Mireille Ducassé/Archil Elizbarashvili

FINDING LEMMAS IN AGGLUTINATIVE AND INFLECTIONAL LANGUAGE DICTIONARIES WITH LOGICAL INFORMATION SYSTEMS

The case of Georgian verbs

Abstract Looking up for an unknown word is the most frequent use of a dictionary. For languages both agglutinative and inflectional, such as Georgian, this can be quite challenging because an inflected form can be very far from the lemmas used by the target dictionary. In addition, there is no consensus among Georgian lexicographers on which lemmas represent a verb in dictionaries. It further complicates dictionaries access. Kartu-Verbs is a base of inflected forms of Georgian verbs accessible by a logical information system. It currently contains more than 5 million inflected forms related to more than 16,000 verbs for 11 tenses; each form can have 11 properties; there are more than 80 million links in the base. This demonstration shows how, from any inflected form, we can find the relevant lemma to access any dictionary. Kartu-Verbs can thus be used as a front-end to any Georgian dictionary.

Keywords E-dictionary; lemma; Georgian language; under-resourced language; inflected forms; logical information systems; semantic web

1. Introduction

The survey reported in Kosem et al. (2019), involving 10,000 people of 26 countries, shows that looking up for an unknown word is the most frequent use of a monolingual dictionary. For languages both agglutinative and inflectional, such as Georgian, this can be quite challenging because an inflected form can be very far from the lemmas used by the target dictionary. As discussed in Ducassé (2020), conjugation can modify any part of the form of Georgian verbs and to go from a conjugated form to a lemma requires a good knowledge of Georgian grammar. Moreover, the lemmatization of verbs in Georgian dictionaries is still an open problem, see Margalitadze (2020) and Gippert (2016). Most dictionaries use “Georgian infinitives”[1] as entries. The “Comprehensive Georgian-English Dictionary” (Rayfield et al. 2006) presents, for all verbs, the infinitive as well as the 3rd person singular in the present and the future. The Georgian-German dictionary of Tschenkéli et al. (1965) uses the abstract verb root under which all sub-paradigms are listed. Thus, depending on the target dictionary, starting from inflected form “vikiraveb” (ვიქირავებ, «I will rent»), a user would have to find Georgian infinitive “kiraoba” (ქირაობა), 3rd person present “kiraobs” (ქირაობს), 3rd person future “ikiravebs” (იქირავებს), or root “kirav” (ქირავ). Note that this verb is a relatively “easy” one.

Kartu-Verbs is a knowledge base of inflected forms of Georgian verbs[2] integrated within a semantic web tool, Sparklis, a platform to easily implement logical information systems.

[1] What we call “Georgian infinitive” is, strictly speaking, a verbal noun. This is the form that comes closest to an infinitive. We use this term because it is easier to understand for non-linguist English-speaking target users.

[2] Kartu-Verbs can be accessed at https://www-semlis.irisa.fr/software/georgian-verb-inflected-forms-base/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Logical information systems allow users to retrieve information on the facets of their choice by progressively refining their queries using suggestions (Ferré 2017). The base can be easily navigated in all directions: from an inflected form to an infinitive, and conversely from an infinitive to any inflected form; from components to forms and from a form to its components. Thanks to our collaboration with Paul Meurer, we are in the process of integrating the very rich knowledge about Georgian verbs underneath the Georgian part of his INESS::XLE-Web platform (Meurer 2007).[3] The integration is not yet completed. Kartu-Verbs contains, however, already more than 5 million inflected forms related to more than 16.000 verbs for 11 tenses. Each form can have 11 properties. There are more than 80 million links in the base. All keywords can be displayed both in Georgian and Latin characters. This version of Kartu-Verbs is almost 3 orders of magnitude larger than the previous version that only contained thousands of inflected forms related to hundreds of verbs. Ducassé (2020) mainly illustrates how to use Kartu-Verbs to gain knowledge about Georgian conjugation. This demonstration shows in details how, from any inflected form, one can find the relevant lemma(s) to access any dictionary. Kartu-Verbs can thus be used to build a front-end to any Georgian dictionaries.

