A uniform data model for joint publication of minor language dictionaries

Martin Haspelmath
2014

1. Introduction

This paper proposes a data model for dictionaries of minor languages that can be applied to any language and that allows publication of many different dictionaries in the same general framework, with identical search options and linking of entries via comparison meanings. Dictionaries of minor languages (i.e. languages which are not taught widely in schools and do not have official support) are of great interest to linguistic scholarship, but publishing such dictionaries in a way that that permits easy use by linguists and speakers requires electronic publication in database form and a relatively uniform data structure.[1]

The basic idea is that it should not be necessary to create a separate digital infrastructure for each new electronic dictionary. While it is possible to create separate infrastructures for commercial dictionaries of big languages such as merriam-webster.com, duden.de and larousse.com, this is not a realistic option for languages that have just a few thousand speakers or less and a handful of linguists working on them. Thus, electronic dictionaries of minor languages must adopt the Wikipedia model of having a single infrastructure for joint publication (e.g. a dictionary journal), and this implies a UNIFORM DATA MODEL.

By *data model* we mean a set of data types and their interrelations, abstracting away from the way in which they are presented visually. Our data model should be able to subsume all the important information found in traditional linear dictionaries, but since it must be uniform and should be relatively simple, the information will not always have the same form as traditional dictionaries. In particular, information that is conveyed by purely linear arrangement must be represented in some different form in our model, because databases do not have a linear structure.

It is immediately apparent that all dictionaries share some key properties: they consist of lists of WORD FORMS, each of which is linked to some additional information about these word forms, in particular a MEANING DESCRIPTION in a metalanguage, a meaning description in the object language itself, GRAMMATICAL INFORMATION such as word-class, gender or inflection class, and so on. Dictionaries also usually contain EXAMPLE SENTENCES that illustrate the use of a word in context. Thus, it should be possible in principle to adopt the same data model for all languages, regardless of language-specific peculiarities of individual languages.

This paper describes such a data model and proposes specific solutions for data modeling issues that dictionary authors may be confronted with. Of course, by using a

[1] See Bell & Bird (2000) for a similar attempt to devise a uniform data model. However, their focus is on formal implementation (in XML) and on capturing everything that existing linear dictionaries can be found to contain. As a result, their model is far more complicated than the model proposed here. The focus here is on a data model for future dictionaries that are created with a view to integrating them into a a larger set of electronic dictionariess.

single data model for all languages, we may sometimes have to adopt a solution that is less than perfect for a particular language. But we think that using different data models will make it far more difficult to store and access the information, so that it is best to accept the limitations of the uniform data model.

2. Data types

Dictionaries are collections of LEXICAL ENTRIES. The most important parts of an entry are the LEMMA and the SENSE. The lemma is the citation form which serves as the name of an entry. In the simplest form, a dictionary is a list of lemma-sense pairs (*arbre*: ‚tree', *maison*: ‚house', *acheter*: ‚buy', etc.).

The most typical lexical entries are SIMPLE LEXEMES such as nouns and verbs, but more complex lexical entries, such as PHRASAL ENTRIES (*take part, point of view,* etc.) and IDIOMS (*all hell broke loose, fall on deaf ears*) must also be treated as lexical entries.

A lexical entry may be associated with one or more senses. While the simplest case of words with a single sense (e.g. *estomac*: ‚stomach') is not uncommon, there are also many cases of words with multiple senses. Thus, the data model needs to accommodate one-to-many correspondences between lemmas and senses, e.g. German *Platz*:

Platz	(i) ‚space'
	(ii) ‚seat'
	(iii) ‚square'

Both lexical entries (§4) and senses (§5) are associated with further categories of information (database fields).

3. Database fields

For concreteness, the categories of information that are attached to lexical entries and senses are called DATABASE FIELDS. The most important question that a dictionary author needs to resolve is which database fields to use. In traditional print dictionaries, the various categories of information are printed in a linear fashion, with no labels for the database fields, and mostly only with typographic distinctions or separation marks such as commas, semicolons or brackets between the fields.

For example, Figure 1 shows an online dictionary of Maung (Australian), where lexical entries have the database fields WORD-CLASS (blue italics, abbreviated) and RELATED ENTRY (green, preceded by *See:*), and each entry can have multiple senses. A sense is a list of one or more SENSE DESCRIPTORS (e.g. the two descriptors „gather together in one place" and „put together" for the single sense of the entry *akpaj*). Each sense may be associated with a SEMANTIC CATEGORY (purple, preceded by *Category:*) as well as with an EXAMPLE (blue).

ajput *nc4.*

1 • sand, beach. *Category:* **Elements.** **Ngana tuka ajput.** I am going to the beach.

