Lemma collection for the first monolingual dictionary

Oddrun Grønvik
University of Oslo

A NUMBER OF THE WORLD'S LANGUAGES AWAIT lexicographical documentation. Many aspiring mother tongue linguists[1] would like to help document their own languages in a product that is useful to their own people. One such product is the first dictionary where headwords and definitions are in the same language – the monolingual mother tongue dictionary. This article discusses how to start on one aspect of mother tongue lexicography, namely leadword collection for a dictionary when there is very little or no printed literature to excerpt. The technical term for this task is lemma collection[2].

Standard literature on lexicography, including the "how to do it"-handbooks[3], is written against a background of centuries old lexicographical tradition for the world's major languages. In this context, compiling a lemma list is about selection, the basic assumption being that the lexicographer has to discard most of the available materials in favour of the essential. This is no help to would-be pioneers of lexicography in preliterate linguistic communities. But the pioneer dictionaries of European languages were edited from scratch centuries ago, with paper and (feather) pens as tools. Today, they are revered as national monuments and remarkable research achievements. A present day pioneer lexicographer will normally have access to computers and suitable software for dictionary

1 The term "mother tongue linguist" here means 'linguist who studies and investigates his or her mother tongue, i.e. his or her first (spoken) language'.
2 The *lemma* is identified by its headword orthographic form, its word class and (if available) its etymology. Some lexicographers prefer the term *lexeme*, cf. Crystal 2008.
3 Cf. For instance Laindau 2001, Svenssén 2009.

making, which makes the job easier and saves time. But the linguistic analysis at the back of a good pioneer dictionary remains the same.

INITIAL ASSUMPTIONS

This article builds on some assumptions based on discussions of standard conditions with African linguists over the years: there is (1) very little or no written literature, and none written by native speakers, (2) no monolingual mother tongue dictionary, (3) perhaps a bilingual dictionary, edited by a non-native speaker, (4) a usable (if not perfect) orthography which shows principles for word division, (5) a (possibly unpublished) rudimentary description of the grammar, (often based on language typology) and written by non-native speakers.

The prospective editors of the first monolingual dictionary are mother tongue speakers who (1) have had some training in linguistics (2) have studied their own language, privately or formally (at school), (3) have studied at least one foreign language (also probably the language of their education) (4) have not produced a dictionary before.

INITIAL TEXT DATA

The best starting point for lemma collection is running text. A small amount, equivalent to ca 2-3000 words in English or French, will do. If there is no written literature, task number one is to find informants who are good storytellers, record 10-15 minutes of speech and transcribe it. From a lexicographical point of view, one person talking continuously in each recording, with few prompts, is the best thing. Continuous speech produces more complete sentences and a wider vocabulary than dialogue.

The first transcription should be phonemic, in order to test the quality and limits of the orthography and allow for initial adjustments. The transcripts must then be thesaurus-excerpted. This means making a table with every word or word element from the transcribed texts in one column, the equivalent in the established orthography in the next column, then the lemma form and finally a word class label. Each word form needs a sequence number so that one can sort the word forms out of and back into original sequence, in order to see them in context. Each word form – or linguistic unit – also needs a word class label, in order to keep apart lemmas that are homographs. Here is an example (Hogg 1989:194):

NUMBER	TRASCRIBED WORD FORM	ORTHOGRAPHIC WORD FORM	LEMMA	WORD CLASS
1	He	He	He	pronoun
2	Gangs	Goes	Go	verb
3	Hame	Home	home	adverb
4	To	To	To	preposition
5	His	His	his	pronoun
6	Ain	Own	own	adjective
7	Ha'	House	house	noun

In a conjunctively written language, the phrasal word forms must be analysed and segmented i.e. *that which is a little bit small* (Mpofu 2009:168).

NUMBER	LEMMA	TRANSLATION	WORD CLASS
1	ra-	That	subject concord
2	-ka	Which	Stative
3	-it-	Be	verb root
4	a		final vowel verb
5	Doko	Small	Adjective
6	svi-	Little	Adjectival
7	-shoma	Bit	Noun

When the table is completed, there will be a list of perhaps 2500 lemmatised word forms, including punctuation marks. Every element should be listed and the table filled in, for each occurrence, for the lexicographer needs to know about frequency and context.

SORT BY LEMMA AND WORD CLASS

The next thing is to sort the table by lemma and word class, resulting in an alphabetised lemma list. This list is likely to show that (1) one lemma can have more than one transcription (variant forms, even within the same dialect), (2) one orthographic word form may represent different lemmas, depending om context (*houses* can be both a noun in plural form and a verb in present singular form) (3) one lemma form may have different word class tags, depending on its function in the text (c.f. the lemma *home* which can be an adjective, noun, verb or adverb).

