Dictionaries of under-researched languages

Ulrike Mosel

Kiel University

Abstract: This chapter compares under-researched language dictionaries (ULDs) with commercial dictionaries of major European languages with respect to the selection of lemmas (headwords), sense division, the types of meaning description, and the source, form and content of examples. Although ULDs only cover a limited range of lexemes and their senses, they can be a valuable source for educational materials and for linguistic and anthropological research, if the lexicographers follow a corpus-based approach in collaboration with native speakers. Since ULDs are not commercial products, their lexicographers are free to develop new types of dictionary and give up the strict distinction between monolingual and bilingual dictionaries and between dictionaries and encyclopedias.

Keywords: dictionaries, lexicography, under-researched languages, language documentation

1. Introduction

Dictionaries of under-researched languages (ULDs) are typically bilingual dictionaries with the under-researched language as the source language (SL) and an international or a national language as the target language (TL). At first sight they may look like concise commercial dictionaries of European languages, but a closer look reveals that there are profound differences that the compilers and the users of ULDs should be aware of. This article explores these differences on the basis of metalexicographical handbooks (Atkins & Rundell 2008, Durkin (ed.) 2016, Jackson (ed.) 2013, Svensón 2009), the study of several unpublished lexical databases and dictionaries of unrelated non-European languages, my own experiences in compiling thematic dictionaries for the Polynesian language Samoan (Mosel & Fulu 1997) and the Meso-Melanesian language Teop spoken in Bougainville, Papua New Guinea (Mosel 2010, 2012b, 2014a, 2014b), and my current work of transforming the Teop lexical database into a general dictionary.

The following sections of this introduction will first briefly describe the distinctive features of ULDs with respect to their production circumstances and, consequently, their content (§2.1), the strategies of data collection (§2.2), and the content of lexical databases (§2.3) from which one or more dictionaries can be derived (§2.4). The next sections discuss the selection of lemmas (also called headwords) in §3, the description of meanings in §4, and the selection of examples in §5. The last section (§6) suggests to inform the user about the semantic fields covered by the dictionary. If the senses of the lemmas are classified into semantic fields, electronic ULDs can facilitate the extraction of the entries of specific semantic fields for further research or the compilation of thematic dictionaries (Coward & Grimes 2000:177, Mosel 2011).

The article does not deal with dictionary making tools, the design of electronic dictionaries, and the inclusion of multimedia files. For the development of multimedia lexical databases and dictionaries of ULDs see Cablitz 2011a, 2011b, Hyslop Malau 2011, Thieberger 2011, for e-lexicography in general see Fuertes-Olivera & Bergenholtz (eds.) 2011 and Granger & Paquot 2012. A survey of ULDs of Australian and South Pacific languages is presented by Thieberger (2015).

2. ULDs and commercial dictionaries

2.1. Distinctive features of ULDs

The form and content of commercial dictionaries "is largely determined by the types of linguistic activities it is meant to be used for, by the needs of the users in this context and by the capabilities of the users" (Svensén 2008:5). Similarly, Tarp (2008:43) says:

> "... dictionaries are objects of use which are produced or should be produced to satisfy specific types of social need. These needs are not abstract - they are linked to specific types of user in specific types of social situation."

In the case of ULDs, however, the form and content is not only determined by their purpose, but also to a considerable degree by the production circumstances. In contrast to commercial dictionaries, ULDs are typically produced in a rather short time with a small team of non-professional, part-time working lexicographers who at the same time compile a corpus and do some other kind of linguistic or anthropological research. Since the corpus which serves as the source of the lexical database consists of texts of only a few themes, genres and registers, and contains much less than a million of words, ULDs cover only a small proportion of the lexemes of the language and their senses and uses. However, whether an ULD qualifies as a useful object or does not meet the expectations of its users, does not depend on its size, but on the quality of the information it offers and how this information can be accessed by searching and filtering options (see §6).

ULDs do not fit into the typology of commercial dictionaries, which roughly may be classified by the following features (for more detailed classifications see Adamska-Sałaciak 2013:214f., Atkins & Rundell (2008:17-44), Béjoint (2000:37-41), and Svensón (2009:12-38)).

1. monolingual dictionaries for native speakers or language learners;
2. passive bilingual dictionaries, i.e. dictionaries for decoding texts in a non-native language (L2) and translating them into the user's native language (L1) or active bilingual dictionaries, i.e. dictionaries for encoding texts in a non-native language (L2);
3. general and specialised dictionaries;

4. dictionaries without and with encylopedic information;
5. dictionaries for adults or children.

When these criteria are applied to ULDs, their special features become immediately obvious. ULDs are, as already mentioned, bilingual dictionaries, but whereas the users of commercial dictionaries are either L1 speakers of the SL or the TL, the users of ULDs may be neither L1 speakers of the SL nor of the TL, but semi-speakers or learners of the SL and L2 speakers of the TL.

If the dictionary is meant to account for these special user types, a meaning description that only uses translation equivalents may not be sufficient, as for example in the case of plant and fish names:

(1) Teop

agevana	pink siris, albizza species
asiata	spangled emperor

For the Teop dictionary project we therefore gave up the rigorous distinction between mono- and bilingual dictionaries, as we do not only provide English translation equivalents, but also SL meaning explanations that are translated into English. This strategy is not only applied to plant and animal names, but also to names of artefacts with English translation equivalents that are presumably unknown to the user, e.g. *vatetekoio* 'upper ridgepole', and for lexical items denoting any kind of culture specific phenomena that lack expressions and referents in English speaking cultures. Giving the explanation both in the SL and the TL documents the conceptualisation of native speaker authors (Mosel 2011) and avoids misinterpretations caused by the lexical anisomorphism between the SL and the TL (Adamska-Sałaciak 2016). This strategy has also been adopted by Cablitz (2011a, b) for the Marquesan dictionary project and The Savosavo Documentation Project (Wegener et al. 2014). For further comments on the meaning description in ULDs see §4. A detailed general treatment of the problem of equivalents in commercial bilingual dictionaries is found in Svensón (2009:253-280); for the inclusion of encyclopedic information see §5.