2 • sugar. **Ngana nganalakpalwarrki ngartu ta ajput.** I am going to buy myself sugar.

ajputajput *nc4.* along the beaches. **Kawarra ajputajput.** They are going along the beaches (looking for turtle eggs).

akarnpa *nc4.* yam root. *Category:* **Food Drink Cooking and Fire.** **Takapa akarnpa wularr.** Those are yam roots. *See:* **igarnpa.**

akarra *nc6.* many, a lot of. *Category:* **Number and Quantity.** *See:* **wigarra.**

akijalk *nc6.* yam, grain or vegetable fruit or body. *Category:* **Plants.** *See:* **ingijalk.**

akiri *nc6.* skin of some nc6 item, yam, grain, etc., foreskin. *Category:* **Plants.** *See:* **ingiri.**

akpaj *vroot.* gather together in one place, put together. *Gram:* used with tv kinima *Category:* **Holding and Transfer.** **Ararrkpi ying inimang akpaj.** A man might come and put them together.

Figure 1: AuSIL Dictionary of Maung[2]

Thus, for the dictionary in Figure 1, the following database fields are used:

lexical entry:
- – lemma (= citation form) (black, boldface)
- – word-class (= part of speech) (blue, italics, serif font)
- – associated entry (green)

sense:
- (– sense number)
- – list of sense descriptors (black serif font, separated by comma)
- – semantic category (purple)
- – example (dark blue, translation in black serif font)

The list and organization of the database fields that are used by a dictionary is its DATA MODEL. Different dictionaries tend to have slightly different data models, but here a single uniform data model is proposed that should be able to accommodate most of the information that authors want to provide in dictionaries of minor languages.

Some fields are applicable to any language in principle (GENERAL FIELDS), while other fields may be language-specific (SPECIFIC FIELDS). The fields used in the Maung dictionary in Figure 1 are all general fields, i.e. they could be used also in a dictionary of any other language. But dictionaries may also contain specific fields that are useful only for the language in question, e.g. information on inflectional class membership.

[2] http://ausil.org/Dictionary/Maung/lexicon/index.htm (a dictionary created with the programme Lexique Pro)

4. Database fields for lexical entries

After laying out the main concepts, the following sections give the uniform data model that is proposed here.

4.1. General fields

The following database fields of lexical entries are immediately clear and do not raise any questions. For dictionaries of minor languages, lemma and part-of-speech are generally considered essential and are present in all dictionaries (hence in boldface below).

– **lemma = headword = citation form**[3]
– lemma in original script (if the language is commonly written in another script)
– pronunciation of lemma (in IPA)
– **part-of-speech**
– general comments
– bibliographical reference

In addition, each entry is obligatorily associated with one or more senses (see §5 for data fields of senses).

Lexical entries may also contain historical information, especially on loanwords, but also other etymological information:

– source language (for loanwords)
– source word (for loanwords) plus gloss
– etymology

It should be noted that lexical entries need not be simple words, but can be idioms of various sizes (e.g. *hot dog, pay lipservice, once and for all, be that as it may*). Such complex entries can be associated with their parts via association fields (§6).

4.2. Specific fields

In addition to the general fields, dictionaries of particular languages may require other kinds of fields, in particular fields for grammatical information:

– verbal valency
– gender
– inflectional class[4]

---

[3] Several different entries may have the same lemma (such cases are called HOMONYMS). For easier distinguishability, they may be accompanied by a homonym number that is part of the lemma (e.g. *bank 1*, *bank 2*)

[4] Inflectional information is often given in the form of specific inflected forms, e.g. unpredictable plural or past-tense forms (called „principle parts“ if these can be used to predict further inflected forms). Such specific forms are association fields, see §6 below.

Other specific fields may concern sociolinguistic information, e.g. concerning dialects, registers, politeness, taboo, speech level (plain vs. honorific), about literary vs. colloquial usage.

These fields cannot be specified in advance, because they may depend on language-specific circumstances. There is thus no way around language-specific fields for lexical entries.

5. Database fields for senses

The most important field for senses is of course the sense description (= definition). As mentioned earlier, a lexical entry can have several different senses, which are often numbered (e.g. in the Maung entry *ajput* in Figure 1).[5]

A sense can be a single sense descriptor, or can consist of a list of sense descriptors, when it is described by multiple words or phrases (e.g. the first sense of *ajput* is characterized by two sense descriptors as ‚sand, beach').

