Kartu-Verbs : un système d'informations logiques de formes verbales fléchies pour contourner les problèmes de lemmatisation des verbes géorgiens

Mireille Ducassé

Démonstration présentée à la conférence francophone sur l'Extraction et la Gestion des Connaissances, janvier 2022.

Kartu-Verbs : un système d'informations logiques de formes verbales fléchies pour contourner les problèmes de lemmatisation des verbes géorgiens

Mireille Ducassé*

*IRISA-INSA Rennes
mireille.ducasse@irisa.fr,

Résumé. La langue géorgienne possède un système verbal complexe, à la fois agglutinant et flexionnel, avec de nombreuses irrégularités. Les formes fléchies d'un verbe peuvent être très différentes les unes des autres. Il faut une bonne connaissance de la grammaire géorgienne pour remonter à l'infinitif (le lemme d'accès des dictionnaires le plus fréquent). L'accès aux dictionnaires pour les débutants est, de ce fait, très difficile. De plus, il n'y a pas de consensus parmi les lexicographes du Géorgien sur les lemmes qui représentent un verbe dans les dictionnaires, ce qui complexifie encore davantage les accès. Nous proposons Kartu-Verbs, une base de formes fléchies de verbes géorgiens accessible par un système d'informations logiques. Cette démonstration [1] montre comment, à partir de n'importe quelle forme fléchie, on peut trouver le lemme pertinent pour accéder à n'importe quel dictionnaire. Kartu-Verbs peut, ainsi, être utilisé comme une interface aux dictionnaires géorgiens.

1 Introduction

Le Géorgien est une langue caucasienne, langue maternelle d'environ 5 millions de personnes. Il a son propre alphabet. La grammaire géorgienne a un système verbal complexe. Certains problèmes sont illustrés ci-dessous du point de vue d'un débutant (pour plus de détails, voir par exemple Anderson (1984) ; Tuite (1998) ; Assatiani et Malherbe (2011) ; Gérardin (2016)). Il existe de nombreux verbes irréguliers et la langue est à la fois agglutinante et flexionnelle. La conjugaison peut modifier toute partie de la forme des verbes. Par exemple, le verbe « travailler » (mushaoba

1. Cette démonstration complète et reprend des élements de l'article suivant. Ducassé,M.(2020).Kartu-verbs : A semantic web base of inflected Georgian verb forms to bypass Georgian verb lemmatization issues. In Z. Gavriilidou, M. Mitsiaki, et A. Fliatouras (Eds.), Proceedings of XIX EURALEX Congress

- მუშაობა), à la première personne du pluriel du présent donne « vmushaobt » (ვმუშაობთ). Notez le « v » au début du verbe pour marquer la première personne et le « t » à la fin pour marquer le pluriel. La première personne du singulier du futur donne « vimushaveb » (ვიმუშავებ). Un « i » a été inséré après le marqueur « v » de la première personne ; cela est assez fréquent pour un grand ensemble de verbes. Notez que « ob » est devenu « eb » ; ce qui est typique d'un sous-ensemble plus petit de verbes. L'apparition du « v » après la racine est plus exceptionnelle. Passer d'une forme conjuguée à un infinitif géorgien [2] demande une bonne connaissance de la grammaire géorgienne. Pour les néophytes, cela représente un véritable défi. Par exemple, pour « chamodikhar » (ჩამოდიხარ, « tu viens en descendant »), l'infinitif géorgien est « mosvla » (მოსვლა, « venir »).

De plus, la lemmatisation des verbes dans les dictionnaires géorgiens est toujours un problème ouvert comme discuté dans Margalitadze (2020) et Gippert (2016). De nombreux projets donnent, en plus de l'infinitif, des exemples de formes fléchies comme lemme(s). Par exemple, la troisième personne du futur du singulier est utilisée dans Daraselia et Sharoff (2016). Le « Comprehensive Georgian-English Dictionary » présente, pour tous les verbes, l'infinitif et la 3e personne du singulier au présent et au futur. Il est cependant encore difficile pour un néophyte de trouver le « chamodikhar » mentionné ci-dessus. Le dictionnaire Géorgien-Allemand de Tschenkéli et al. (1965) utilise la racine verbale abstraite sous laquelle tous les sous-paradigmes sont répertoriés. Si cette représentation est très instructive pour les linguistes, elle est trop lourde pour les débutants, d'autant plus que de nombreuses racines ne sont constituées que d'un ou deux caractères.