2.2. Strategies of data collection

The work on a dictionary starts with the collection of words. There are several methods of collecting words, which can be combined:

1. translating a wordlist of the dominant language into the under-researched language;
2. compiling wordlists for selected semantic fields with native speakers;
3. extracting words from a corpus.

Moe (2003, 2007) developed the Rapid Word Collection method (http://rapidwords.net/) which is based on semantic fields and promises:

> “Rather than the default language worker’s process of collecting words over a period of years and then publishing a work containing 5,000 words or so, RWC workshops consistently achieve a total of 10,000 or more raw entries during a brief two-week period.”

But even with translations 10 000 rapidly collected words are far away from what could serve as a lexical database for the creation of a dictionary, because word lists do not say anything about how the words are used in context and their meaning because "there is no meaning of a lexical item apart from some context, linguistic or cultural" (Nida 1997). Consequently, it makes only sense to collect words in context. Just as for this very reason corpus-based dictionaries have become the standard in commercial lexicography (Hanks 2012, Kilgariff 2013, Kozem 2016, Moon 2010), dictionaries of under-researched languages must be based of corpora.

In the beginning of the documentation and analysis of a previously unresearched language the contexts will be phrases and simple clauses, later stories, descriptions and conversations from which the words are extracted and listed as lemmas in the lexical database. Words, phrases and simple clauses should be elicited by methods that avoid translations from the TL into the SL and that always provide some kind of context (Chelliah & Reuse 2011:227-249, Majid 2012, Mosel 2006, Mosel 2012a, Pawley 2011:267). Under certain conditions, collections of isolated sentences may be used as examples in ULDs (cf. §5.3).

2.3. The lexical database

Language documentation projects focus on the collection of transcribed and translated audio and video recordings, grammaticography or specialised linguistic and anthropological research, but usually not on lexicography. An exception is the Marquesan dictionary project (Cablitz 2011a, b). Otherwise a lexical database often merely serves as a tool for translation and glossing and, consequently, may be quite different from a dictionary:

1. the lemmas and their senses have not undegone a thorough semantic analysis yet;
2. the orthography hasn't been standardised;
3. the selection of lemmas has not considered multi-word lexemes (see §3.4);
4. the sense division is more or less intuitively guided by the TL translation equivalents (see §4);
5. the quotations from the corpus may not be suitable as illustrative examples (see §5.4).

There may be hundreds of gaps, errors, and inconsistencies because the lexical database was not meant to be published. Since it only served as a tool for the researchers, it has not been carefully reviewed with respect to the needs of potential other users, be it other researchers or members of the speech community.

2.4. Creating a dictionary

The above mentioned characteristics of lexical databases imply that the creation of a dictionary requires more than a few weeks' work (Pawley 2011:267-268, Cablitz 2011b, Lichtenberk 2008:2-3).

Since the language is under-researched, an ideal ULD does not just present a wordlist with parts of speech labels and glossings, but

1. includes multi-word expressions in the lemma list, whenever they seem to be common collocations or lexicalisations (cf. §3.4);
2. gives spelling variants in case that the spelling has not been standardised;
3. informs on derivational morphology by including derivational affixes in the lemma list;
4. provides authentic examples that cover as many types of construction as possible (§5.2) and gives not only free translations, but also, wherever necessary for the understanding of the construction, a literal translation (cf. §5.3);
5. gives references to the source of all examples (cf. §5.1);
6. gives cross-references to other entries with semantically related lexical units such as antonyms, homonyms, hyperonyms, etc.

3. Lemma

3.1. A definition of lemma

The lemma is an orthographical word or a sequence of orthographical words that functions as the heading of a lexical entry, e.g. *go*, *went*, *I*, *me*, *left*, *UFO*, *I'm*, *'m, e-* (as in *e*-store), *-ed, light-hearted*, *light year* in the Collins Cobuild Advanced Dictionary (CCAD). As the preceding examples show, the notion of lemma is distinct from the notion of lexeme, as not only lexemes, but also irregularly inflected wordforms, abbreviations, clitics and affixes may figure as lemmas. Furthermore, a lemma may head an entry with several sublemmas each of which is representing a distinct lexeme, e.g. the lemma *quick* and the sublemma *quickly* in CCAD.

3.2. Subentries

In print dictionaries the lexical entry may contain subentries each of which is introduced by a sublemma. The main lemma *light* (adjective) is, for instance, followed by the sublemmas *lightly* (adverb) and *lightness* (noun) in the CCAD. The subentries save space on paper, but since in electronic dictionaries there is no need of saving space, subentries are avoided as they would make the structure of the database unnecessarily complex.

The avoidance of subentries implies that in non-inflecting languages lexical units that have the same form and related meanings, but are assigned to distinct parts-of-speech, function as lemmas of separate entries. A typical example is the treatment the Teop wordform *kapa*:

(2) Teop

kapa$_1$	n. skin of animals and root crops / shell of turtles, shellfish and nuts / bark of trees
kapa$_2$	v.t. remove the skin / shell / bark of something

The distinction of the parts of speech of SL lexemes must not be influenced by their translation equivalents in the TL, but must exclusively be based on the grammatical characteristics of the SL lexemes. Since *kapa* ‘skin / shell / bark’ and *kapa* ‘remove skin / shell / bark’ differ with respect to their modifiers, the distinction between *kapa*$_1$ and *kapa*$_2$ and their classification as a noun and a verb is justified.