There is no hard and fast rule for when to set up a new sense, as opposed to describing a single sense with multiple descriptors. In general, multiple senses are used to represent POLYSEMY, i.e. a situation where a lexical form has a number of clearly distinct senses which cannot be readily subsumed under a single meaning from which the senses can be contextually derived. Thus, while both squares and seats are kinds of spaces, it is not possible to predict that German *Platz* ‚space, seat, square' (mentioned above in §2) should have precisely these three senses (and not, for instance, the sense ‚apartment', even though an apartment is also an important kind of space, cf. English *place*, which can have this sense). Deciding between these two kinds of representations is not straightforward at all in practice, but what is important for our data model is that both ways of representing a word's meaning are often used by lexicographers and hence need to be recognized by the data model.

In addition to the sense description in the metalanguage (English), senses can be given in the object language itself (as in monolingual dictionaries), and in some other metalanguage that is relevant for the minor language in question (most likely a national language such as Spanish or Russian). Many dictionaries also give a semantic domain.

– **sense description (English)**
– sense description in object language („native definition")
– sense description in another relevant metalanguage (e.g. Spanish, Russian)
– semantic domain
– comment

Finally but importantly, each sense may be associated with an EXAMPLE. Examples are particularly helpful when a word has multiple senses, and dictionaries tend to give

[5] There is no hard and fast rule for when to set up homonyms and when to assume a single entry with multiple senses. In general, homonyms are set up when the grammatical information is different and/or when the senses are not related at all. Given our data model, when the grammatical information is different, it is necessary to set up a new entry (see §7).

examples especially for such polysemous words. An example may be a clause or a phrase (e.g. a verb with an object, or a complex noun phrase).

Examples minimally consist of the primary text and the free translation, but ideally they should also have an interlinear morpheme-by-morpheme translation.[6] There may be multiple examples for each sense, and an example may illustrate multiple lexical entries, so there is a many-to-many relation between examples and senses. Examples thus need to be in a separate database table.

It appears that the data fields for senses are all general, i.e. there are no language-specific fields here.

6. Association fields

One important type of information for many lexical entries (and senses) is association with some other lexical entry. For example, we may want to say that a lexical entry is part of, or contains, or is derived from, some other lexical entry. For example:

(lexical entry:)	*cloth*
– part of (lexical entry):	*table cloth*

(lexical entry:)	*tablecloth*
– contains (lexical entries):	*table, cloth*

(lexical entry:)	*hot dog*
– contains (lexical entries):	*hot, dog*

(lexical entry:)	*amusement*
– derived from (lexical entry):	*amuse*

For senses, we may say that they are synonymous with other senses, or antonyms, or hyponyms, etc.

(sense:)	‚stop' (a sense of the entry *cease*)
– synonym(ous with):	‚stop' (a sense of the entry *stop*)

Association fields are also used to represent unpredictable inflectional information, or information about associated classifiers etc.

(lexical entry:)	*mouse*
– plural form:	*mice*

[6] It is mainly for typographic reasons that examples in dictionaries of minor languages tend not to have interlinear translations, whereas this is now completely normal for grammars of minor languages. But they are of course no less necessary in dictionaries, because readers cannot be assumed to know the language well enough to understand the example without such interlinear glosses.

(lexical entry:)	*mǎ* ‚horse' (Mandarin Chinese)
– numeral classifier:	*pǐ*

Some associated entries such as inflected forms like *mice* need not be given a full lexical entry. *Mice* would be an ASSOCIATED ENTRY, which does not have word-class and sense information, but only has the association field „plural form of", which links it to the regular entry *mouse*.

In electronic dictionaries with hypertext links, such associations can simply be implemented by links to the corresponding lexical entries (and senses) via the lemmas.

7. One-to-many relationships

As was noted in §2, a lexical entry may be associated with multiple senses, as is again illustrated below.

lexical entry:	*macinare* (Italian)
sense:	grind
sense:	spend (a lot of money)
sense:	eat (a lot of food)

And of course a sense can be associated with multiple examples. Moreover, examples may of course be associated with multiple senses of different entries, illustrating simultaneously the uses of several words.

Multiple senses of an entry must occur in some linear order, and this linear order is sometimes given some significance in traditional dictionaries. For instance, Hausa *ɗari* has the two senses ‚1. dry cold' and ‚2. chills due to illness' (Newman's Hausa dictionary, see Figure 3 in §9.x below). Presumably the ‚dry cold' meaning is given first here because the other meaning represents some kind of semantic extension. In order to capture the significance of linear order, our data model would have to add a rank field to each sense. It is not clear whether this is necessary, because in many cases, senses seem to be linearized (and numbered) without implying any rank (primary vs. secondary meaning).

