Retro-digitization in Greek dialectology and lexicography: Challenges of morpho-phonetic representation of the Cappadocian dialect

Io Manolessou, Athanasios Karasimos & Georgia Katsouda
Academy of Athens

Στόχος της παρούσας μελέτης είναι η παρουσίαση της πρώτης ψηφιακής τυποποιημένης κωδικοποίησης για την σύνταξη ενός ιεραρχικά σχολιασμένου ψηφιακού λεξικού της Καππαδοκικής διαλέκτου. Δεδομένης της ανάγκης για βαθύτερη κατανόηση και αναπαράσταση των διαλεκτικών δεδομένων στα διαφορετικά επίπεδα γλωσσικής ανάλυσης, αλλά και της έλλειψης ενός προτύπου κωδικοποίησης, έχει δημιουργηθεί ένα σχήμα/πρότυπο κωδικοποίησης για πολυεπίπεδη γλωσσική επισημείωση και λεξικογραφική αναπαράσταση. Για τη δημιουργία αυτού του προτύπου αντιμετωπίστηκαν τα ακόλουθα ζητήματα: α.) σύστημα φωνητικής μεταγραφής, τόσο στο IPA όσο και στο ελληνικό αλφάβητο (χρήση ειδικών γραφημάτων, συμμόρφωση στο πρότυπο Unicode, συμβατότητα με προηγούμενες προσεγγίσεις μεταγραφής, φιλικό προς το χρήστη), β.) σύστημα γεωγραφικού προσδιορισμού και γεωθεσίας (τυποποίηση τοπωνυμίων βάσει εξαντλητικής έρευνας των πηγών και αξιολόγηση των παραλλαγών, εναλλακτική υλοποίηση ανάλογα με το μέσο [έντυπο vs ψηφιακό], σύνδεση με ψηφιακό χάρτη και τοπωνύμια από γεωπληροφοριακά συστήματα), και γ.) σύστημα ταξινόμησης και παρουσίασης παραδειγμάτων (βάσει διαφόρων τύπων παραδειγμάτων: αποσπάσματα απομαγνητοφωνημένων προφορικών συνεντεύξεων, παλαιότερες γραπτές πηγές, απλή αφήγηση *vs.* φράσεις vs. παροιμίες vs. άσματα) και δ.) ένα συνολικό σύστημα παρουσίασης δεδομένων (συμβατότητα με τη μορφή μεγάλων διαλεκτικών λεξικών όπως το Ιστορικό Λεξικό της Νέας Ελληνικής (ΙΛΝΕ), το Ιστορικόν Λεξικόν των ελληνικων ιδιωμάτων της Κάτω Ιταλίας (ΙΛΕΙΚΙ), το Ιστορικόν Λεξικόν της Ποντιακής διαλέκτου (ΙΛΠΔ) και το Λεξικό Τσακώνικης Διαλέκτου). Το προτεινόμενο πρότυπο συγκρίνεται με αντίστοιχες διεθνείς προσπάθειες στη διαλεκτική ηλεκτρονική λεξικογραφία (π. χ. την ηλεκτρονική έκδοση του Innsbruck EDD Online 3.0), καθώς και με προηγούμενα ελληνικά έργα μικρότερης κλίμακας (π. χ. το Τριδιαλεκτικό Λεξικό).

Keywords: *Cappadocian dialects, dialectology, lexicography, historical dictionary, digital humanities*

1. Introduction

In this paper, we present various aspects of a major on-going digital project funded by the Hellenic Foundation for Research and Innovation (ΕΛΙΔΕΚ). The project, entitled DicaDland (Digitizing the Cappadocian Dialectal Landscape), runs from

2018 until 2021 and aims to produce a full documentation of the Cappadocian dialect, including the varieties of Pharasa and Silli, by exploiting all the available sources (older, written and new, oral) and the latest advances in the domain of digital humanities. It constitutes an intersection of a humanities/social sciences discipline (linguistics) and informatics, as it involves both the examination of how digital tools can be applied to specific domains such as dialectal Lexicography and Modern Greek dialectology, and how these, as domains of application, can influence the development of information technology.

The final products of the project will be (or rather, are, since we are more than half-way-through by now) two state-of-the-art major reference works, namely an interactive electronic dialectal atlas (which will constitute the first such effort in the domain of Greek Linguistics — you'll hear about it later on) and a comprehensive historical dictionary of the Cappadocian dialects (again lacking until now — you'll hear all about it now).

The project[1] is hosted by the Laboratory of Modern Greek Dialects of the University of Patras and co-ordinated by Dimitra Melissaropoulou, Associate Professor at the Aristotle University of Thessaloniki. The second Principal investigator is Athanasios Karasimos, from the Academy of Athens and the Centre for the Greek language, and the research team is complemented by a host of specialists in dialectology, including Io Manolessou and Georgia Katsouda from the Academy of Athens, as well as Angela Ralli, Mark Janse, Petros Karatsareas, Metin Bağrıaçık, Symeon Tsolakidis, Christos Papanagiotou, Georgios Chairetakis, Stavros Bompolas, and Dimitris Papazachariou.

The importance of the Cappadocian dialect both as a source of linguistic data (invaluable for investigating language-specific issues such as the diachronic evolution and dialectal differentiation of Greek and theoretical issues such as the mechanisms of language change and language contact) and as a depository of the cultural heritage of the now lost communities of the Greek-speaking populations of Asia Minor needs no comment. We will only say that Cappadocian was the only major dialectal variety of Modern Greek (the other being Pontic, Tsakonian, and Grico) which until now lacked a unified lexicographical treatment and that it is the only dialectal variety for which we now possess a wealth of new data, unavailable to previous scholarship, thanks to the recent re-discovery of 3rd generation native speakers. These two last factors were the original trigger for planning this project.

[1] See http://cappadocian.upatras.gr/en (access date 12.11.2022).