3.3. The lemma form in dictionaries of inflecting languages

Dictionaries of inflecting languages use different grammatical forms as lemma forms, e.g. for verbs the infinitive (French, German), the form of the 1st person singular present indicative active (Latin, Classical Greek), or the root (Biblical Hebrew). In ULDs the selection of the lemma form should take the native speakers’ preferences into consideration. When writing the dictionary of San Lucas Quiavini Zatopec, a prefixing Amerindian language, Munro (2002:98-99) initially suggested to use an unprefixed bare stem form as the lemma form for verbs, but as such unpronouncable lemmas were not accepted by her collaborator who was a native speaker, they chose an inflected form with the prefix *r-* . This had the consequence that all verbal lemmas start with *r-*.

3.4. Multi-word lemmas

In dictionaries of well researched languages, “the basic prerequisite for according lemma status to a multi-word item is that it has undergone some kind of lexicalisation, i.e. that it has been stored in our mental lexicon as a unit” (Svensén 2009:102f). For the authors of a ULD it may, however, be difficult to decide whether a collocation has been lexicalised or not. Since space does not matter in the electronic Teop-English dictionary, I decided to include multi-word expressions as lemmas that occur more than once in the corpus and have a meaning that from the perspective of the TL is unpredictable and may be interesting for later research:

(3) Teop

ani kahi	lit. ‘eat from’	‘leave (food) uneaten ’
mate kahi	lit. ‘die from’	‘die and leave someone behind’

Furthermore, Teop determinative compounds that correspond to single-word expressions are used as lemmas although their meaning is fully predictable (4):

(4) Teop

hum aniani	lit. ‘eating place’	‘restaurant’
hum komkom	lit. ‘stepping place’	‘steps (in a house)’
hum vaavon	lit. ‘selling place’	‘shop’

For a brief overview of corpus-based studies of English multi-word units see Greaves & Warren 2010.

4. The meaning of lexical items

4.1. Forms of meaning descriptions

In bilingual dictionaries the meanings of lexical items can be given by simple words, compounds, phrases, clauses, sentences, or even longer descriptions. The selection of the one or the other kind of meaning description is determined by the degree of formal, semantic and pragmatic isomorphism between the SL and TL lexicons and by the purpose of the dictionary.

4.2. Translation equivalents and explanations

With the exception of internationally defined terminologies, the meaning and usage of TL words that are used as translation equivalents rarely fully match their counterparts in the SL so that the meaning of a SL word cannot be translated by the same TL word in all contexts as shown in (2). Furthermore, the TL may even totally lack a translation equivalent of the SL word.

The more distant the culture of the SL speakers is from that of the TL speakers, the more gaps and not fully matching translation equivalents will be found, because the TL culture either lacks the respective referent or its speakers conceptualise the referent in a different way. In both cases the meaning needs to be described by multi-word expressions as in the following Teop-English example. The Teop wordform *koopu* represents a noun with three senses and a verb that is derived by conversion from the noun in its second sense.

(5) Teop
koopu$_1$ n. 1. a bamboo species
2. container made of bamboo for food or water
3. knife made of bamboo
koopu$_2$ v. put something into a bamboo container

If the compilers of the Teop-English dictionary were professional lexicographers, they would probably have known the English or the scientific name of the bamboo species called *koopu* and used it as a translation equivalent instead of the explanation ‘a bamboo species’. But the second and the third sense can only be rendered by the compounds ‘bamboo container’ and ‘bamboo knife’ or by explanations. We preferred to use the explanations given in (5), because the compounds ‘bamboo container’ and ‘bamboo knife’ can be wrongly interpreted as ‘container for bamboo’ and ‘knife for cutting bamboo’, respectively.

In the translations of texts the vague compounds ‘bamboo container’ and ‘bamboo knife’ are sufficient, if misinterpretations are excluded by the context. Consequently, it can be argued that the selection of a vague compound or an encyclopedic explanation is a matter of the purpose of the dictionary. For rapid translations, the users will prefer compounds, whereas users who want to know how the culture of the speech community is encoded in the lexicon are better served with explanatory meaning descriptions.

For a further discussion about the challenges of documenting cultural and lexical knowledge see Haviland 2006 and Cablitz 2011a, b.

4.3. Sense division

In contrast to monolingual dictionaries, the sense division in bilingual dictionaries does not necessarily reflect the polysemy of SL words, but “indeed, many bilingual dictionaries divide the semantic space of source lexical items as a function of the target language” (Fontenelle 2016:46) so that a monosemous SL word may look like a polysemous word because it has different translation equivalents in the TL, e.g.

(6) German *Ruder*, English 1. *rudder*, 2. *oar* (Svensén 2009:272)

On the other hand, a polysemous SL word looks monosemous, because it has exactly the same kind of polysemy as its TL counterpart and therefore is only given a single translation equivalent.

(7) German *liquidieren.* English *liquidate* (Svensén 2009:279)

The reason for this TL-oriented sense division is of a practical nature. It helps the user to quickly find a suitable translation equivalent and saves space (Svensén 2009:278-279, Adamska-Sałaciak 2013:222-226, Lev 2013:289-291).