The association relations described in §6 may also be one-to-many. For example, a lexical entry can be said to comprise several other lexical entries (as in the case of *tablecloth*, which comprises *table* and *cloth*).

However, other one-to-many (or many-to-many) relationships need not be recognized. For example, it is not necessary to allow entries with two different parts of speech, such as:

lexical entry:	*stop* (English)
part-of-speech:	verb
part-of-speech:	noun

lexical entry:	*lingüista* (Spanish)
gender:	masculine
gender:	feminine

Linguists might want to posit precisely such one-to-many relationships, but such entries would lead to too much complexity in the data model, which is rarely needed for dictionaries of minor languages. If there are two different parts-of-speech, then these two are presumably also associated with two different senses, so that one would have to express that the lemma has sense 1 if it has part-of-speech 1, and sense 2 if it has part-of-speech 2, and so on. This would effectively mean the creation of a subentry, but we do not allow subentries, as discussed in §8. In the case of nouns with two genders, an alternative (if one wants to avoid multiple lexical entries) might be to set up a new gender category „epicene".

8. No subentries

Our data model does not allow for subentries. Traditional dictionaries often have a structure such as:

put
  put aside
  put off
  put up

That is, the formally complex entries *put aside*, *put off* and *put up* are subentries of the formally simple entry *put*. This makes sense in linearized, paper-based dictionaries, where searching is done by alphabetic scanning.

In electronic dictionaries, this does not make much sense, because complex entries otherwise have the same properties as simple entries. The important piece of information that is traditionally expressed by the relationship between main entry and subentry is the part-of and the contains relationship:

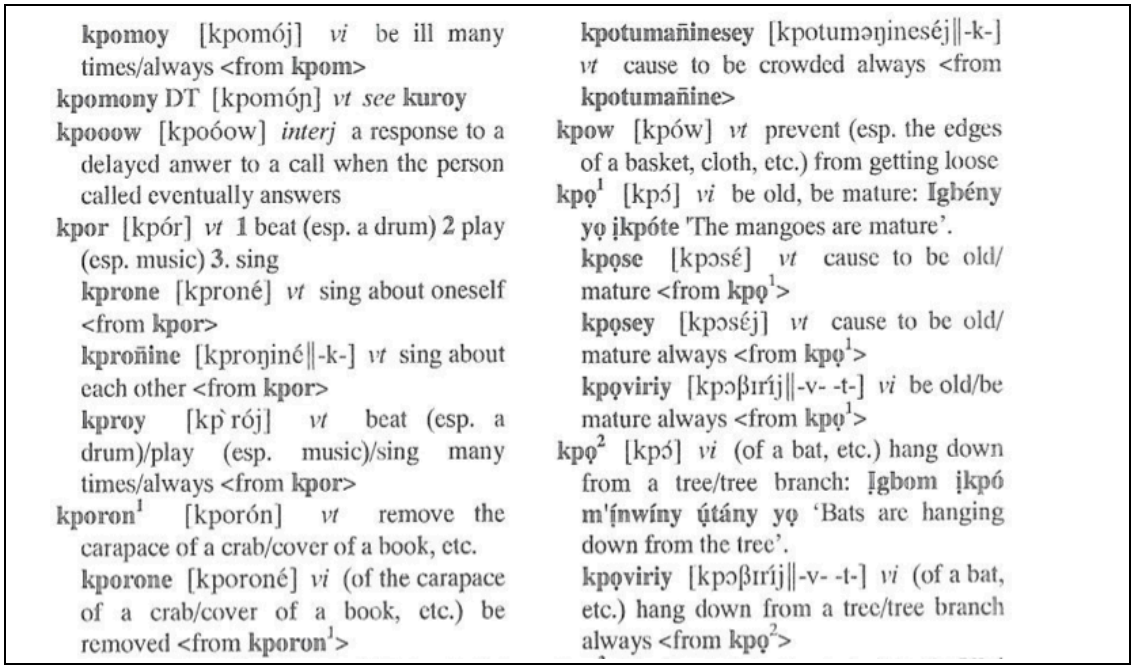
entry: *put*
– part of: *put aside*

entry: *put aside*
– contains: *put, aside*

9. Some example dictionaries and how our data model applies to them

9.1. Kari's Degema dictionary

In Ethelbert Kari's (2008) dictionary of Degema, each entry is followed by the pronunciation in IPA and the word-class, plus one or more numbered senses (three senses occur with *kpor*, for example), and an example.