Both word forms and lemmas can be homographs. The number and frequency of homographs depends to some extent on the language. In general, languages with many short words and a limited inflection system will have more homographs than languages with long word forms and a rich morphological system.

English has more homographs than Norwegian, and Norwegian more than French. The fact that a language has a conjunctive orthography does not take away the homography factor, for the dictionary will need entries for affixes as well as for free-standing word forms.[4]

Another point to note are the cases where what seems to be word forms of one lemma, appears with different senses in different contexts. Ultimately the lexicographer will have to decide whether such cases should be handled as polysemy (several meanings to one lemma), or whether homograph separation[5] is appropriate.

SORT BY WORD CLASS AND DIVIDE INTO FUNCTION AND CONTENT CATEGORIES

Vocabulary can be roughly sorted into function and content units according to word class. In Germanic languages, the function units typically comprise pronouns, conjunctions, prepositions and adverbs. In a Bantu language the function units comprise for instance concords, conjunctives, noun prefixes and verbal extensions. The most typical feature of function units is that they are few, but frequent. A study of real language, shows that their use is more various than the description in a standard grammar.

Content words are nouns, adjectives, verb, interjections and, in Bantu languages, ideophones. These are open word classes where new words are added at need. Each word class has some highly frequent lemmas, some of which will have grammaticalised functions as well (examples are the verbs *be* and *have*, or the nouns *time, lot*). Most lemmas in the open word classes are fairly infrequent, and even multimillion word language collections may not have a usage example to show. About half the word forms in the Norsk Ordbok 2014 Corpus (NC) occur only once[6].

It is essential to get the description of the most important function units into place early in the dictionary project, as the description of these elements accumulate into a description of how the language rules work in practice. A carefully worked out lemma list of the first 500 lemmas in any continuous text will render a large number of function units, some with many usage examples, and bits of most function systems.

As a test, I have made a lemma list of ca 2200 units out of three Norwegian dialect texts (Aasen (1853): A1, B4, C19) and sorted it by word class (including

4 Duramazwi Guru reChiShona has four separate entries for *nda* as a noun and an ideophone, and *nda-* as a subject concord and .

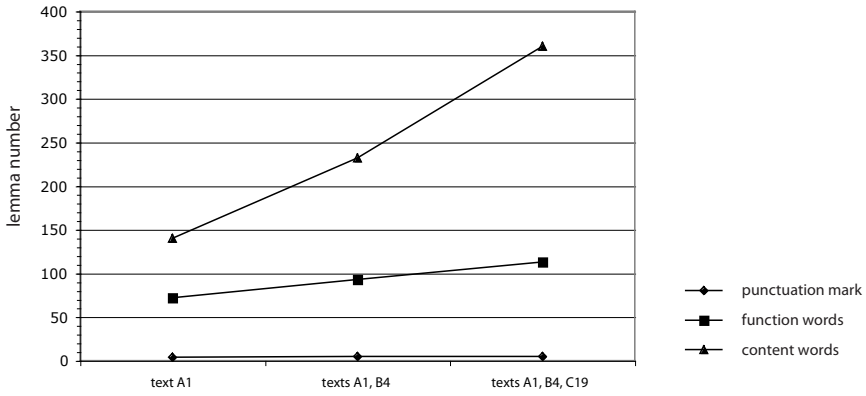
5 The neuter noun *fly* in Norwegian can mean 'aeroplane' or 'small flying insect'. Although word form and ultimate etymology are the same and thre is a sense relationship dictionaries list them as separate lemmas.

6 The Norsk Ordbok corpus in May 2010 had 49,3 mill tokens and 985000 different tokens of which 541000 (55 %) occurred only once. 205000 tokens occurred five times or more.

punctuation marks). In each of the three little texts there were about 50 % function words, 40 % content words and 10 % punctuation marks. But if you look at the numbers of different lemmas within each type, the picture changes :

TYPE	TEXT 1	TEXT 1, 2	INCREASE	TEXTS 1, 2, 3	INCREASE
Punctuation mark	4	5	1	5	0
Function lemma	72	93	21	113	20
Content lemma	140	232	92	360	129
Total	216	330	114	478	149

The number of different punctuation marks is small, and stops increasing very soon. The number of function lemmas represents a third of the total in the first text, but after that, the addition of new lemmas is limited. In contrast, the number of content lemmas increases sharply, as shown in the graph below:



The heterogenity of content words i text becomes more pronounced the more text is added. In the Norwegian language collections[7], which cover about 500 000 different lemmas, the function lemmas comprise less than 2 % of the total, but a look at words in context changes the picture. The fifty most frequent function lemmas represent 15 % of the text in NC[8], and the 1000 most frequent word forms make up 30 % of the text total.