2. The electronic dictionary of the Cappadocian dialect

2.1. The dictionary and its sources

Turning to specific issues, we should firstly present the dictionary itself, its sources, format and content. The dictionary aspires to become the ultimate reference work on Cappadocian, and thus it attempts to incorporate and present all the available data on the Cappadocian language. This includes:

(i) older written sources starting from the 19th century onwards (dictionaries, glossaries, linguistic descriptions, collections of primary texts such as folktales, songs, narrations, riddles etc.), most of which were collected and digitized thanks to an earlier project of the Laboratory of Modern Greek Dialects, AMiGre, and are available online[2].

(ii) new oral recordings from current 3rd generation native speakers (descendants of Cappadocian refugees) collected the last decade. Special emphasis was placed on the exploitation of this new material, so that the dictionary under preparation is not a simple "super-container and comparative presenter" of already available but scattered data, but an opportunity for the presentation of new information. This allows also for a "diachronic" examination of the evidence, as we have the possibility to examine side-by-side data which may be divided by more than 100 years.

However, the dichotomy between written and oral sources has given rise to a major problem: whereas older sources are roughly equally distributed with respect to geographical provenance (i.e., data is available for almost all Cappadocian settlements, ca. 20 in number), oral data, from current speakers, are available only from 2–3 major communities, and mostly from that of Misti, which was the largest. This creates an imbalance in the lexicographical treatment of words, phenomena and senses. Another issue requiring special attention is the fact that the older material was in part collected by amateurs, or at a time when linguistic descriptive tools had not yet been sufficiently developed, and therefore it is to a certain extent

[2] See http://amigredb.philology.upatras.gr/ (access date 12.11.2022).

less reliable than the oral material, containing many inaccuracies which can no longer be assessed.

Considering the large size of the corpus of data on which the dictionary is based, it is not surprising that the size of the dictionary is equally large: we are currently in the middle of letter S, and our lemma-list contains ca. 7000 entries. It is therefore projected that at the end of the project the dictionary will include ca. 8500 entries. This makes it much larger than the, similar in conception, dictionary of South Italian Greek (ca. 6.000 entries) but still smaller than the Tsakonian (ca. 13.000 entries) and the Pontic dictionaries (20.000 entries) — of course the data in each Cappadocian entry is much larger than in either of these last two and is similar to Karanastasis (1984). Luckily, "size does NOT matter" in the case of electronic publications, so we are letting ourselves run free in this respect. However, apart from the on-line edition, the "Historical Dictionary of the Cappadocian dialect" will also appear in print form, which will probably be realized as a four-volume-set.

The entry compilers of the Historical Dictionary of the Cappadocian dialect are, in alphabetical order, Gogo Katsouda, Io Manolessou, Symeon Tsolakidis, Christos Papanagiotou and George Chairetakis. Io Manolessou also acts as editor-in-chief, while Athanasios Karasimos and Io Manolessou are responsible for the creation of the dictionary's DTD, i.e. the parametrization of the lexicographical software and the implementation of the dictionary's electronic platform.

2.2. The e-dictionary

The format of the Dicadland dictionary attempts to conform to standards of state-of-the-art academic-level Dictionary Writing Systems (DWS), after careful evaluation of available options, and adheres to the most recent advances in the domain of electronic lexicography (see e.g. Granger & Paquot 2012). It is built using the powerful professional dictionary editing software TLex Suite, one of the most widely-used state-of-the-art DWS internationally. The decision to use a professional lexicographical tool, rather than one of the freely available online lexicographic platforms such as "Lexonomy" or "Matapuna" was forced by the size and complexity of the data to be treated, since free platforms do not leave much room for parametrization. On the other hand, the solution of creating from scratch an in-house custom-made lexicographic tool capable of handling our complex dialectal material was deemed impractical given the relative short time-range of the project.

In effect, the creation of a digital platform for the "Historical Dictionary of the Cappadocian Dialect" was something half- way between a retro-digitization and a new digitization proper (hence the title of the present paper). This is because we took the conscious decision to adhere the microstructure of the other major dialectal dictionaries of Greek we have mentioned above (Pontic, Tsakonian, South Italian), which in turn are all based on the "mother of all dialectal dictionaries", the *Historical Dictionary of Modern Greek* (ILNE) of the Academy of Athens. Note that also the Dictionary of Medieval Greek of Kriaras (1968–) shares the same structure, as it too, was based, as a concept, on the *Historical Dictionary of Modern Greek*.

διακλύζω Σιν. *δακλύζω* Κερ.Κοτ.Τραπ. Χαλδ. *δακλύω* Χαλδ. *δακλῡ̃ω* Οἰν. *δακλῶ* Χαλδ. *δακλῶ* Χαλδ. *δακλύγω* Σάντ.Χαλδ. Παθητ. *δακλύσκουμαι* Μετοχ. *δακλυσμένος*, *δακλυγμένος*.

Ἀπὸ τὸ ἀρχ. *διακλύζω*=ἐκπλύνω, ξεπλύνω. 1) Πλύνω καὶ ξεβγάλλω τι μὲ καθαρὸ νερό: *Δακλύζω τὸ ποτήριν*. 2) Καθαρίζω σκεῦος τι διακινῶν αὐτὸ καὶ ἐκτινάσσων τὸ περιεχόμενον ὕδωρ. 3) Πλύνω: *Δακλύζω τὸ στόμα μ'*.

Picture 1: Papadopoulos (1958: 263)

***διάβασμα** το, *δέβασμα* ΒΧ, *δᾶβασμα* ΒΧ, *ζβάισμα* ΜΠΤ, *ζιβάισμα* Μ.

Οι δύο τελευταίοι τύποι από τον τύπ. *ζβαΐχου* <*διαβάζω*.

