CQPweb is being widely used by numerous researchers and universities for corpus exploration. The software is not proprietary so that there is a perspective for maintenance and continuous development, extending features and functions. The meta-data can be used in a very flexible way. The functions are very simple and intuitive. CQPweb offers various functions, e. g. for calculating collocations.



Your query "nuk*" returned 625 matches in 240 different texts (in 136,768 words [377 texts]; frequency: 4,569.78 instances per million words) [0.007 seconds - retrieved from cache]

Navigation: < << >> > Show Page: 1 Line View Show in random order Choose action... Go!

| No | Text | Solution 601 to 625 | Page 13 / 13 |
|-----|--|---------------------|---|
| 601 | 1364 uon atyne të Korintit . Capitulo decimo . Vëllazënë , ën virgjënashit | nukë | kam ordhënë ën Sinëzot ; epur u juve kshillonj , porsì paqë |
| 602 | 1365 ën gruojet , mos lypë gruo . E në marsh gruo , | nukë | bie ëndë kat ; e vajza virgjënë në martonetë , nukë bie |
| 603 | 1365 nukë bie ëndë kat ; e vajza virgjënë në martonetë , | nukë | bie ëndë kat ; për të vërtetë këta të kenë mundim mishit |
| 604 | 1365 flakësë ëndhezunë qì muo më ëngjëroi , e ëndë vjedmis zjamit u | nukë | u dogjëshë ; e ën së naltët bark së ferësë , e |
| 605 | 1366 e pesë t urta . Ato të këmbyme muorë luqeretë e | nukë | muorë voj me vetëhenë ; e t urtatë muorë voj ëndë kanderat |
| 606 | 1366 u përgjegjunë u thanë : Për të vërtetë juve thom se u | nukë | u njoh . Bdinì prà , përse ju nukë dinì as ditënë |
| 607 | 1366 thom se u nukë u njoh . Bdinì prà , përse ju | nukë | dinì as ditënë as orënë qì të vinjë i Biri i njeriut |
| 608 | 1369 të dinì , të mos zemërkeqì porsì edhe të tjerë , qì | nukë | kanë shpënesë ëmbë jetët amëshuome qì kà me ardhunë . Përse r |
| 609 | 1369 qì gjellimë e qì ëmbetunë jemi ëmbë t ardhunit të Tinëzot , | nukë | t u dalmë përpara atyne qì ënflënë . Përse vetë Zotynë me |
| 610 | 1369 tek u ; e aj qì të vinjë tek u , u | nukë | ta qes jashtë . Përse u nukë jam djergunë ën qiellshit me |
| 611 | 1369 vinjë tek u , u nukë ta qes jashtë . Përse u | nukë | jam djergunë ën qiellshit me bàm vollundetë teme , por me bàm |
| 612 | 1371 de jë gruojet së keqe , këta aspak mos kloftë . As | nukë | dinì se aj qì afëronetë gruosë keqe , anshë një korp me |
| 613 | 1371 qì ban furëqinë , përkatënonetë ëndë korp të tì . E as | nukë | dinì ju se korpi yj anshë shtëpia e t ëmbetunit të Shpiritit |
| 614 | 1371 Shpiritit shenjtë , qì ju e kinì ën Sinëzot , e ju | nukë | ini të vetëveti ? Përse ju ini blëm ën së madh tregu |
| 615 | 1371 faj ? E aj tue u përgjegjunë atyne , tha : As | nuk | kinì ju zgjedhunë ënde e Shkruomit se aj qì bani ën së |
| 616 | 1371 e të jenë dy ëndë një mish ? E prashu mà | nukë | janë dy , por një mish . E ta qish Zotynë bani |
| 617 | 1372 e i dërejtë , qì dro Tenëzonë , e qì të keqë | nukë | i pëlqen , por ikën së keqi ? E ju përgjegji Satani |

The above figure shows the result of simple searching with regular expressions for word forms starting with nuk. The star stands for any letter that can occur one or more times or not at all.

5. Conclusions

The presented work is only an initial step for preparing diachronic corpora of the Albanian language, which can be useful for medieval linguists, but not only for them. As the work with raw texts is going as planned, confirmed by the work with Buzuku data, it is promising to continue the work with the remaining data. For the work with linguistic annotation – of course – a mediaevalist scholar (team) is needed.

6. References

- Fiedler, Wilfried (2004): Das Albanische Verbalsystem in der Sprache des Gjon Buzuku (1555). Botime te Vecanta LV., Libri 25. Prishtinë: Akademia e Shkencave dhe e Arteve e Kosovës.
- Gipert, Jost et al.: (1995–) *The Thesaurus Indogermanischer Text- und Sprachmaterialien (TITUS)*, Text Database, online <http://titus.fkidg1.uni-frankfurt.de>, Universität Frankfurt. URL: <http://titus.uni-frankfurt.de> (last access December 12, 2019).
- 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.
- Morozova, Maria & Rusakov, Alexander (2020) “The early Albanian texts in an annotated language corpus: An attempt of processing and analysis.” (In this volume, pp. 91-102).