If the purpose of the ULD is to document the semantics and the usage of lexical items rather than to serve as a tool for rapid translation, the division of senses should be based on the SL and not be influenced by translation equivalents. As evidence for polysemy and, consequently, as a justification of sense division one counts distinct grammatical and collocational features. Thus the Teop lexeme *beera* '1. big; 2. important' is split into two lexical units, because it is only *beera* in its second meaning 'important' that can be coordinated with *rutaa* 'small':

(8) Teop (Aro_08E.032)

o	*toro mohina*	*o*	*rutaa*,	*evehee*	*o*	*beera*
ART	island	ART	small	but	ART	important

'the island (is) small, but important'

As long as such evidence is missing in the corpus, a sense division would - in my view - not be justified as, for instance, in the case of the Teop noun *kapa* 'skin of animals and root crops, but not of humans / shell of turtles, shellfish and nuts / bark of trees' (2). But others may have a different view. As Atkins & Rundell (2008:269) state, "there is little agreement about what word senses are or how broad their scope should be, and no definite way of knowing when one sense ends and another begins."

4.4. Homonyms

To save space in commercial bilingual print dictionaries an entry may comprise distinct, but formally and semantically related lexemes that belong to different parts of speech as, for example, the entry of *skin* in an English-German dictionary which is quoted here in an abbreviated form:

(9) English - German (LGED)
skin I 1. Haut 2. Fell, Pelz 3.Haut, Schale, Hülse, Schote, Rinde ... III 14. schälen

In an electronic ULD it is more user-friendly to present such homonyms as separate lemmas as in (2) and (5).

5. Examples

5.1. The function of examples

As in all other kinds of dictionaries, examples in ULDs complement the translation equivalents or meaning descriptions “because they show how a word is actually used, thus returning the word to its natural environment in context after decontextualising it by listing in isolation” (Kosem 2016:90). In addition, they have the function of proving the existence of the word in the under-researched language.

5.2. Authentic examples

Since ULDs aim at documenting lexemes and their meanings of a language that is not widely spoken, the examples do not only have the function of illustrating the grammatical, semantic and pragmatic features of words, but also the function of providing evidence for

- their authenticity,
- the adequacy of translation equivalents or meaning descriptions, and
- the adequacy of the grammatical information given in the entries.

The requirement of authenticity implies that the source of all examples is given. The examples should not be invented by the lexicographer although there are good arguments for inventing examples in learners’ dictionaries of European languages (Adamska-Sałaciak 2013:227).

If the lexicographers see the need of additional examples for some words, they can ask native speakers to create a little story, a short description or just a few sentences that contain these words. Before such sentences are entered into the dictionary, they should be checked with other native speakers and stored in special files of the corpus where they are accompanied by their specific metadata. In the Teop Language Corpus, for instance, each of these example files only contains the examples of a single author, and in most cases the examples of a file exclusively relate to a single topic such as the description of fish, fishing or house building.

Even if fieldworkers do not remember where their examples come from, they could number and store them in a file called “fieldnotes”, give them the metadata of “unknown origin”, and then use them in the dictionary with their references. Similarly, if native speakers want to edit examples and, for instance, shorten them or exchange difficult words by easier ones, I recommend to accept this, but document the changes because they show the native speakers’ metalinguistic intuitions and may provide interesting data for future research.

5.3. The form examples

Since the meaning of a lexical item manifests itself in context, examples should show as much context as necessary for the user to understand its grammatical construction and semantic relations. If the language allows argument ellipsis across sentence boundaries, this requirement could lead to a sequence of several sentences. An alternative to such long and user-unfriendly examples is to only quote the relevant clause and use the translation to clarify the context. In the example below, for instance, the missing argument is represented in the translation by the pronoun ‘she’ in brackets.

(10) Teop (Aro_01E(Eno).080)
Me=paa abana vo te=a ruene, ...
and=TAM jump GOAL PREP=ART water
‘And (she) jumped into the river, ...’

Since the examples should be authentic, the Teop quotation couldn’t be changed by the insertion of a Teop pronoun in (10). But the original sentence can be shortened if as in (10) the relevant construction, here the goal construction of *abana* ‘jump’, is not affected.

All SL examples should have a free translation in the TL, which may be accompanied by a literal translation to show the differences between the SL and TL constructions.

(11) Teop (Rum_01R.039)
Ahiki be= naa skul vamataa.
not.exist COMPL 1SG.PRON school well
‘My school attendance was not all that good.’; lit. ‘It was not the case that I attended school well.’

Note that in the dictionary the examples are not glossed, because the morphological segmentation and glossing would be confusing for the native speakers and too time consuming for the lexicographer.

5.4. Good examples

Good examples illustrate the use of a lexeme in its different senses in a user-friendly manner (Atkins & Rundell 2008:458-461, Kosem 2016:90, Prinsloo 2013). In particular, good examples:

- are authentic
- are short, syntactically independent clauses;
- do not contain anaphoric pronouns or zero anaphora;
- illustrate the different senses of lexemes;
- show the different syntactic constructions of the lexeme;
- show typical collocations (Atkins & Rundell 2008: 363-373, Svensón 2009:158-187);
- contain frequent, simple words to make the examples easily comprehensible.

If these criteria were equally weighted, most examples in ULDs would not be good ones. But the evaluation of examples in ULDs should take into account that authenticity outweighs other criteria and that additional criteria could outrank the criteria of shortness and simplicity.