kpomoy [kpomój] *vi* be ill many times/always <from **kpom**>
kpomony DT [kpomǫ́n] *vt see* **kuroy**
kpooow [kpoóow] *interj* a response to a delayed anwer to a call when the person called eventually answers
kpor [kpór] *vt* **1** beat (esp. a drum) **2** play (esp. music) **3.** sing
kprone [kproné] *vt* sing about oneself <from **kpor**>
kproñine [kproŋiné‖-k-] *vt* sing about each other <from **kpor**>
kproy [kp᷄rój] *vt* beat (esp. a drum)/play (esp. music)/sing many times/always <from **kpor**>
kporon[1] [kporón] *vt* remove the carapace of a crab/cover of a book, etc.
kporone [kporoné] *vi* (of the carapace of a crab/cover of a book, etc.) be removed <from **kporon**[1]>
kpotumañinesey [kpotumǝŋineséj‖-k-] *vt* cause to be crowded always <from **kpotumañine**>
kpow [kpów] *vt* prevent (esp. the edges of a basket, cloth, etc.) from getting loose
kpǫ[1] [kpɔ́] *vi* be old, be mature: **Igbény yǫ ịkpóte** 'The mangoes are mature'.
kpǫse [kpɔsé] *vt* cause to be old/ mature <from **kpǫ**[1]>
kpǫsey [kpɔséj] *vt* cause to be old/ mature always <from **kpǫ**[1]>
kpǫviriy [kpɔβɪríj‖-v- -t-] *vi* be old/be mature always <from **kpǫ**[1]>
kpǫ[2] [kpɔ́] *vi* (of a bat, etc.) hang down from a tree/tree branch: **Ịgbom ịkpó m'ịnwíny ụ́tány yǫ** 'Bats are hanging down from the tree'.
kpǫviriy [kpɔβɪríj‖-v- -t-] *vi* (of a bat, etc.) hang down from a tree/tree branch always <from **kpǫ**[2]>

Figure 2: An excerpt from Kari (2008) (A dictionary of Degema)

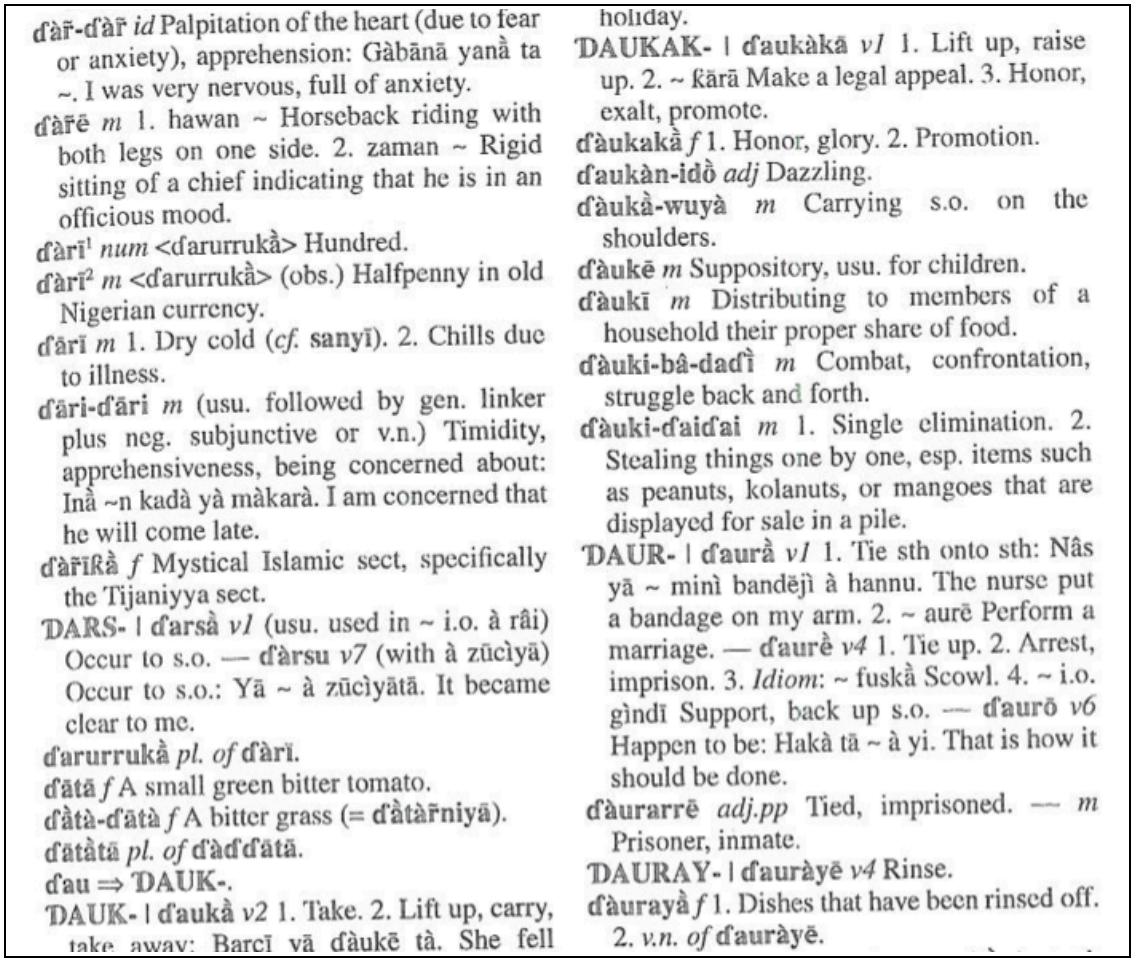
What is notable in Figure 2 is that many entries have subentries for morphologically related items, e.g. *kpǫ 1* (‚be old'), which has the subentries *kpǫse* (‚cause to be old'), *kpǫsey* ‚cause to be old always', and *kpǫviriy* ‚be old always'. Since our model does not allow subentries (§8), these would have to be separate entries, but they would be linked to the entry *kpǫ 1* by association fields (§6), e.g.