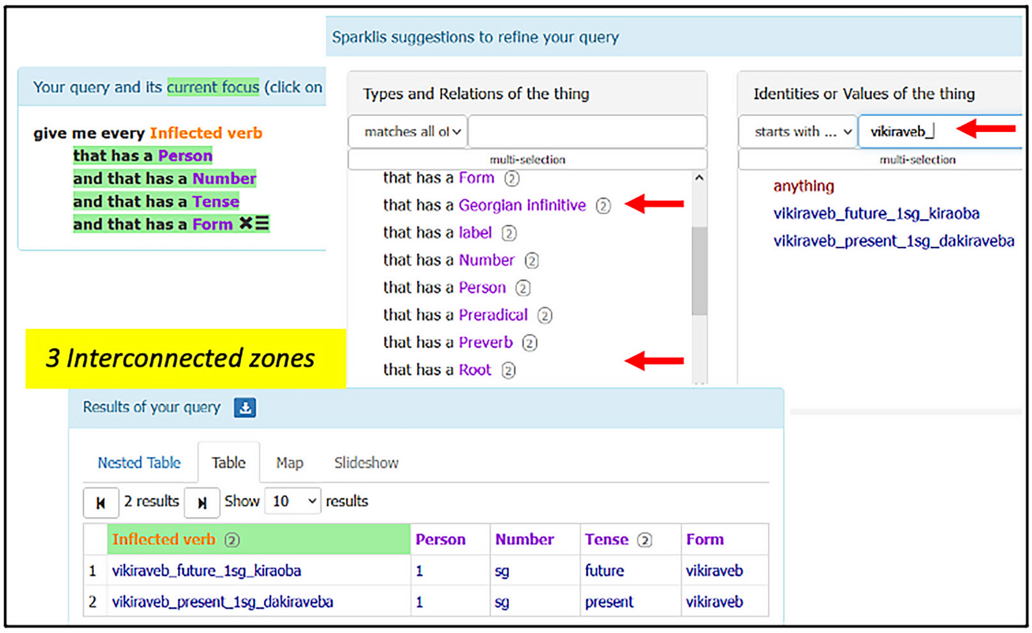
2. Finding relevant entries from inflected forms

This section explains how to use Kartu-Verbs to find in a few clicks 4 different lemmas for 4 different dictionary organizations, starting from an inflected form. In addition, we show how to get verb variants from a given root. We use a Latin transliteration following Georgian government recommendations.[4] The base can also be searched using the Georgian alphabet.

Figure 1 shows the 3 areas of the Kartu-Verbs user interface, from left to right and top to bottom: the query, the suggestions and the results. A basic query is displayed to find 4 of the features of inflected forms: person, number tense and (surface) form. The “Suggestions” area, is itself divided into sub-areas, only 2 are of interest to us for this demonstration. On the left, the “Types and Relationships” sub-area offers features that can still be added to the query; on the right, the “Identities or values” box suggests some of the inflected form identifiers that match the query. Suppose the user is looking for lemmas for “vikiraveb” (“ვიქირავებ”). We can see in the figure that, by entering “vikiraveb_” in the search part of the “Identities or Values of the thing” sub-area, two forms are accessible for two different tenses, their identifiers are “vikiraveb_future_1sg_kiraoba” and “vikiraveb_present_1sg_dakiraveba”. Note that the results area has already adjusted to the suggested values. Sometimes this is enough to find the information we are looking for. As discussed above, depending on the targeted dictionaries, the Georgian infinitive, the 3rd-person singular present or future, as well as the root, would be needed. In the relations sub-area of suggestions, on the left, we see that the “Georgian infinitive” and the “root” features are accessible.

[3] INESS::XLE-Web is a powerful tool dedicated to linguists. It is able to parse sentences and produce syntactic trees for a number of languages, including Georgian. It contains a lot of interesting information but presented in a form not really accessible to beginners.