Since a ULD is typically a dictionary of a language whose speakers have a culture that is also under-researched, good examples would not only illustrate linguistic features, but also convey encyclopedic information. In the corpus encyclopedic information is found in the descriptions of animals, plants and artefacts, daily activities and customs, so that these descriptions or parts of them can be quoted as examples with translations to complement the information given by the translation equivalents and meaning explanations. Compare (1) with (12):

(12) Teop (Sii_17W_trees.022)
O *naono* *vai* *o* *agevana* *na* *pura-pura* *bata=na*
ART tree DEM ART pink.iris TAM RED-grow CONT=3SG.IPFV
'This tree, the pink siris, grows'

te=a *maa* *apao ...*
PREP=ART PLM old.garden
'in old overgrown gardens ...'

Another possibility would be to quote encyclopedic descriptions and their translations in special entry fields.

6. Semantic fields

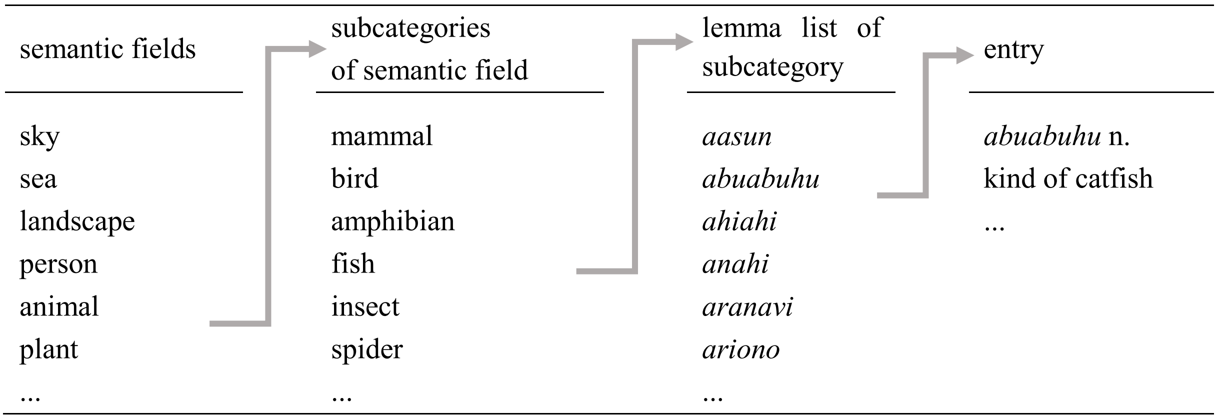
6.1. The function of semantic fields

Since size and content of a ULD heavily depends on its production circumstances, its users cannot expect to always find what they are looking for. Therefore, to avoid disappointment and facilitate further research, an electronic dictionary could provide an overview of the semantic fields covered by the dictionary. Each of them would then in turn be linked to a list of subcategories, if these exist, and either the semantic fields or the subcategories would be linked to a lemma list and each lemma of this list to the entry in which the respective lexical item is found.

6.2. The definition of semantic fields

A semantic field consists of semantically related lexical units and, consequently is language specific though language comparison shows that the semantic fields of different languages show a great deal of overlapping as Lehrers (1974:155-167) investigation of cooking words demonstrates. For some types of semantic fields see below §6.3.

The semantic fields are independent of parts-of-speech. One and the same semantic field may contain lexical units of different parts of speech as, for example, the semantic field NUMERAL in English and Russian (Corbett 1978).



semantic fields	subcategories of semantic field	lemma list of subcategory	entry
sky	mammal	*aasun*	*abuabuhu* n.
sea	bird	*abuabuhu*	kind of catfish
landscape	amphibian	*ahiahi*	...
person	fish	*anahi*	
animal	insect	*aranavi*	
plant	spider	*ariono*	
...	...	...	

Figure 1. Linked semantic fields, subcategories of semantic fields, lemma list and entry (the example comes from the Oceanic Teop language spoken in Papua New Guinea

In a ULD with the target language English the labels of the semantic fields are in English and, consequently, designate English concepts. Compilers of a ULD cannot assume that any concept that comes to their mind in English would have an equivalent in the minds of the speakers of this language. For English speakers, for example, the semantic domain FISH comprises all fishnames and for some people also *dolphin* and *whale*, but not *turtle* and *octopus*, whereas in Teop generally fish as well as whales, dolphins, turtles and octopusses are considered as subcategories of *iana* 'animal that swims' which is not an equivalent of *fish.*

6.3. Types of semantic fields

The semantic fields are defined in terms of semantic relations that hold between the senses of lexical units. For instance, a semantic field can be:

- a superordinate concept that comprises subordinate concepts of the same kind of entity, event or property, e.g.

  pig, *eagle*, *turtle*, *frog*, *tuna*, *wasp*, *beatle* denote a kind of ANIMAL

  butcher, *cut*, *chop*, *carve*, *slice* denote a kind of CUTTING

- the concept of a whole, that consists of several parts of different kinds, e.g.

  HOUSE: *roof*, *thatch*, *ridgepole*, *wall*, *door*, *window*

  TREE: *branch*, *twig*, *leaf*

- the concept of a group or collection that consists of two or more members of different kinds, e.g.

  mother, *father*, *child*, *sister*, *brother* denote members of a FAMILY, they are related to each other by KINSHIP

  Monday, *Tuesday*, *Wednesday*, *Thursday*, *Friday*, *Saturday*, *Sunday* denote the days of the week; each of them is a WEEKDAY

- the concept of SPACE, e.g.

  top, *front*, *back*, *inside*, *in*, *under*, *behind*, *above*, etc.

- the concept of a situation that characteristically involves particular activities, people, things, a particular place and perhaps a particular time of the day or the year, e.g. COOKING and HARVEST.