1. Διάβασμα κοιν. *Το δέβασμα σ' καλέ ρ, δεβάζ' καλά* Χ, *το ζβάισμά σι καλέρ ένι, ένι ζβαΐχου κα* Μ, το διάβασμά του καλό είναι, διαβάζει καλά. *Όσι έχου ζβαΐσματα σάμερε* Μ, δεν έχεις διαβάσματα σήμερα; **2.** Ευχή θρησκευτική κοιν. *Ένι θέου ζβάισμα, να μόλει ο παπά να νι ζβαΐσει* Μ, θέλει διάβασμα (ευχής), νά 'ρθει ο παπάς να τον διαβάσει. **3.** Καθοδήγηση, υποβολή συμπεριφοράς κοιν. *Ν' εφωνιάε για ζβάισμα, ό καταβήτθερε* Μ, τον φώναξε να τον «διαβάσει», δεν κατάλαβες;

Picture 2: Costakis (1983: 262)

δέχομαι (δέχομαι) 'Απουλ. (Καλημ. Μαρτάν.) γ' ἐνικ. *δέχεται* 'Απουλ. (Καλημ.) *δέφεται* Καλαβρ. (Μπόβ.) *δέχει* 'Απουλ. (Μαρτάν.) παθ. ἀόρ. γ' ἐνικ. *ἐδέφτη* 'Απουλ. (Μαρτάν.) 'Απαρ. παθ. ἀορ. *δεφτῆ* 'Απουλ. — Rohlfs Beitr., 151 — Sprichw., 143 (438).

Τὸ ἀρχ. *δέχομαι*.

1) Δέχομαι, κάνω δεχτὸν κάποιον ὅ.π.: *'è σ-σε σώdζω δεφτῆ* = δὲν ἠμπορῶ νὰ σὲ δεχτῶ Rohlfs Beitr., 151. *Τοῦσ' ὁ καιρὸ φσυχρὸ 'é μ-μου dέχεται* = αὐτὸς ὁ ψυχρὸς καιρὸς δέ μου εἶναι εὐπρόσδεκτος Rohlfs, Sprichw., ὅ.π. 'Η σημ. ἤδη ἀρχ. 2) 'Ως τριτοπρόσωπο, ταιριάζει, πρέπει ὅ.π. : *Μοῦ dέχεται 'είσ' ἡ gραβάτα ;* = μοῦ ταιριάζει αὐτὸς ὁ λαιμοδέτης ; Μαρτάν. *'εῖτ-τ' ἀδ-δάϊ 'é σ-σου dέχεται* = αὐτὸ τὸ καπέλλο (σκιάδι) δὲ σοῦ πάει αὐτόθ. *'ἤ d-dέχει οὔτ-το πρᾶμα* = τῆς ταιριάζει αὐτὸ τὸ πρᾶμα Καλημ. *'é τ-του dέχει νὰ πῆ οὔτ-το πρᾶμα* = δέν του ἁρμόζει νὰ πῆ αὐτὸ τὸ πρᾶμα αὐτόθ. Πβ. τὸ Νεοελλην. *μοῦ δέχεται* = μὲ συμφέρει Πελοπν. (Κορινθ.). || Φρ. *Μοῦ dέχεται ἄρια ἀτ-τὴ τ-τάλασ-σα* = μὲ ὠφελεῖ ὁ ἀέρας τῆς θάλασσας Καλημ. *Μοῦ dέφεται* = μοῦ χρησιμεύει Μπόβ. β) Δίνεται ἡ εὐκαιρία, ἡ σύμπτωση Μαρτάν. : *'Εdέφτη καὶ βόρασα ἅ σ-σάκ-κο σ-σιτάρι* = δόθηκε ἡ εὐκαιρία καὶ ἀγόρασα ἕνα σακκὶ σιτάρι.

Picture 3: Karanastasis (1986: 253)

ἀγναντίζω (II) Λαρδαν. ('Οφρύν.) Θρᾴκ. (Σάκκ.) *ἀγναδίζω* Κρήτ. *ἀγναδίζου* Θρᾴκ. Ἴμβρ. Σαμοθρ. *ἀγνατίζω* Κύπρ. Μεγίστ. *ἀγναδῶ* Καππ. (Σίλατ.) *ἀγνατοῦ* Λυκ. (Λιβύσσ.)

Ἐκ τοῦ Τουρκ. *ağnadım* διαλεκτ. ἀορ. τοῦ ρ. *anlamak*.

1) Ἐννοῶ, καταλαμβάνω ἔνθ' ἀν. : *Ἐνι ξουκέφαλους κὶ δὲν ἀγναδίζ' τὶ dούν λέγου* (εἶναι ξεροκέφαλος καὶ δὲν ἐννοεῖ τί τοῦ λέγω) Σαμοθρ. Συνών. ἰδ. ἐν λ. *ἀγνατεύω*.

2) Μετβ. κάμνω τινὰ νὰ ἐννοήσῃ Θρᾴκ. (Σάκκ.) Σαμοθρ. : *Δὲ δουλῶ ν' ἀγναδίσου τοὺ χάλι μ'* (δὲν δύναμαι νὰ παραστήσω τὴν κατάστασίν μου) Σαμοθρ. *Ἂν μὲ τ' ἀγνάντιζες, θὰ τὸ ἔκαμνα* Σάκκ.

Picture 4a: From ILNE (1938: entry *ἀγναντίζω* II)

διάρμισμα τό, *διάρμισμα* [ðjármizma] Αἰολ. (Κυδων.) Θρᾴκ. ἐ.Ἑλλ. (Αἶν.) Ἰων. (Βουρλ.) Κάρπ. Κρήτ. Νάξ. (Ἀπέραθ.) Σέρρ. — *Λεξ.* Πρω. Δημητρ. Σταμ. Τσιούν. Ὑπερλ. *διάρμ'σμα* [ðjármzma] Ἴμβρ. Λέσβ. *γιάρμισμα* [jármizma] Κάρπ. *τζάρμισμα* [dzármizma] Κάρπ. *ϑιάρμισμα* [ϑjármizma] Κάλυμν. *διάρμοσμα* [ðjármozma] Κάρπ.

Ἀπὸ τὸ ρ. *διαρμίζω* (θ. ἀορ. *διαρμισ-*), ὅπου καὶ τύπ. *γιαρμίζω*, *τζιαρμίζω*, καὶ τὸ παραγωγ. ἐπίθμ. *-μα*. Ὁ τύπ. *διάρμοσμα* μὲ παρετυμολογ. ἐπίδρ. τοῦ ρ. *ἁρμόζω*.