Nous proposons, Kartu-Verbs, une base de formes fléchies de verbes géorgiens, contenant actuellement plus de 15 000 formes fléchies liées à 278 verbes pour 10 temps [3]. Les formes fléchies sont intégrées au sein d'un outil du web sémantique, Sparklis, une plate-forme qui permet d'implémenter facilement des systèmes d'informations logiques. Les systèmes d'informations logiques permettent aux utilisateurs de récupérer des informations sur les facettes de leur choix en affinant progressivement leurs requêtes grâce à des suggestions (cf Ferré (2017)). La base peut être facilement parcourue dans toutes les directions : du Géorgien au Français et à l'Anglais ; d'une forme fléchie à un infinitif, et inversement d'un infinitif à toute forme fléchie ; des composants aux formes et d'une forme à ses composants.

Cette démonstration montre comment, à partir de n'importe quelle forme fléchie, on peut trouver le lemme pertinent pour accéder à n'importe

2. Ce que nous appelons « infinitif géorgien » est à proprement parler un nom verbal. C'est la forme qui se rapproche le plus d'un infinitif. Nous utilisons ce terme car il est plus facile à comprendre pour les utilisateurs cibles, français ou anglais non linguistes.

3. Kartu-Verbs est accessible à `https://www-semlis.irisa.fr/software/georgian-verb-inflected-forms-base/`.

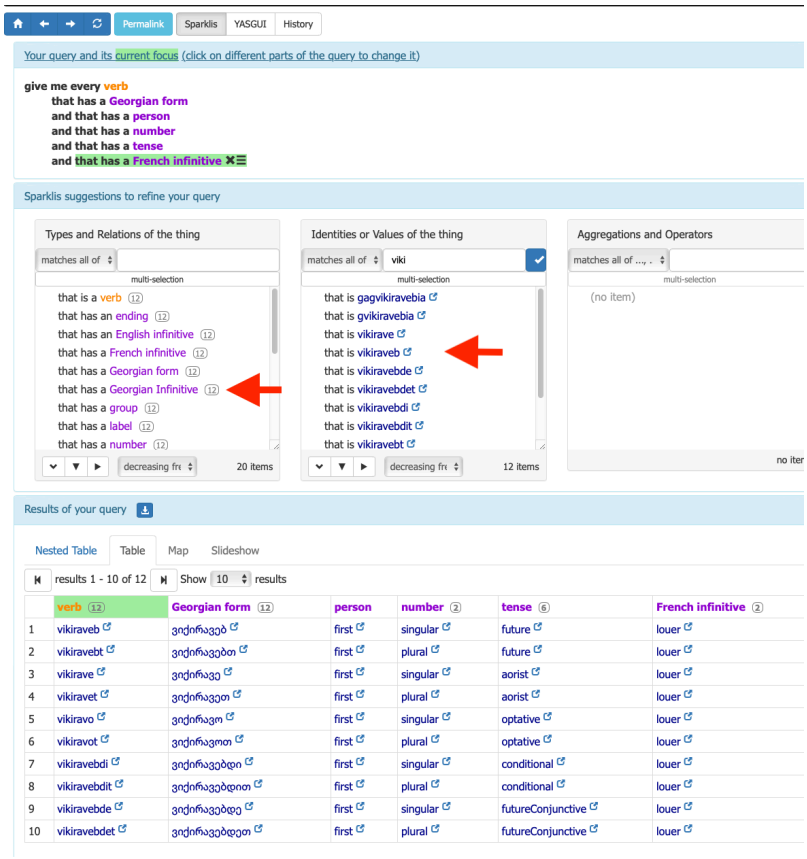

Fig. 1 – *Recherche d'une forme fléchie.*

quel dictionnaire. Kartu-Verbs peut, ainsi, être utilisé comme interface aux dictionnaires géorgiens.