The semantic fields are not disjunct categories. A lexical unit, for example *chicken*, may be assigned to the semantic field BIRD, because it shares salient features with the hyperonym *bird* and its co-hyponyms *eagle*, *hawk*, and *sparrow*, but at the same time its shares the feature of being eaten with other kinds of food which are shared with *fish* and *chips*, but not with *eagle*, *hawk* and *sparrow.* Consequently, it also belongs to the semantic field FOOD.

6.4. Lists of semantic fields in ULDs

Since "there is no real consensus on what constitutes a semantic field or semantic domain, nor how it can be identified" (Majid 2015:366), the approach suggested here is an opportunistic one that accepts the impossibility of classifying the senses of all lexical units in a dictionary and leaves the selection of semantic fields to the dictionary compilers though they should follow some principles as suggested below.

The kind, number and labels of semantic fields depends on the content of the dictionary and to some extent on linguistic and anthropological conventions. There are several lists on the internet, which may help to find suitable labels for one's own list of semantic fields. However, I recommend not to copy them, but create a list that accounts for the content of the dictionary and is consistent and user friendly; compare the filtering options by "category" of the 12 dictionaries presented on the web page of the Australian Society for Indigenous Languages (http://www.ausil.org.au/node/3717).

In the case of hierarchical semantic relations a lexical unit can be related to superordinate categories of different levels. A turtle, for example, is a kind of reptile, a kind of vertebrate, a kind of animal, and a kind of life form, and similarly, a fingernail is a part of the finger, a part of the hand, a part of the arm, and a part of the body. Which level is chosen to define a semantic field is a matter of practical considerations. If the lexical database only contains 10 animal names, it would not make sense to split the semantic field ANIMAL into subfields. But if there are hundreds of animal names, a division into subcategories provides a useful overview.

A semantic field like ANIMAL may contain some large subfields as for instance FISH and INSECT, while others like MAMMAL or SHELLFISH are very small. This imbalance can be accounted for by having the smaller subfields subsumed in a single field labelled UNCLASSIFIED:

semantic field	semantic subfield	examples of lexical units
ANIMAL	FISH	*halfbeak*, *moray*, *mullet*, *ray*, *shark*, *tuna*, *etc.*
ANIMAL	INSECT	*ant*, *bee*, *beetle*, *cricket*, *grasshopper*, *etc.*
ANIMAL	UNCLASSIFIED	*alligator*, *clam*, *dolphin*, *monkey*, *pig*

Table 1. Illustrative classification of the semantic field ANIMAL

Another reason for having a subfield UNCLASSIFIED may simply be that the compiler doesn't know how to classify lexical units, which may happen especially with verbs:

semantic field	semantic subfield	examples of lexical units
ACTION	BREAK	*break*, *chip*, *crack*, *crash*, *crush*, etc.
ACTION	EAT	*eat*, *munch*, *crunch*, *devour*, etc.
ACTION	UNCLASSIFIED	*play*, *work*, etc.

Table 2. Illustrative classification of the semantic field ACTION

Once a semantic field has been defined, all items of the same kind should be assigned to this category. A list that has a semantic field FISH and at the same time a separate category on the same level labelled SHARKS, RAYS as recommended by The Language Explorer (FLEx) Lexicon is inconsistent.

7. Concluding remarks

A dictionary is an essential contribution to the documentation of an under-researched language and the basis for all kinds of future linguistic and anthropological research provided that the lexicographers and the users are aware of its special characteristics with respect to the lemma list (§3), the meaning description and equivalents (§4), and the examples (§5).

As in modern commercial lexicography for European languages, the prerequisite for making a dictionary of an under-researched language is the existence of a corpus (§2.2). The lexical items are extracted from the corpus and analysed on the basis of their contexts. Correspondingly, the examples given in the dictionary articles to illustrate their syntactic and semantic characteristics are citations from the corpus (§5.2). They are presented with a reference to their sources so that they also provide evidence of the existence of the lexical item and the adequacy of the information given in the dictionary article.

Examples that come from unidentifiable fieldnotes or that have been modified or invented by native speakers can be given labels for their identification, stored in special files which are integrated in the corpus with metadata, and then be quoted with their reference labels so that other researchers are informed about their origin.

Dictionaries for non-academic, educational purposes for the speech communities can be derived from the scientific ones, by abandoning everything that is believed to impair the userfriendliness for their potential users. A different type of derived dictionary that may be attractive for the speech community is a dictionary for rapid translations. For this purpose encyclopedic meaning descriptions would be replaced by simple translation equivalents irrespective of to which extent their accuracy is dependent on the context (cf. §4) and the examples would be deleted.

Since scientific ULDs are not commercial products, the lexicographers are free to invent new types of dictionaries as long as they conform to the standards of authenticity and verifiability. They may, for example, give up the strict distinction between monolingual and bilingual dictionaries and the ambition to write a comprehensive dictionary. Instead they may choose a single theme and produce a small dictionary in which the meanings of lexemes are explained in the source and the target language.

8. Exercises

Exercise 1

Search the internet for three bilingual on-line ULDs and write down for each language:

- the web address,
- the name or names of the source language,
- its genetic affiliation, and
- the name of the place or places where it is spoken,
- the name of the target language and its genetic affiliation.

Exercise 2

Choose one of the three languages and answer the following questions:

- Does the dictionary have an introduction? What kind of information is given in the introduction? Is there any important information missing?
- How many entries does the dictionary contain?
- What kind of information is given in the entries?
- What kinds of lemma do you find in the lemma list?
- What kinds of meaning descriptions are given in the entries?
- Does each entry contain at least one example? Are the examples translated? Is there any information about their origin?
- What kind of search options does the dictionary offer?

- Do you think that the dictionary can be used for linguistic or anthropological research? What kind of research questions could it be used for?