1) Ἡ ἐνέργεια καὶ τὸ ἀποτέλεσμα τοῦ ρ. *διαρμίζω* **2**, ἡ τακτοποίηση, ἡ τάξη Αἰολ. (Κυδων.) Θρᾴκ. ἐ.Ἑλλ. (Αἶν.) Ἴμβρ. Ἰων. (Βουρλ.) Κάλυμν. Κάρπ. Κρήτ. Λέσβ. Σέρρ. — *Λεξ.* Πρω. Δημητρ. Σταμ. Τσιούν. Ὑπερλ.: *Θέλω καλὸ διάρμισμα στὸ σπίτι, ὄχι τσαπατσούλικα!* Ρεθύμν. (Μαλάκ.) *Μωρέ, χαρὰ στὸ γιάρμισμα ἀπού 'καμες τοῦ σπιτίου!* Κάρπ. Συνών. βλ. λ. *διάρμιση*. Ἀντίθ. *ἀδιαρμισιά*. **2)** Ἡ ἐνέργεια καὶ τὸ ἀποτέλεσμα τοῦ ρ. *διαρμίζω* **1**, ἡ φύλαξη Ἰων. (Βουρλ.) **3)** Ἡ ἐνέργεια καὶ τὸ ἀποτέλεσμα τοῦ ρ. *διαρμίζω* **3**, τὸ ἀνακάτωμα Νάξ. (Ἀπέραθ.): *Ἤκαμές του πάλι σήμερα διάρμισμα τοῦ νεροῦ! Εἶδα λοᾶται καὶ τὸ μετατόπισες ἀ' τό 'να μπιθάρι στ' ἄλλο;*

Picture 4b: From ILNE (2016: entry διάρμισμα)

δέμα(ν) το, Παράφρ. Μανασσ. Β 300, Ερμων. Κ 195, Διήγ. παιδ. (Tsiouni) 627, Λιβ. Ρ 1427, Λιβ. Sc. 377, 752, Λιβ. Esc. 1494, 1844 (κριτ. υπ.), Λιβ. Ν 1343, 1645, Αχιλλ. Ν 377, Ψευδο-Γεωργηλ., Ἀλ. Κων/π. 590, Σκλέντζα, Ποιήμ. 3ᵃ, Πεντ. Αρ. XXX 3, 13, 15, Στάθ. Γ' 439, Ροδολ. (Μανούσ.) Γ' [428], Ροδολ. Ε' [81], Συμμ., Παστ. φίδ. Γ' [831], Ε' [216], Τζάνε, Κρ. πόλ. 272^{26}· *δ ἡ μ μ α ν*, Μαχ. $20^{17.23}$, 22^{12}, $252^{14.25}$, 654^{28}.

Το μτγν. ουσ. *δέμα*. Η λ. και σήμ. (Δημητράκ., λ. *δέμα*) και ως τοπων. (Βλ. Σαρρή Ι., Αθ. 40, 1928, 136).

1) Αυτό με το οποίο δένομε κ., σκοινί, ταινία (Η σημασ. μτγν., Ι.-Σ, λ. *δέμα* **1**): *ἔλυσε με τα χέρια-της τον πιττακίον το δέμαν* Λιβ. Sc. 377. **2)** Δέμα, δεμάτι (Η σημασ. τον 5. αι., Lampe, Lex., λ. *δέμα* **1** και σήμ., Δημητράκ., ὁ.π. 3): *λάβρες, βερτόνια, δέματα καλὰ ἐτσι σοβεμένα* Ροδολ. (Μανούσ.) Γ' [428]. **3)** Δεσμός: *εσὺ το ἀδετον το δέμα της Τριάδος εἶσαι και μονοούσιος ἀγάπη της μονάδος* Σκλέντζα, Ποιήμ. 3ᵇ· *μέσα εις τη φιλιά ποὺ 'ναι σ' εσάς η τόση απόμενε παντοτινὸ δέμαν ἀναμεσάσας* Ροδολ. Ε' [81]. **4)** Δεσμά: *'χ τα χέρια λὺσετέ-την κι' από τ' ανάξια δέματα λεὐ-*

Picture 5: Kriaras (1977: 9)

The entry structure or microstructure of all the above-mentioned projects is, as you can see, tripartite: It consists of the following "sections", which, were translated in the XML format of the Cappadocian dictionary's template as "Elements" (for more information, see Karasimos et al. 2020):

(i) A **Formal Section**, where the variant dialectal forms are set out, with phonetic transcription, part of speech characterization and geographical distribution, realized as "attributes", and with bibliographic source data available as a drop-down list.

(ii) An **Etymological section**, where the word's origin (native/loanword) and dating are recorded. This is our only free-text field, as etymology remains a domain where XML encoding has yet to be standardized.

(iii) A **Senses section**, with numbered senses and sub-senses, which include, as attributes, definitions, examples, quotations and documentation from oral and written corpora, and also including "special" types of examples, such as proverbs, songs and riddles. The latter type of examples required additional customization of the DTD, as they involved on the one hand "double" translations (literal and metaphorical) and on the other special formatting (verses).

3. Special issues of morphophonetic representation

3.1. Transcription

We have already touched upon the issue earlier, when we mentioned the problem that the older written sources of the dictionary employ a variety of symbolisms, not consistent with each other and not always easily interpretable. All of them had to a) be unified and homogenized and b) be represented in a system easily accessible and comprehensible. That is, the transcription needed to be given both in the International Phonetic Alphabet (for reasons of scientific clarity and accessibility to the international academic community) and in some form of the Greek alphabet (for reasons of accessibility to the general public).

In this respect, the project was greatly assisted by the previous work done on the topic by the *Historical Dictionary of Modern Greek* and its transcription system, which the Cappadocian historical dictionary also adopts. The specialized groundwork of the Historical Dictionary provides correspondence tables for all the variant transcription systems to be found in the standard dialectological publications on Cappadocian.