(lexical entry:)	*kpǫsey*	‚cause to be old always'
– contains root:	*kpǫ 1*	‚be old'
(lexical entry:)	*kpǫ 1*	‚be old'
– habitual form:	*kpǫviriy*	‚be old always'

9.2. Newman's Hausa dictionary

In Paul Newman's (2007) dictionary of Hausa (see Figure 3), most entries have a word-class (in italics: *m, f, id, num, v1, v2*, etc), and quite a few have different senses, especially the verbs. Examples of multiple sense descriptors for a single sense are ‚honor, glory' (one sense of the entry *dàukaka*), and ‚lift up, carry, take away' (one sense of *dauka*). Quite a few of the entries have example phrases or clauses. Some of the nouns have unpredictable plural forms given in angle brackets. Occasionally there are sociolinguistic annotations such as „(obs)" (obsolete) for *dari*[2].

Since Hausa verbs typically come in „grades" of semantically related verbs derived from the same root, all verbs are treated as subentries of verb roots in the dictionary. Thus, the verbs *daura* ‚tie' (grade 1), *daure* ‚tie up' (grade 4), and *dauro* ‚happen to be' (grade 6) are all found under the entry *DAUR-*. In our datamodel, which does not allow subentries, the verb root would be an associated entry (like English mice, cf. §6), and it would be linked to the individual verbs (and vice versa) via association fields.