9. Abbreviations

9.1. Glosses

1	1st person
3	3rd person
ART	article
CONT	particle denoting continuous aspect
COMPL	complementiser
GOAL	preposition marking the goal of motions
DEM	demonstrative
IPFV	imperfective aspect marker
PLM	plural marker
PREP	the multi-purpose preposition *te*
PRON	independent or clitic pronoun
SG	singular
TAM	tense-aspect-mood marker preceeding the head of the verb complex

9.2. Further abbreviations

SL	source language
TL	target language
ULD	under-researched language dictionary

10. References

10.1. Books and articles

Adamska-Sałaciak, Arleta. 2013. Issues in compiling. bilingual dictionaries. In In Howard Jackson (ed.) *The Bloomsbury companion to lexicography.* London etc.: Bloomsbury, pp. 213-231.

Adamska-Sałaciak, Arleta.2016. Explaining meaning in bilingual dictionaries. *The Oxford handbook of lexicography.* Oxford: OUP, pp. 145-169.

Atkins, B.T. Sue & Michael Rundell. 2008. *The Oxford guide to practical Lexicography.* Oxford: OUP.

Béjoint, Henri.2000. *Modern lexicography.* Oxford: OUP.

Cablitz, Gabriele. 2011a. Documenting cultural knowledge in dictionaries of endangered languages. In *International Journal of Lexicography* 24.4, pp. 446-462.

Cablitz, Gabriele. 2011b. The making of a multimedia encyclopedic lexicon for and in endangered speech communities. In Geoffrey L. J. Haig, Nicole Nau, Steafan Schnell & Claudia Wegener (eds.). *Documenting endangered languages.* Achievements and perspectives. Berlin/Boston: De Gruyter Mouton, pp. 223-261.

Chelliah, Shobhana L. & Reuse, Willem J. 2011. *Handbook of descriptive linguistic fieldwork.* Dordrecht, Heidelberg, London, New York: Springer.

Corbett, Greville G. 1978. Universals in the syntax of cardinal numerals. *Lingua* 46, pp. 355-368.

Coward, David F. & Grimes, Charles. 2000. *Making dictionaries*. SIL International, Waxhaw, North Carolina.

Durkin, Philip (ed.). 2016. *The Oxford handbook of lexicography.* Oxford: OUP.

Fontenelle, Thierry. 2016. Bilingual dictionaries; history and devekopment; current issues. In Philip Durkin (ed.) *The Oxford handbook of lexicography.* Oxford: OUP, pp. 44-61.

Fuertes-Olivera & Henning Bergenholtz. 2011. *e-Lexicography.* London, New Delhi, New York, Sydney: Bloomsbury.

Greaves, Chris & Martin Warren. 2010. What can a corpus tell us about multi-word units? In Anne O'Keeffe & Michael McCarthy. *The Routledge handbook of corpus linguistics.* Abingdon: Routledge, pp. 212-226.

Hanks, Patrick. 2012. Corpus evidence and electronic lexicography. In Sylviane Granger & Magali Paquot. *Electronic lexicography.* Oxford: OUP, pp. 57-82.

Hyslop Malau, Catriona. 2011. Sustaining Vures: making products of language documentation accessible to multiple audiences. In Geoffrey L. J. Haig, Nicole Nau, Steafan Schnell & Claudia Wegener (eds.). *Documenting endangered languages.* Achievements and perspectives. Berlin/Boston: De Gruyter Mouton, pp. 305-329.

Jackson, Howard (ed.). 2013. *The Bloomsbury companion to lexicography.* London, New Delhi, New York, Sydney: Bloomsbury.

Kilgarriff, Adam. 2013. Using corpora as data sources for dictionaries. In Howard Jackson (ed.). *The Bloomsbury companion to lexicography.* London etc.: Bloomsbury, pp. 77-96.

Kosem, Iztok. 2016. Interrogating a corpus. In Philip Durkin (ed.). *The Oxford handbook of lexicography.* Oxford: OUP, pp. 76-93.

Lehrer, Adrienne. 1974. *Semantic fields and lexical structure.* North Holland Linguistic Series 11. Amsterdam etc.: North-Holland Publishing Company.

Lev, Robert. 2013. Identifying, ordering and defining senses. In In Howard Jackson (ed.). *The Bloomsbury companion to lexicography.* London etc.: Bloomsbury, pp. 284-302.

Majid, Asifa. 2012. A guide to stimulus-based elicitation for semantic categories. In Nicholas Thieberger. *Linguistic Fieldwork.* Oxford: OUP, pp. 54-71.

Majid, Asifa. 2015. Comparing lexicons cross-linguistically. In John R. Taylor (ed.). 2015. *The word.* Oxford: OUP, pp.364-379.

Moe, Ronald. 2003. Compiling dictionaries using semantic domains. In *Lexicos* 13 (AFRILEX-reeks / series 13:2003), pp. 215-223.

Moe, Ronald. 2007. Dictionary development program. In *SIL Forum for Language Fieldwork 2007-009. December 2007.*

Moon, Rosamund. 2010. What can a corpus tell us about lexis? In Anne O'Keeffe and Michael McCarthy (eds.). *The Routledge handbook of corpus linguistics.* Abingdon: Routledge, pp. 197-211.

Mosel, Ulrike. 2006. Fieldwork and community language work. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.). *Essentials of language documentation.* Berlin, New York: Mouton de Gruyter, pp. 67-85.

Mosel, Ulrike. 2011. Lexicography in endangered language communities. In Austin, Peter & Sallabank, Julia (eds.). *The handbook of endangered languages*. Cambridge: CUP, pp. 337-353.