Καππαδοκική

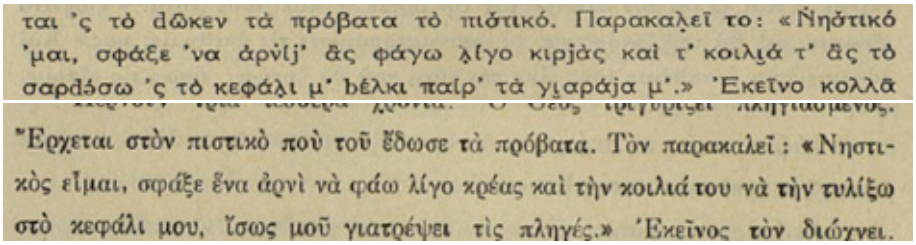
Καρολίδης 1885	Lagarde 1886	Dawkins 1916	Ανδριώτης 1948	Κοσσυφίδου 1951	Μπαμπινιώτης - Κακούρ. 1960	Κωστούρος 1964	Αναστασιάδης 1976	Janse 2004, 2011	IANE	ΔΦΑ
		ä	ä		ä	ä	ä		ä	æ
		ö			ö		ö	ö	ö	ø
		ü			ü	ü	ü	ü	ü	y
		ə		ə	ə		ə	ɣ	ı	ɯ
			ṕ, ť, ḱ	ṕ, ť, ḱ	ṕ, ť, ḱ	ṕ, ť, ḱ	ṕ, ť, ḱ		ṕ, ť, ḱ	p^h, t^h, k^h
			ǩ	ǩ	ǩ	ǩ	ǩ		ǩ	k
				ǧ	ǧ		ǧ	c	ǧ	c
		q							q	q
b, d, g		b, d, g	b, d, g	b, d, g	b, d, g	b, d, g	b, d, g	b, d, g	b, d, g	b, d, g
							ɣ́ ǩ	g	ɣ ǩ	g
		kh	x̌	x̌	x̌	x̌	x̌	x	x̌	x
		x́		x̂	x̂		x̂	ç	x̂	ç
		gh	ɣ̌	ɣ̌	ɣ̌	ɣ̌	ɣ̌		ɣ̌	ɣ
		ɣ́		j	j		j	j		j
			λ̌	λ̌	λ̌	λ̌	λ̌		λ	l
		λ́						lj	λ̌	ʎ
			ṅ	ṅ	ṅ	ɣ, ṅ	ṅ	ń		n
			ŋ	ŋ	ŋ		ŋ	ŋ	ṅ	ɲ
		ɣ	ɣ (?)	ṅ (?)	ṅ (?)		ṅ (?)			ŋ
sch	σσ	ǯ, ξ	ǯ	ǯ	ǯ, ð	ǯ, ð	ǯ	š	ǯ	ʃ
		ζ	ζ	ζ	ζ	ζ	ζ	ž	ζ	ʒ
ć	τσ	τσ							τσ	ts
ć	πζ	č	τǯ	τǯ	τð, τǯ	τð	τǯ	tš	τǯ	ʧ
ǵ	ζζ	j	τζ	τζ		τζ	τζ	dž	τζ	ʤ

Picture 6: The variant transcription of Cappadocian texts and manuscripts (Manolessou et al. 2012)

If one is not aware of the potential variety of representation practices in the written sources, one can easily misread the data. To give a couple of concrete examples.

3.1.1. The case of Kentro Mikrasiatikon Spoudon approach

The system employed by the publications of the Kentro Mikrasiatikon Spoudon employ the symbols λ with superimposed or undersided dot, and ν with superimposed dot (Mavroxalybidis & Kesisoglou 1960).


ται 'ς τὸ dῶκεν τὰ πρόβατα τὸ πιότικό. Παρακαλεῖ το: «Νηότικό
'μαι, σφάξε 'να ἀρνίj' ἂς φάγω λίγο κιρjὰς καὶ τ' κοιλιά τ' ἂς τὸ
σαρdάσω 'ς τὸ κεφάλι μ' bέλκι παίρ' τὰ γιαράjα μ'.» Ἐκεῖνο κολλᾶ

Ἔρχεται στὸν πιστικὸ ποὺ τοῦ ἔδωσε τὰ πρόβατα. Τὸν παρακαλεῖ: «Νηστι-
κὸς εἶμαι, σφάξε ἕνα ἀρνὶ νὰ φάω λίγο κρέας καὶ τὴν κοιλιά του νὰ τὴν τυλίξω
στὸ κεφάλι μου, ἴσως μοῦ γιατρέψει τὶς πληγές.» Ἐκεῖνος τὸν διώχνει.

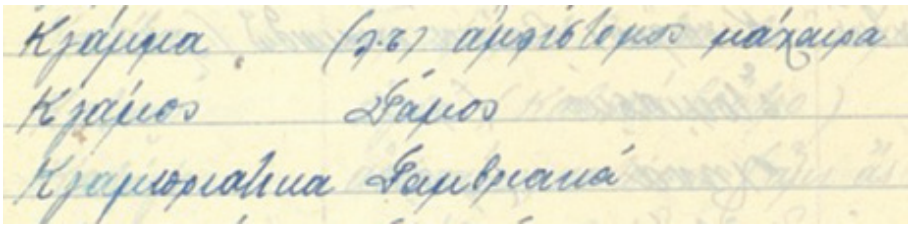
Picture 7: Text sample from Mavroxalybidis & Kesisoglou (1960)

This representation is in fact a trap: one is led to believe that it records a special kind of pronunciation of /l/ and /n/, when in fact it is nothing of the kind: these symbols represent the "normal" sounds [l] and [n]. The special notation starts from the assumption that the "default" realization of the phonemes /l/ and /n/ in Standard Modern Greek, when they are followed by the front vowels /e/ and /i/, is as the palatal allophones [ʎ] and [ɲ] — i.e. that Standard Modern Greek is like the local dialect of Patras or Eastern Crete. So a "special notation", the dot in this case, is used to signal the ABSENCE of palatalization from Cappadocian. The dot just means that the words in question are supposed to be read: [paraka'li] [niʃti'ko] [ar'ni] ['liɣo] and not [paraka'ʎi] [ɲiʃti'ko] [ar'ɲi] ['ʎiɣo].