[4] http://www.enadep.gov.ge/uploads/Bulletin_II_2019-2020.pdf.


Sparklis suggestions to refine your query
Your query and its current focus (click on
3 Interconnected zones
Types and Relations of the thing
matches all of
multi-selection
Identities or Values of the thing
starts with ...
multi-selection
Results of your query
Nested Table
Table
Map
Slideshow
Show 10 results

give me every Inflected verb
that has a Person
and that has a Number
and that has a Tense
and that has a Form ✖☰

that has a Form ②
that has a Georgian Infinitive ②
that has a label ②
that has a Number ②
that has a Person ②
that has a Preradical ②
that has a Preverb ②
that has a Root ②

vikiraveb_

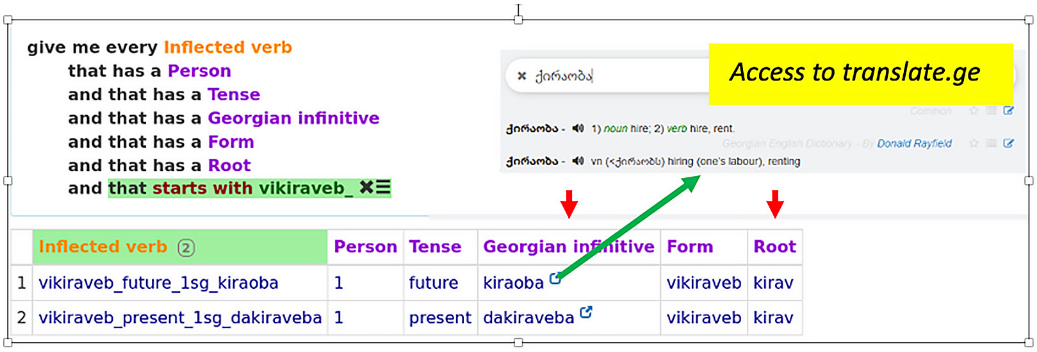
anything
vikiraveb_future_1sg_kiraoba
vikiraveb_present_1sg_dakiraveba

2 results

	Inflected verb ②	Person	Number	Tense ②	Form
1	vikiraveb_future_1sg_kiraoba	1	sg	future	vikiraveb
2	vikiraveb_present_1sg_dakiraveba	1	sg	present	vikiraveb

Fig. 1: Searching for inflected form “vikiraveb”

Figure 2 shows the query and result areas after the user clicked on the 3 suggestions. The query has been automatically updated. We see that the inflected form is “vikiraveb”. The “Georgian infinitive” and “root” features have been added. In the results area, two columns have appeared giving the two pieces of information sought. The Georgian infinitive is “kiraoba” (ქირაობა) for the present form and “dakiraveba” for the future form. The root is “kirav” for both forms. We can also see that both forms correspond to the first-person singular. Clicking on the arrow at the right of each of the Georgian Infinitive opens a new window to access the translate.ge site. We can see that “kiraoba” can mean “to hire” or “to rent”.



give me every Inflected verb
that has a Person
and that has a Tense
and that has a Georgian infinitive
and that has a Form
and that has a Root
and that starts with vikiraveb_ ✖☰

ქირაობა

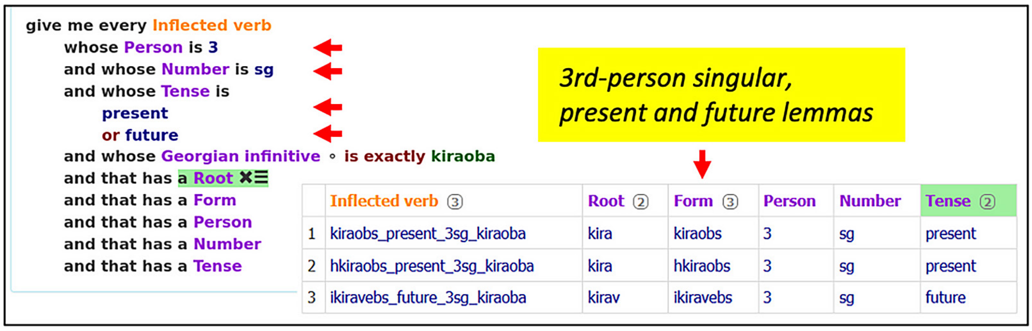
ქირაობა - 1) noun hire; 2) verb hire, rent.