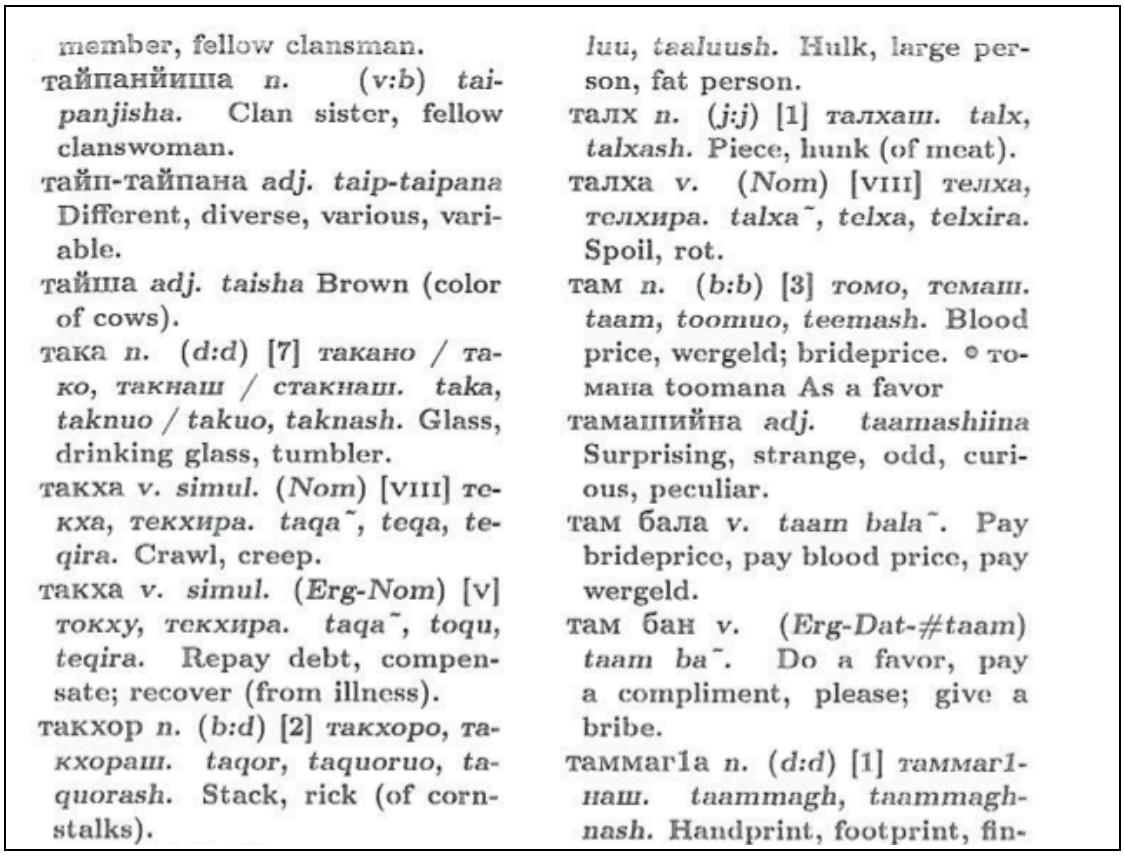
ɗàř-ɗàř *id* Palpitation of the heart (due to fear or anxiety), apprehension: Gàbānā yanå ta ~. I was very nervous, full of anxiety.
ɗàřē *m* 1. hawan ~ Horseback riding with both legs on one side. 2. zaman ~ Rigid sitting of a chief indicating that he is in an officious mood.
ɗàrī[1] *num* <ɗarurrukå> Hundred.
ɗàrī[2] *m* <ɗarurrukå> (obs.) Halfpenny in old Nigerian currency.
ɗārī *m* 1. Dry cold (*cf.* sanyī). 2. Chills due to illness.
ɗāri-ɗāri *m* (usu. followed by gen. linker plus neg. subjunctive or v.n.) Timidity, apprehensiveness, being concerned about: Inå ~n kadà yà màkarà. I am concerned that he will come late.
ɗàřīßå *f* Mystical Islamic sect, specifically the Tijaniyya sect.
DARS- | ɗarså *v1* (usu. used in ~ i.o. à râi) Occur to s.o. — ɗàrsu *v7* (with à zūcìyā) Occur to s.o.: Yā ~ à zūcìyātā. It became clear to me.
ɗarurrukå *pl. of* ɗàrī.
ɗātā *f* A small green bitter tomato.
ɗàtà-ɗātà *f* A bitter grass (= ɗàtàřniyā).
ɗātàtā *pl. of* ɗàɗɗātā.
ɗau ⇒ DAUK-.
DAUK- | ɗaukå *v2* 1. Take. 2. Lift up, carry, take away: Barcī vā ɗàukē tà. She fell holiday.
DAUKAK- | ɗaukàkā *v1* 1. Lift up, raise up. 2. ~ ƙārā Make a legal appeal. 3. Honor, exalt, promote.
ɗàukakå *f* 1. Honor, glory. 2. Promotion.
ɗaukàn-idò *adj* Dazzling.
ɗàukà-wuyà *m* Carrying s.o. on the shoulders.
ɗàukē *m* Suppository, usu. for children.
ɗàukī *m* Distributing to members of a household their proper share of food.
ɗàuki-bâ-daɗì *m* Combat, confrontation, struggle back and forth.
ɗàuki-ɗaiɗai *m* 1. Single elimination. 2. Stealing things one by one, esp. items such as peanuts, kolanuts, or mangoes that are displayed for sale in a pile.
DAUR- | ɗaurå *v1* 1. Tie sth onto sth: Nâs yā ~ minì bandējì à hannu. The nurse put a bandage on my arm. 2. ~ aurē Perform a marriage. — ɗaurè *v4* 1. Tie up. 2. Arrest, imprison. 3. *Idiom*: ~ fuskå Scowl. 4. ~ i.o. gìndī Support, back up s.o. — ɗaurō *v6* Happen to be: Hakà tā ~ à yi. That is how it should be done.
ɗàurarrē *adj.pp* Tied, imprisoned. — *m* Prisoner, inmate.
DAURAY- | ɗauràyē *v4* Rinse.
ɗàurayå *f* 1. Dishes that have been rinsed off. 2. *v.n. of* ɗauràyē.