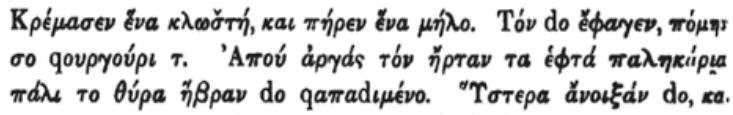
3.1.2. Three Cappadocian sub-dialects

The second example is from three Cappadocian sub-dialects, those of Malakopi, Flogita and Silata, possess a realization of /k/ or /ɣ/ which is very velar, in fact [q]. This sound is represented in a variety of ways in the sources:

As <ϰ̇> with a superimposed dot, as plain <ϰ> or as <ϰγ>:


Κράμμα (γ.ε.) ἀμφιθέατρο μάχαιρα
Κράμος Δάμος
Κραμοριάλικα Δαμβριανά

Picture 8: Text sample from Anthopoulos (1937–1938)


Κρέμασεν ἕνα κλωβάτή, και πήρεν ἕνα μήλο. Τόν do ἔφαγεν, πόμη
σο qουργούρι τ. Ἀπού ἀργάς τόν ἤρταν τα ἐφτά παληκούρια
πάλι το θύρα ἤβραν do qαπαδιμένο. Ὕστερα ἄνοιξάν do, κα.

Picture 9: Text sample from Dawkins (1916) (Silata)

All three representations <κ>, <κ̇> and <κγ>, are in fact the same sound, [q], which may correspond to two different phonemes, /k/ or /ɣ/. The *Historical Dictionary* of the Academy of Athens also offers a custom-made system of phonetic representation, based on the Greek alphabet, with some additional symbols used as diacritics, which combines one the one hand systematic one-to-one mapping with the IPA and at the other increased readability and comprehensibility through the retention of standard Greek orthography. We cannot stress how crucial this is for dialects such as Cappadocian, where massive phonetic and morphological changes have rendered even common words of the standard completely unrecognizable.

For example, it is by far preferable to spell *αυίτσα* or *ευίτσα* instead of *αβίτσα* /avitsa, since the etymological connection to *αυγή*, *αυγίτσα* becomes immediately obvious and the sense 'dawn', 'morning' transparent. Similarly, *νεγγιριώνα* spelt with double <γγ> instead of <γκ> or <g> provides an insight to the origin of the word, that it has something to do with *αγγούρια* 'cucumbers', and thus can help one to guess its meaning 'vegetable garden'

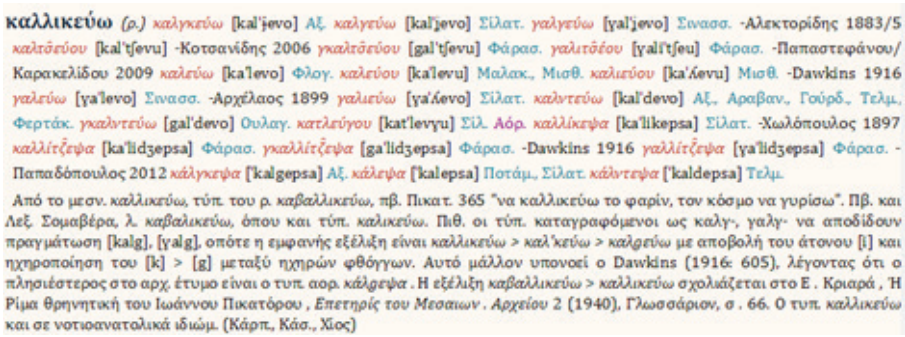
Of course, the use of diacritics adds a level of difficulty in digitization, as these are not always easily rendered with the standard fonts available, especially when they stack one on top of the other. Whereas the *Historical Dictionary of Modern Greek* employs a custom-made font of its own, **Athens Academy fonts**, it was a conscious decision early on in our planning not to use any fonts or glyphs which are not part of the Unicode Standard, and to employ only fonts which belong to Microsoft's ClearType collection (such as Cambria, Calibri, Corbel) because these allow for much better alignment of diacritics on top of main letter glyphs.

4. Lemmatization

One of the most serious problems we face in compiling the Cappadocian dictionary is that of lemmatization, i.e. the selection of the headword This is because this dialect has undergone a great number of phonetic and morphological changes, more than any

other Greek dialectal variety, both due to the length of time of its separation / isolation from the other forms of Greek, about a thousand years, (from the 11th c.), and due to intense contact with Turkish. This entails that a) its forms frequently are very distanced from the Standard and most other dialectal varieties of Greek and b) more importantly, each local Cappadocian form is distanced from all others; each settlement may present its own variant (see also Manolessou & Katsouda 2021; Forthcoming).

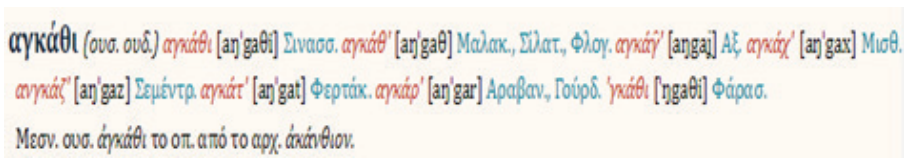
Let's give a concrete example:


καλλικεύω (ρ.) καλγκεύω [kal'ɟevo] Αξ. καλγεύω [kal'ɟevo] Σίλατ. γαλγεύω [γal'ɟevo] Σινασσ. -Αλεκτορίδης 1883/5 καλτσεύου [kal'ʧevu] -Κοτσανίδης 2006 γκαλτσεύου [gal'ʧevu] Φάραισ. γαλιτσέου [γali'ʧeu] Φάραισ. -Παπαστεφάνου/Καρακελίδου 2009 καλεύω [ka'levu] Φλογ. καλεύου [ka'levu] Μαλακ., Μισθ. καλιεύου [ka'ʎevu] Μισθ. -Dawkins 1916 γαλεύω [γa'levo] Σινασσ. -Αρχέλαος 1899 γαλιεύω [γa'ʎevo] Σίλατ. καλντεύω [kal'devo] Αξ., Αραβαν., Γούρδ., Τελμ., Φερτάκ. γκαλντεύω [gal'devo] Ουλαγ. κατλεύγου [kat'levγu] Σίλ. Αόρ. καλλίκεψα [ka'likepsa] Σίλατ. -Χωλόπουλος 1897 καλλίτζεψα [ka'liʤepsa] Φάραισ. γκαλλίτζεψα [ga'liʤepsa] Φάραισ. -Dawkins 1916 γαλλίτζεψα [γa'liʤepsa] Φάραισ. -Παπαδόπουλος 2012 κάλγκεψα ['kalgepsa] Αξ. κάλεψα ['kalepsa] Ποτάμ., Σίλατ. κάλντεψα ['kaldepsa] Τελμ.