ქირაობა - vn (<ქირაობს) hiring (one's labour), renting

	Inflected verb ②	Person	Tense	Georgian infinitive	Form	Root
1	vikiraveb_future_1sg_kiraoba	1	future	kiraoba	vikiraveb	kirav
2	vikiraveb_present_1sg_dakiraveba	1	present	dakiraveba	vikiraveb	kirav

Fig. 2: Two lemmas for inflected form “vikiraveb” (Georgian infinitive and root)

We have thus found in a few clicks two of the four lemmas. Figure 3 displays a refined query to find the two other lemmas for verb “kiraoba”, its 3rd-person singular forms present and future. From the previous query, we have kept the Georgian infinitive and the number by clicking on “kiraoba” and “singular” in the table of results. We specified (by clicking in the suggestions) that we wanted third-person forms. We took advantage of the logical capabilities of Sparklis queries to ask for tense to be either present or future (using the ≡ menu icone next to query elements). We have also asked to see 5 features in the result table: Root, Form, Person, Number and Tense. The results area displays 3 form identifiers because the verb has two different forms at present. Note that if we had not specified the tense, we would have gotten a conjugation table with the third person singular forms of the verb "kiraoba" at all tenses. We can see on Figure 3 that the same verb can have different roots at different tenses ("kira" at present or "kirav" at future.) We could investigate further which root is used for which tenses by clicking on one root not specifying the tense.



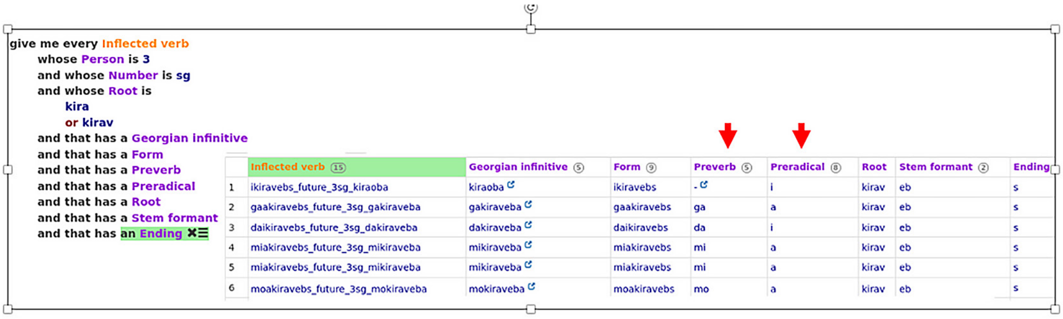
```
give me every Inflected verb
    whose Person is 3
    and whose Number is sg
    and whose Tense is
        present
        or future
    and whose Georgian infinitive ∘ is exactly kiraoba
    and that has a Root ✖☰
    and that has a Form
    and that has a Person
    and that has a Number
    and that has a Tense
```

3rd-person singular, present and future lemmas

	Inflected verb ③	Root ②	Form ③	Person	Number	Tense ②
1	kiraobs_present_3sg_kiraoba	kira	kiraobs	3	sg	present
2	hkiraobs_present_3sg_kiraoba	kira	hkiraobs	3	sg	present
3	ikiravebs_future_3sg_kiraoba	kirav	ikiravebs	3	sg	future