2 Trouver des entrées pertinentes à partir de formes fléchies

Kartu-Verbs a été créé afin de permettre à tous les usagers d'accéder à leur dictionnaire quelles que soient leurs connaissances linguistiques courantes. Cette section illustre comment utiliser Kartu-Verbs pour trouver, à partir d'une forme fléchie, 4 lemmes différents pour 4 organisations différentes de dictionnaire et démontre la puissance de l'outil.

La figure 1 montre les 3 zones de l'interface utilisateur de Kartu-Verbs, de haut en bas : la requête, les suggestions et les résultats. Une requête

Fig. 2 – *Deux lemmes à partir d'une forme fléchie.*

de base est affichée ici. Elle permet de retrouver 5 des traits des formes fléchies : forme en alphabet géorgien, personne, nombre, temps et infinitif français. La zone « Suggestions », au centre, est elle-même divisée en trois sous-zones, seules les 2 de gauche nous intéressent pour cette démonstration. À gauche, la sous-zone « Types et relations » propose des traits qui peuvent encore être ajoutés à la requête ; au centre, la zone « Identités ou valeurs » suggère certaines des formes verbales qui correspondent à la requête. Supposons que l'utilisateur soit à la recherche de lemmes pour la forme « vikiraveb » (« ვიქირავებ »). On voit sur la figure, qu'en rentrant « viki » dans la partie de recherche de la sous-zone valeurs des suggestions, « vikiraveb » est accessible, il suffit de cliquer dessus pour l'intégrer dans la requête. On peut noter que la zone résultats s'est déjà ajustée aux valeurs suggérées. Il arrive que cela suffise pour trouver les informations que l'on cherche. Comme vu plus haut, en fonction des dictionnaires ciblés, il faudrait l'infinitif géorgien, la 3ème personne du singulier au présent ou au futur, ainsi que la racine. Dans la sous-zone relations des suggestions, à gauche, on voit que le trait « Georgian infinitive » est accessible. De même en descendant dans cette liste, on pourrait voir que le trait « root » est également accessible.

La figure 2 affiche les zones de requête et de résultats après que l'utilisateur ait cliqué sur les 3 suggestions. La requête a été automatiquement mise à jour. On voit que la forme fléchie est « vikiraveb ». Les traits « Georgian infinitive » et « root » ont été ajoutés. Dans la zone de résultats, deux colonnes ont été ajoutées donnant les deux informations recherchées, l'infinitif Géorgien est donc « kiraoba » et la racine « kira ». On peut voir, de plus, la forme en alphabet géorgien, « ვიქირავებ », et que la forme correspond à la première personne du singulier au futur du verbe louer.

	verb (10)	Georgian form (10)	tense (10)
1	kiraobs	ქირაობს	present
2	kiraobda	ქირაობდა	imperfect
3	ikiravebs	იქირავებს	future
4	ikirava	იქირავა	aorist
5	ikiravos	იქირავოს	optative
6	ikiravebda	იქირავებდა	conditional
7	ikiravebdes	იქირავებდეს	futureConjunctive
8	kiraobdes	ქირაობდეს	presentConjunctive
9	ekirava	ექირავა	pastPerfect
10	ukiravebia	უქირავებია	presentPerfect

Fig. 3 – *Troisième lemme et table de conjugaison à partir d'une forme fléchie.*

On a donc trouvé en quelques clics deux des quatre lemmes cherchés. Pour trouver le troisième, la forme à la troisième personne du singulier au présent, on va garder l'infinitif géorgien et le nombre en cliquant sur « kiraoba » et « singular » dans le tableau de résultats. La figure 3 affiche la requête affinée, où on a, de plus, précisé (en cliquant dans les suggestions) qu'on voulait des formes à la troisième personne. La zone de résultats affiche les formes de la troisième personne du singulier du verbe « louer » à tous les temps. En plus du lemme recherché, on a donc une table de conjugaison. Pour trouver le quatrième lemme, la forme à la troisième personne du singulier au futur, on procède comme précédemment, en changeant le temps.