Από το μεσν. καλλικεύω, τύπ. του ρ. καβαλλικεύω, πβ. Πικατ. 365 "να καλλικεύω το φαρίν, τον κόσμο να γυρίσω". Πβ. και Λεξ. Σομαβέρα, λ. καβαλικεύω, όπου και τύπ. καλικεύω. Πιθ. οι τύπ. καταγραφόμενοι ως καλγ-, γαλγ- να αποδίδουν πραγμάτωση [kalg], [γalg], οπότε η εμφανής εξέλιξη είναι καλλικεύω > καλ'κεύω > καλγεύω με αποβολή του άτονου [i] και ηχηροποίηση του [k] > [g] μεταξύ ηχηρών φθόγγων. Αυτό μάλλον υπονοεί ο Dawkins (1916: 605), λέγοντας ότι ο πλησιέστερος στο αρχ. έτυμο είναι ο τυπ. αορ. κάλγεψα. Η εξέλιξη καβαλλικεύω > καλλικεύω σχολιάζεται στο Ε. Κριαρά, Ή Ρίμα θρηνητική του Ιωάννου Πικατόρου, Επετηρίς του Μεσαιων. Αρχείου 2 (1940), Γλωσσάριον, σ. 66. Ο τυπ. καλλικεύω και σε νοτιοανατολικά ιδιώμ. (Κάρπ., Κάσ., Χίος)

Picture 10: The entry καλλικεύω from TLex software entry

For the verb 'to ride' each settlement presents different variants, such as καλγκεύω, καλτσεύου, καλιεύου, γαλεύω, γκαλντεύω, κατλεύγω. The only way to unify them is to subsume them under an unattested headword καλλικεύω, which corresponds to the word's medieval etymon, the point of origin of all the changes (n.b. in turn this comes from the Standard word καβαλλικεύω). Whether the headword will be an unattested reconstructed one or not to a certain extent depends on the available evidence of the sources. One of the Cappadocian dialects, that of Sinasos, is much more conservative than all the others (or much more affected by the standard?). If the word happens to be preserved in the texts from Sinasos, then the headword will probably be attested. For example:


αγκάθι (ουσ. ουδ.) αγκάθι [aŋ'gaθi] Σινασσ. αγκάθ' [aŋ'gaθ] Μαλακ., Σίλατ., Φλογ. αγκάϑ' [aŋgaj] Αξ. αγκάχ' [aŋ'gax] Μισθ. αγκάζ' [aŋ'gaz] Σεμέντρ. αγκάτ' [aŋ'gat] Φερτάκ. αγκάρ' [aŋ'gar] Αραβαν., Γούρδ. 'γκάθι ['ŋgaθi] Φάραισ.

Μεσν. ουσ. άγκάθι το οπ. από το αρχ. άκάνθιον.

Picture 11: The entry αγκάθι and its variant formations

The problem becomes more acute when the phonetic change involves the initial vowel or consonant of the word, as this affects the alphabetical order of the dictionary (this is less of a problem in an electronic than in a print dictionary, but still there are many readers who "browse" instead of just "search").

A concrete example involves the adaptation of loanwords from Turkish which start with /k/ (e.g. καμπούρης 'hunchback' from Turk. *kambur*). Since the local Turkish dialects of Eastern Anatolia present a change /k/ > /ɣ/, loanwords entering Cappadocian from local Turkish varieties often show initial /ɣ/, e.g. γαμπούρ' Misti.

καμπούρης *(επίθ.)* καμπούρης [kam'buris] Γούρδ. *καμπούρ'* [qam'bur] Μαλακ. *γαμπούρ* [yam'bur] Μισθ., Φάρασ. *γκαμπ'ρ* [gam'bur] Ουλαγ. *θηλ. γαμπουρούτσα* [yambu'rutʃa] Φάρασ.
Από το τουρκ. επίθ. *kambur* = καμπούρης.

Picture 12: The entry καμπούρης και its variant formations with initial /k/ and /ɣ/

If there exists at least one subdialect which does not present this phenomenon, the case is straightforward, the headword will be an attested word starting with <κ>. But if not, then the headword, unless one wishes to resort to reconstruction again (less justifiable in the case of loanwords), will start with <γ>.

γαπαχλούς *(ουσ. αρσ.)* *γαπαχλούς* [ɣapax'lus] Φάρασ.
Από το τουρκ. επίθ. *kabaklı*.
1 Κολοκύθι ως γλυκό κουταλιού

γαπαχουφαΐ *(ουσ. ουδ.)* *γαπαχουφαΐ* [ɣapaxufa'i] Φάρασ.
Από τα ουσ. καμπάκι, όπου και τύπ. *γαπάχι*, και *φαΐ*.
Φαγητό με κολοκύθια.

Picture 13: The entries γαπαχλούς and γαπαχουφαΐ

Here the derivative and the compound are to be found only in Pharasa, and so start with <γ>. However, the simplex word is to be found in many areas, and therefore the headword starts with <κ>, something which disrupts the word-family.

καμπάκι *γκαμπάκ'* [gabak] Ουλαγ. *φαμπάχ'* [qa'bax] Μαλακ., Ποτάμ. *γαμπάχ'* [ya'bax] Αραβαν., Μισθ. *γαπάχι* [ya'paxi] Φάρασ. Πληθ. *γαμπάια* [ya'baja] Μισθ.
Από το τουρκ. ουσ. *kabak* = κολοκύθι.
1 Κολοκύθι ό.π.τ.: *Γαμπαχιού γούτσια* (Σπόρια κολοκυθιού) Μισθ. -Κοτσανίδης 2006 *Εγώ μι γαπάϊα να σπέρου* (Εγώ με κολοκύθια θα σπείρω) Μισθ. -ΑΠΥ-Καρατσ. *Έχου γαμπάχια, ε ὀάνου δα φέτις φέτις, τεγανίζου δα τρώου· μετά εκείνου του γαμπάχ' ξύνου δου, ξύνου δου, μάζου λίου αλεύρ' ὀάνου κιοφτάδις* (Έχω κολοκύθια, τα κάνω φέτες-φέτες, τα τηγανίζω τα

Picture 14: The entry καμπάκι

It really is questionable whether one should create an unattested form such as "καμπακοφάι" in order to keep the family together. As in the previous case, the issue is again connected with the availability of attestations. If we had more sources discussing food in other Cappadocian villages, perhaps our lemmatization would be different, as some forms with initial /k/ may have been available.