Fig. 3: Two more types of lemmas for vikiraveb

In Georgian, many verb variants can be constructed with a given root. Figure 4 shows a query to find all the 3rd person singular future of verbs whose root is either "kira" or "kirav". There are 15 different forms with different preverbs and preradicals. This shows the importance of morphological information for agglutinative languages. We could click on any of the morphemes to refine the query.



```
give me every Inflected verb
    whose Person is 3
    and whose Number is sg
    and whose Root is
        kira
        or kirav
    and that has a Georgian infinitive
    and that has a Form
    and that has a Preverb
    and that has a Preradical
    and that has a Root
    and that has a Stem formant
    and that has an Ending ✖☰
```

	Inflected verb ⑮	Georgian infinitive ⑤	Form ⑨	Preverb ⑤	Preradical ⑧	Root	Stem formant ②	Ending
1	ikiravebs_future_3sg_kiraoba	kiraoba	ikiravebs	-	i	kirav	eb	s
2	gaakiravebs_future_3sg_gakiraveba	gakiraveba	gaakiravebs	ga	a	kirav	eb	s
3	daikiravebs_future_3sg_dakiraveba	dakiraveba	daikiravebs	da	i	kirav	eb	s
4	miakiravebs_future_3sg_mikiraveba	mikiraveba	miakiravebs	mi	a	kirav	eb	s
5	miakiravebs_future_3sg_mikiraveba	mikiraveba	miakiravebs	mi	a	kirav	eb	s
6	moakiravebs_future_3sg_mokiraveba	mokiraveba	moakiravebs	mo	a	kirav	eb	s

Fig. 4: Finding all the verbs with given roots

3. Discussion

In this demonstration, even for a relatively simple verb we have encountered a number of "homonyms", namely entries with identical surface forms with different attributes (for example, "vikiraveb" can be either at present or future). Figure 1 illustrates how the form identifiers help to distinguish homonyms (for example "vikiraveb_future_1sg_kiraoba" vs "vikiraveb_present_1sg_dakiraveba"). There can also be "synonyms", namely entries sharing their properties with different surface forms (for example "kiraobs" and "hkiraobs" are two different surface forms for verb “kiraoba” at 3rd singular present). In Figures 3 and 4 we can see that synonyms can easily be captured thanks to the query mechanism that can traverse the base in all directions and easily go from components to entries.

The queries can use any combination of logical operators “AND”, “OR”, “NOT” at any place as the user can specify the focus (scope) of an operator. There are also aggregate operators, not demonstrated here. We have illustrated how to add features in the query. The user can also easily remove features that are not of interest to him at a given moment by clicking on the “x” attached to them as can be seen next to “tense” in Figure 3. As illustrated above, all queries are constructed using suggestions. Users have nothing to invent. They can use filters to help Kartu-Verbs come up with relevant suggestions, then queries are built only by clicking on suggestions that are necessarily relevant. The benefits are three-fold, first, it is easier to find something in a list than typing it, second, users cannot make typos, and finally, as a direct result, queries can never yield an empty result. This is a very strong property due to the powerful mechanisms of Sparklis.

Some linguists provide comprehensive tables of inflected forms, for example the “Georgische Verbtabellen” by Chotiwari-Jünger et al. (2010) or the books of the “Biliki” series by Nana Shavtvaladze[5]. The latter offer conjugation tables of several types as an appendix to the lessons. They contain invaluable information. However, learners have to go through different books to find relevant information. When searching for an inflected form, learners must check each of the over 10,000 entries. Furthermore, the inflected forms use the Georgian alphabet, which is a big hurdle for beginners. Exceptions, quite common, cannot always be anticipated from the sample tables. Unlike tables, in Kartu-Verbs there are no predefined usages. Any field can be used to search the database at any moment and the whole information is accessible from the query.

Other systems provide all inflected forms for inflectional and agglutinating languages. For example, in Verbel, the inflection dictionary for Polish, morphological structure and inflectional properties are used to structure verb forms into “flexemes” (Czerepowicka 2021). We still have to investigate if this notion of flexeme could help us integrate more data with reasonable performances. In DeCesaris (2021), the author advocates to add more morphological information in monolingual dictionaries, in particular because users may not have enough knowledge about inflectional rules and derived words may have acquired additional nuances of meaning. Morphological information is also considered crucial by de Schryver/Chishman/DaSilva (2019) for languages with complex morphology such as Bantu. In Kartu-Verbs, not only do we provide morphological information, but each morpheme, like any piece of information, is an URI that can be used to lead to another piece of information.