Commentaires. On a illustré comment ajouter des traits dans la requête. L'utilisateur peut aussi facilement enlever des traits qui ne l'intéressent pas à un instant donné en cliquant sur les "x" attachés aux traits comme

on peut voir à côté de « tense » sur la figure 3.

Les requêtes données ici n'utilisent que l'opérateur logique « ET ». Les opérateurs logiques « OU » et « NON » sont également disponibles pour des requêtes plus sophistiquées, de même que des opérateurs d'agrégat. D'autres possibilités illustrant ces opérateurs sont détaillées dans Ducassé (2020) : comment obtenir des informations à partir d'un infinitif anglais ; comment construire un échantillon de conjugaison ; comment obtenir des informations de conjugaison à partir d'une terminaison donnée ; comment acquérir des connaissances en comparant des formes ou des racines similaires ; comment vérifier des hypothèses sur des constructions à l'aide d'opérateurs logiques ainsi que des requêtes plus sophistiquées pour acquérir des méta-connaissances sur la base en utilisant des agrégats.

Certains linguistes fournissent des tableaux exhaustifs de formes fléchies, par exemple les « Georgische Verbtabellen » de Chotiwari-Jünger et al. (2010) ou les livres de la série « Biliki » de Nana Shavtvaladze[4]. Ces derniers proposent des tableaux de conjugaison de plusieurs types en annexe des leçons. ls contiennent des informations inestimables. Cependant, les apprenants doivent parcourir différents livres pour trouver des informations pertinentes. Lors de la recherche d'une forme fléchie, les apprenants doivent vérifier chacune des plus de 10 000 entrées. De plus, les formes fléchies utilisent l'alphabet géorgien, ce qui est un gros obstacle pour les débutants. Les exceptions, assez courantes, ne peuvent pas toujours être anticipées à partir des exemples de tableaux. Contrairement aux tables, dans Kartu-Verbs il n'y a pas d'utilisations prédéfinies. N'importe quel champ peut être utilisé pour chercher dans la base et toutes les informations sont accessibles depuis la requête.

Comme illustré ci-dessus, toutes les requêtes sont construites à l'aide de suggestions. Les utilisateurs n'ont rien à inventer. Ils peuvent utiliser des filtres pour aider Kartu-Verbs à proposer des suggestions pertinentes, puis des requêtes sont construites uniquement en cliquant sur des suggestions qui sont nécessairement pertinentes. Les avantages sont triples, premièrement, il est plus facile de trouver quelque chose dans une liste que de le taper, deuxièmement, les utilisateurs ne peuvent pas faire de fautes de frappe, et enfin, en conséquence directe, les requêtes ne peuvent jamais donner un résultat vide. C'est une propriété très forte due aux mécanismes de base de Sparklis.

3 Perspectives et conclusion

Notre projet est encore en développement. Actuellement, la base contient plus de 15 000 formes fléchies liées à 278 verbes pour 10 temps. Les formes fléchies ont été générées et testées par des étudiants dont la langue maternelle est le Géorgien.

4. Biliki, Georgian Language For English Speakers. See http ://lsgeorgia.com.

Les perspectives à court terme sont les suivantes. INESS ::XLE-Web de Meurer (2007) est un outil dédié aux linguistes. Il est capable d'analyser des phrases et de produire des arbres syntaxiques pour un certain nombre de langues, dont le Géorgien. Il contient un grand nombre d'informations intéressantes mais présentées sous une forme peu accessible aux débutants. Nous travaillons actuellement, en collaboration avec Paul Meurer, à intégrer les centaines de milliers de formes verbales géorgiennes existant sur cette plate-forme. Une bibliothèque de requêtes usuelles est en cours de construction. Une interface graphique dédiée est à l'étude. Des translittérations phonétiques et orientées vers l'anglais sont prévues afin d'aider les utilisateurs non francophones. Des liens vers un dictionnaire électronique seront insérés (par exemple vers le « Comprehensive Georgian-English Dictionary » de D. Rayfield et al., sur le site de la Bibliothèque Parlementaire Nationale de Géorgie.)