Figure 3: An excerpt from Newman's (2007) (A dictionary of Hausa)

A few further observations: (i) One entry (*ɗari-ɗari*) has a miscellaneous grammatical comment (in parentheses) that should probably be in a general comments field. (ii) The entry *ɗàre* does not have a sense, but is associated with two (what appears to be) fixed phrases, *hawan ɗare* and *zaman ɗare*. Since their meaning (apparently) cannot be compositionally derived, they are idioms and hence they should get their own entries, linked to the association entry *ɗàre*. (iii) The dictionary treats some homonyms as senses, e.g. *ɗàuraya* is said to have two senses: ‚dishes that have been rinsed off', and ‚verbal noun of *ɗauraye*'. The latter should be an association entry in our datamodel, linked to the verb *ɗauraye* ‚rinse', so it cannot be another sense of the same entry.

9.3. Nichols's Chechen dictionary

Johanna Nichols's Chechen dictionary gives all Chechen words both in the Cyrillic spelling and in the pronunciation (using a non-IPA transcription). Nouns have gender information in round brackets (*v:b, j:j, b:b* etc.), declension-class information in square brackets, as well as the principal parts ergative singular and nominative plural (e.g. *taam, toomuo, teemash*). Verbs have conjugation class information in square brackets and valency information in round brackets. In addition, principal parts are given for many verbs (present tense, witnessed past tense, e.g. *taqa˜, teqa, teqira*).



member, fellow clansman.
тайпанийша *n.* (*v:b*) *taipanjisha.* Clan sister, fellow clanswoman.
тайп-тайпана *adj. taip-taipana* Different, diverse, various, variable.
тайша *adj. taisha* Brown (color of cows).
така *n.* (*d:d*) [7] *такано / тако, такнаш / стакнаш. taka, taknuo / takuo, taknash.* Glass, drinking glass, tumbler.
такха *v. simul.* (*Nom*) [VIII] *текха, текхира. taqa˜, teqa, teqira.* Crawl, creep.
такха *v. simul.* (*Erg-Nom*) [V] *токху, текхира. taqa˜, toqu, teqira.* Repay debt, compensate; recover (from illness).
такхор *n.* (*b:d*) [2] *такхоро, такхораш. taqor, taquoruo, taquorash.* Stack, rick (of cornstalks).
luu, taaluush. Hulk, large person, fat person.
талх *n.* (*j:j*) [1] *талхаш. talx, talxash.* Piece, hunk (of meat).
талха *v.* (*Nom*) [VIII] *телха, телхира. talxa˜, telxa, telxira.* Spoil, rot.
там *n.* (*b:b*) [3] *томо, темаш. taam, toomuo, teemash.* Blood price, wergeld; brideprice. ◦ томана toomana As a favor
тамашийна *adj. taamashiina* Surprising, strange, odd, curious, peculiar.
там бала *v. taam bala˜.* Pay brideprice, pay blood price, pay wergeld.
там бан *v.* (*Erg-Dat-#taam*) *taam ba˜.* Do a favor, pay a compliment, please; give a bribe.
таммаг1а *n.* (*d:d*) [1] *таммаг1наш. taammagh, taammaghnash.* Handprint, footprint, fin-

Figure 4: An excerpt from Nichols (2010) (A dictionary of Chechen)

Different senses are separated by semicolons (e.g. *taam* ‚blood price, wergeld; brideprice'). Fixed phrases are given their own entries, as in our datamodel (e.g. *taam bala˜* ‚pay brideprice').

The main challenge of this dictionary for our datamodel is that there are multiple cases where forms are given both in transcription and in Cyrillic orthography. This means that we need both a primary field and a Cyrillic field for a number of categories of information: lemma, ergative of nouns, plural of nouns, present tense of verbs, witnessed past of verbs. The alternative would be to put the orthography in the same field as the transcription, but this would mean that the dictionary cannot be sorted by the orthography.

Reference

Bell, John & Bird, Steven. 2000. A preliminary study of the structure of lexicon entries. Paper presented at the Workshop on Web-based Language Documentation and Description, 12-15 December 2000, Philadelphia (USA).