A dedicated front-end to dictionaries is under development so that users do not have to see the queries for the usual searches, while being able to access sophisticated queries when needed. We also plan to integrate more features, in particular from the INESS::XLE-Web platform (Meurer 2007). It raises issues both in terms of performances and data structures.

4. Conclusion

In this article, we have illustrated how versatile and powerful LIS navigation mechanisms are. We have also shown how these mechanisms can help users of Kartu-Verbs easily obtain

[5] Biliki, Georgian Language For English Speakers. See http://lsgeorgia.com.

information about the verbs they encounter in Georgian texts whatever their form. In particular, finding the relevant lemma(s) for an entry in a given dictionary is no longer a problem. Kartu-Verbs can thus be used as an interface for any Georgian dictionary, regardless of the lemmatization principles the dictionary uses for verbs.

References

Chotiwari-Jünger, S. et al. (2010): Georgische Verbtabellen. Hamburg.

Czerepowicka, M. (2021): The structure of a dictionary entry and grammatical properties of multiword units. In: Electronic lexicography in the 21st century: eLex 2021, pp. 200–215.

de Schryver, G.-M./Chishman, R./DaSilva, B. (2019): An overview of digital lexicography and directions for its future: An interview with Gilles-Maurice de Schryver. Universidad do Vale de Rio dos Sinos.

Ducassé, M. (2020): Kartu-verbs: A semantic web base of inflected Georgian verb forms to bypass Georgian verb lemmatization issues. In: Gavriilidou, Z./Mitsiaki, M./Fliatouras, A. (eds.): Proceedings of XIX EURALEX International Congress, Volume 1, pp. 81–89.

Ferré, S. (2017): Sparklis: An expressive query builder for SPARQL end-points with guidance in natural language. In: Semantic Web: Interoperability, Usability, Applicability 8 (3).

Gippert, J. (2016): Complex morphology and its impact on lexicology: the Kartvelian case. In: Margalitadze, T./Meladze, G. (eds.): Proceedings of the 17th EURALEX International Congress, Tbilisi, Georgia, pp. 16–36. Ivane Javakhishvili Tbilisi University Press.

Kosem, I./Lew, R./Müller-Spitzer, C./Ribeiro Silveira, M./Wolfer, S./Dorn, A./Gurrutxaga, A./Ceberio, K./Etxeberria, E./Lefer, M.-A. et al. (2019): The image of the monolingual dictionary across Europe. In: International Journal of Lexicography 32 (1), pp. 92–114.

Margalitadze, T. (2020): Lexicography of Georgian: a brief overview. Language@Leeds Working Papers in Linguistics, University of Leeds.

Meurer, P. (2007): A computational grammar for Georgian. In: Logic, language, and computation. 6th International Tbilisi Symposium on Logic, Language, and Computation. Berlin, pp. 1–15.

Rayfield, D./Apridonze, S./Chanturia, A./Amirejibi, R./Broers, L./Chkhaidze, L./Margalitadze, T. (2006): A Comprehensive Georgian-English Dictionary. London.

Tschenkéli, K./Marchev, Y./Flury, L. (1965): Georgisch-deutsches Wörterbuch, Volume 2. Zürich.

Contact information

Mireille Ducassé
IRISA-INSA Rennes
mireille.ducasse@irisa.fr

Archil Elizbarashvili
Ivane Javakhishvili Tbilisi State University
archil.elizbarashvili@tsu.ge

Acknowledgements

We are indebted to Paul Meurer who let us transfer the invaluable data from his private Clarino base. We thank Mikheil Sulikashvili for his technical support to transfer Clarino information into Csv files. We warmly thank Sébastien Ferré for his strong support to use Sparklis.