Mosel, Ulrike, 2012. Morphosyntactic analysis in the field - a guide to the guides. In Nick Tieberger (ed.). *The Oxford handbook of linguistic fieldwork*. Oxford: OUP, pp.72-89.

Munro, Pamela. 2002. Entries for verbs in American Indian language dictionaries. In William Frawley, Kenneth C. Hill & Pamela Munro (eds.). *Making dictionaries. Preserving Indigenous Languages of the Americas*. Berkeley, Los Angeles, London: University of California Press, pp. 86-107.

Nida 1997. The Molecular Level of Lexical Semantics. In *International Journal of Lexicography* 10.4, pp. 265-274.

Pawley, Andrew. 2011. What does it take to make an ethnographic dictionary? In Geoffrey L. J. Haig, Nicole Nau, Steafan Schnell & Claudia Wegener (eds.). *Documenting endangered languages.* Achievements and perspectives. Berlin/Boston: De Gruyter Mouton, pp.263-287.

Prinsloo, Daniel. 2013. New developments in the selection of examples. In Rufus Hjalmar Gouws, Ulrich Heid, Wolfgang Schweickard, Herbert Ernst Wiegand eds.). 2013. *Supplemnetary Volume Dicitionaries, an international encyclopedia of lexicography*. Berlin, Boston: De Gruyter Mouton, pp. 509-516.

Svensén, Bo. 2009. *A handbook of lexikography.* Cambridge: CUP.

Tarp, Sven. 2008. *Lexicography in the borderland between knowledge and Non-knowledge*. Tübingen: Max Niemeyer Verlag.

Thieberger, Nicholas. 2011. Building a lexical database with multiple outputs: Examples fro legacy data and from multimodal fieldwork. IN *International Journal of Lexicography*. Vol. 24 No. 4, pp. 463-472.

Thieberger, Nicholas. 2015. The lexicography of indigenous languages in Australia and the Pacific. In Patrick Hanks & Gilles-Maurice de Schryver (eds.). *International Handbook of Modern Lexis and Lexicography*. Berlin: Springer.

10.2. Dictionaries

CCAD 2009. Collins Cobuild Advanced Dictionary. Glasgow: Harper Collins Publishers.

LGED 2004. *Langenscheidt Muret-Sanders Großwörterbuch Englisch-Deutsch.* Herausgegeben von der Langenscheidt-Redaktion. Berlin etc.: Langenscheidt.

Lichtenberk, Frantisek.2008. *A dictionary of Toqabaqita.* (Solomon Islands). Canberra: Pacific Linguistics. Research School of Pacific and Asian Studies.

Mosel, Ulrike (ed.). 2010. *A inu.* The Teop-English dictionary of house building. Kiel: Seminar für Allgemeine und Vergleichende Sprachwissenschaft. pp.49.

Mosel, Ulrike (ed.). 2012b. *Vagana.* The Teop-English encyclopedia of fishing. Kiel: ISFAS Allgemeine Sprachwissenschaft.

Mosel, Ulrike (ed.). 2014a. *A Iana.* The Teop-English encyclopedia of fishes and other swimming creatures. Kiel: ISFAS Allgemeine Sprachwissenschaft.

Mosel, Ulrike (ed.). 2014b. *Kehaa.* Shellfish. Kiel: ISFAS Allgemeine Sprachwissenschaft.

Mosel, Ulrike (ed.). (in prep.) *Amaa naovana bara amaa meha amaa taba vaa rasuu bara kasuana ae vaan.* The Teop-English children's encyclopedia of birds and other animals of the jungle, the beach and the village. Kiel: ISFAS Allgemeine Sprachwissenschaft.

Mosel, Ulrike (ed.). (in prep.) *O Naono.* The Teop-English plant encyclopedia. Kiel: ISFAS Allgemeine Sprachwissenschaft.

Mosel, Ulrike & Mose Fulu. 1997. *O le fale. O le tusi faamatala upu o le gagana Samoa*. Apia, Western Samoa: Matagaluega Autalavou Taaloga ma Aganuu (Monolingual Samoan dictionary on architecture) Apia, Western Samoa: Ministry of Youth, Sports and Culture.

Wegener, Claudia (in collaboration with Bill Adamson, Thomas Omele, Anthony Pisupisu, Jim Planet, Andrew Rara, Wilson Sungi, John Ninizepo, Charles Ghorosavo, Mary Joseph Abuluvu, Claudette Vangere, Salome Livoni, Laurensia Sipiu, Vangelina Vuvunga, Seserio Togha, Jovita Antairopo, Raphael Lavungana, Colman Maravi, James Pulusala, Joel Sasapa Viriala, Felix Narasia, Edmond Gagavo, Aurélie Cauchard and Ian Scales). 2014. *Savosavo Dictionary*. Bielefeld: Savosavo Documentation Project. https://corpus1.mpi.nl/media-archive/dobes_data/Savosavo/Savosavo/Materials/Annotations/Savosavo_Dictionary.pdf

10.3. Websites

The list of animal names for users of The Language Explorer (FLEx) Lexicon: http://semdom.org/book/export/html/ (accessed 2016/02/08)

information on the SIL Rapid-Word-Collection method is found on http://rapidwords.net/ (accessed 2016/02/08)

A collection of lists of semantic domains, put together by David Nash: http://www.anu.edu.au/linguistics/nash/aust/domains.html (accessed 2016/02/08)

The Australian Society for Indigenous Languages presents 12 Dictionaries on their web-page http://www.ausil.org.au/node/3717 (accessed 2016/02/08)