5. Concluding remarks

In the short time slot available to us we have hoped to show that the Dicadland project is a major, state-of-the-art endeavour in the field of dialectal lexicography, which will produce a concrete and highly useful work of reference. The data discovered and presented allow for important advances in our knowledge of the Greek dialects. This includes both the interpretation of known data (homogeneization, phonetic analysis, new and improved etymologies, explanation of changes, establishment of semantic ranges through comparative examination) and the addition of new and original data. The methodology adopted brings to the fore many thorny issues of dialectal lexicography, which we hope, if not to solve, at least to highlight from several aspects.

To conclude, we provide a golden standard template for future dialectical dictionaries, computational lexicography and digital humanities and we introduce a new era of Digital Humanities in Dialectology and Lexicography by following the dominant trends and schemata of this new discipline. Moreover, we build a concrete and fully adaptable basis for the preservation of endangered dialects of Modern Greek and their cultural heritage. Finally, the open data — open access policy will help the extension of the research community and engage the native speakers to provide more data in the near future.

References

Anthopoulos, Th. 1937–1938. *Φλογητά* [*Flogita*]. Unpublished manuscript, ILNE archive No. 1547. Athens: Academy of Athens [in Greek].

Costakis, Th. 1986. *Λεξικό της τσακωνικής διαλέκτου* [*Dictionary of the Tsakonian Dialect*]. Vol. 1. Athens: Academy of Athens [in Greek].

Dawkins R. M. 1916. *Modern Greek in Asia Minor. A Study of the Dialects of Sílli, Cappadocia and Pharasa with Grammar, Texts, Translations and Glossary*. Cambridge: Cambridge University Press.

Granger, M. & S. Paquot. 2012. *Electronic Lexicography*. Oxford: Oxford University Press.

ILNE 1933 — *Ἱστορικὸν Λεξικὸν τῆς Νέας Ἑλληνικῆς, τῆς τε κοινῶς ὁμιλουμένης καὶ τῶν ἰδιωμάτων* [*Historical Dictionary of Modern Greek, both of the Standard and the Dialects*]. Vol. 1. Athens: Academy of Athens [in Greek].

ILNE 2016 —. *Ἱστορικὸν Λεξικὸν τῆς Νέας Ἑλληνικῆς, τῆς τε κοινῶς ὁμιλουμένης καὶ τῶν ἰδιωμάτων* [*Historical Dictionary of Modern Greek, both of the Standard and the Dialects*]. Vol. 7. Athens: Academy of Athens [in Greek].

Innsbruck EDD Online 3.0 — English Dialect Dictionary, University of Innsbruck. Available at: http://eddonline-proj.uibk.ac.at/edd/termsOfUse.jsp (accessed on 28.11.2022).

Karanastasis, A. 1986. *Ιστορικόν λεξικόν των ελληνικών ιδιωμάτων της Κάτω Ιταλίας* [*Historical Dictionary of the Greek Dialects of South Italy*]. Vol. 2. Athens: Academy of Athens [in Greek].

Karasimos, A., I. Manolessou I. & D. Melissaropoulou. 2020. Creating a DTD template for Greek dialectal lexicography: the case of the Historical Dictionary of the Cappadocian dialect. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds), *Proceedings of the 19th Congress of the European Association for Lexicography- EURALEX XIX, Alexandroupolis, September 2021*. Vol. 1. Komotini: Democritus University of Thrace, 305–314.

Kriaras, E. 1968–. *Λεξικὸ τῆς μεσαιωνικῆς Ἑλληνικῆς δημώδους γραμματείας, 1100–1669* [*Dictionary of Greek Medieval Vernacular Literature, 1100–1669*]. Vol. 1–22. Thessaloniki: Centre for the Greek Language [in Greek].

Kriaras, E. 1977. *Λεξικὸ τῆς μεσαιωνικῆς Ἑλληνικῆς δημώδους γραμματείας, 1100–1669* [*Dictionary of Greek Medieval Vernacular Literature, 1100–1669*]. Vol. 5. Thessaloniki: Centre for the Greek Language [in Greek].

Manolessou, I. & G. Katsouda. 2021. Drawing the line between synchrony and diachrony in historical and dialectal lexicography. In Z. Gavriilidou, L. Mitits & Sp. Kiosses (eds.), *Proceedings Book of Euralex XIX (Congress of the European Association for Lexicography)*. Vol. 2. Komotini: Democritus University of Thrace, 83–92.

Manolessou, I. & G. Katsouda. Forthcoming. On lemmas and dilemmas again: problems in historical dialectal lexicography. In *Proceedings of the 11th International Conference on Historical Lexicography and Lexicology (Universidad de La Rioja, 15–18 June 2021).*

Manolessou, I., S. Beis & C. Basea-Bezantakou. 2012. *Η φωνητική απόδοση των νεοελληνικών διαλέκτων* [*The phonetic transcription of the modern Greek dialects*]. *Λεξικογραφικόν Δελτίον* [*Lexicographic Bulletin*] 26: 161–222 [in Greek].

Mavrochalyvidis, G. & I. Kesisoglou. 1960. *Το γλωσσικό ιδίωμα της Αξού* [*The Dialect of Axos*]. Athens: Centre for Asia Minor Studies.

Papadopoulos, A. A. 1958. *Ἱστορικὸν λεξικὸν τῆς ποντικῆς διαλέκτου* [*Historical Lexicon of the Pontic Dialect*]. Vol. 1: Α–Λ. Athens: Epitropi Pontiakon Meleton [in Greek].