Dans cet article, nous avons illustré à quel point les mécanismes d'interrogation de Sparklis sont polyvalents et puissants. Nous avons aussi montré comment ces mécanismes peuvent aider les utilisateurs de Kartu-Verbs à obtenir facilement des informations sur les verbes qu'ils rencontrent dans des textes géorgiens quelle que soit leur forme. En particulier, trouver le lemme pertinent pour une entrée dans un dictionnaire donné n'est plus un problème. Kartu-Verbs peut ainsi être utilisé comme interface pour n'importe quel dictionnaire géorgien, quels que soient les principes de lemmatisation que le dictionnaire utilise pour les verbes.

Remerciements : Nous remercions Irma Grdzelidze et Tinatin Margalitadze de l'Université d'État Ivane Javakhishvili de Tbilissi pour nous avoir aidé à prendre du recul sur ce travail Grâce au programme Erasmus+ Mobilité Internationale de Crédits, les étudiants géorgiens suivants ont contribué au projet : Keti Meipariani, Mariam Asatiani, Mikheil Maisuradze, Tamari Kldiashvili, Tamar Sharabidze, Beka Chachua, Ana Idadze, Veriko Nikuradze, Tornike Tchanturia, Ana Elchishvili et Aleksandre Jajanidze. Enfin, nous tenons à remercier chaleureusement Sébastien Ferré pour son soutien à l'utilisation de Sparklis.

Références

Anderson, S. R. (1984). On representations in morphology case, agreement and inversion in Georgian. *Natural Language & Linguistic Theory 2*(2), 157–218.

Assatiani, I. et M. Malherbe (2011). *Parlons Géorgien*. l'Harmattan.

Chotiwari-Jünger, S., D. Melik'išvili, et L. Wittek (2010). *Georgische Verbtabellen*. Buske.

Daraselia, S. et S. Sharoff (2016). Enriching Georgian dictionary entries with frequency information. In T. Margalitadze et G. Meladze (Eds.), *Proceedings of the 17th EURALEX International Congress*, Tbilisi, Georgia, pp. 321–327. Ivane Javakhishvili Tbilisi University Press.

Ducassé, M. (2020). Kartu-verbs : A semantic web base of inflected georgian verb forms to bypass georgian verb lemmatization issues. In Z. Gavriilidou, M. Mitsiaki, et A. Fliatouras (Eds.), *Proceedings of XIX EURALEX International Congress*, Volume 1, pp. 81–89. SynMorPhoSe Lab, Democritus University of Thrace.

Ferré, S. (2017). Sparklis : An expressive query builder for SPARQL endpoints with guidance in natural language. *Semantic Web : Interoperability, Usability, Applicability 8*(3), 405–418.

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

Gérardin, H. (2016). *Les verbes intransitifs primaires et dérivés en Géorgien : Description morphosyntaxique, sémantique et dérivationnelle*. Ph. D. thesis, Institut National des Langues et Civilisations Orientales. UMR 7192 – « Proche-Orient-Caucase : langues, archéologie, culture ».

Margalitadze, T. (2020). Lexicography of Georgian : a brief overview.

Meurer, P. (2007). A computational grammar for Georgian. In *International Tbilisi Symposium on Logic, Language, and Computation*, pp. 1–15. Springer.

Tschenkéli, K., Y. Marchev, et L. Flury (1965). *Georgisch-deutsches Wörterbuch*, Volume 2. Amirani-Verlag Zürich.

Tuite, K. (1998). *Kartvelian morphosyntax : Number agreement and morphosyntactic orientation in the South Caucasian languages*. Lincom Europa Munich.

Summary

The Georgian language has a complex verbal system, both agglutinative and inflectional, with many irregularities. Inflected forms of a given verb can differ greatly from each other. It takes a good knowledge of Georgian grammar to go back to infinitive (the most common access lemma for dictionaries). Access to dictionaries for beginners is therefore very difficult. In addition, there is no consensus among Georgian lexicographers on which lemmas represent a verb in dictionaries. It further complicates dictionnaries access. We propose Kartu-Verbs, a base of inflected forms of Georgian verbs accessible by a logical information system. 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.