7 Figures from The Meta Dictionary, digital index to the Nynorsk language collections, 30.7.2010.
8 Figures from frequency data presented at the summer seminar of Norsk Ordbok 4.6.2010.

The three small dialect texts in my sample have rendered a total of 478 lemmas, some of which are bound to be homographs, so it is not far off from 500. If the aim is a dictionary of 5000 entries, the first 10 % would now be covered. As the important function words occur many times, it is also possible to start drafting the entries for them.

EXTENDING THE LEMMA LIST

On the basis of the initial lemma list, it is possible to make lists of semantic sets, i.e. words that can be grouped together because they share important features of function or meaning. From a lexicographer's point of view, first dictionary will need draft semantic sets (as controls) for (1) function units, (2) sets of content words and (3) terms of mother tongue description – the metalanguage of the dictionary.

The initial lemma list is a good starting point. It will typically contain some, but not all, of for instance personal pronouns (or subject and object concordances), noun prefixes, words for time and timekeeping, verbs of motion etc. The linguist who is also a mother tongue speaker will see the lacunae and be able to fill them in, much more easily than a linguist who is not a mother tongue speaker. A future task is to secure independent documentation of the filled-in material, since introspection is not trustworthy enough for lemma collection. One always wants independent documentation.

TYPES OF SEMANTIC SETS

Semantic sets are can be open or closed. The closed ones have a fixed set of units, the open ones can be added to indefinitely. In any dictionary it is important to cover the well known and much used fixed sets. To publish a dictionary that lacks the word for a day of the week or an important pronoun is not only embarrassing, it is the sort of embarrassment a pioneer dictionary cannot afford.

Starting with the initial lemma list, one must consider what semantic set each lemma might belong to, and draft lists of sets that are as complete as possible. Typical closed sets of content lemmas are calendar words (days, months, feast days and annual events), words for indicating time, words for regular meals and mealtimes, kinship terms, colours, numbers. But "closed" is not a absolute term. Different dialects may for instance have different words for the same phenomenon, in which case all should be noted.

My three test sample texts (cf section 5) were a place legend, a fairy tale and a homestead description. They yielded a number of topographical nouns and verbs of perception, which would have been a natural starting point for expanding semantic sets.

Semantic sets of content words take their origin and usage support from the outside world, both from phenomena that are grouped together (for instance cooking tools and processes), and from the concept models that language users have created on the basis of experience. Lexicographers will therefore encounter clashes between different concept models when they collect materials. In Germanic languages two measuring systems collide, the decadic system and older systems based on body part measures (in which for instance one foot equals ca 30 centimeters). All concept models present in the language, with their systems of interrelated terms, should be described fairly and exactly. None should be suppressed as "wrong" or "outdated". In this way one can document traditional culture through vocabulary without colliding with school syllabus needs.

Semantic sets of function words can be difficult to delimit. This may be a homograph issue. Germanic languages are rich in small words which can function as prepositions, adverbs and conjunctions – three lemmas or one? The same goes for Bantu language affixes. Function word may in certain uses have a schema form, for instance. *Either, or* and *either … or …* Is it proper to describe these word forms as two or three lemmas, and if two, under which entry should the third one be edited? The doubts on how to handle single word forms may also spring from grammatical schemas. Should *we* and *us* be covered in one entry (with *us* treated as an oblique form of *we*), or in two?

The analysis of function words and units as lemmas has to be language specific, and will depend on the chosen word class system. Independent of language, the best thing is to start with an extensive registration of function word forms at lemma level, and plan separate entries for each form. There are two reasons for this. (1) User convenience. A first mother tongue dictionary addresses itself to inexperienced users who want to find their information quickly. They will not have the patience to search through long and complex entries. (2) work convenience: adding small entries to each other is easier than prising apart a comprehensive entry. In editing smaller, separate entries, the lexicographer leaves space for easy revision.

One particular semantic set concern the description of the language in the dictionary. A mother tongue dictionary, especially a pioneer dictionary, must have user instruction that tell readers what the dictionary contains, how the entries are constructed, and how to search for information. It should also say something about sources (material collection) and editorial decisions on what is included and what is omitted.

In order to write user instructions in the mother tongue, a *metalanguage* is needed, with words for 'lemma', 'word class', 'verb' and so forth. Other terms needing a mother tongue equivalent are *alphabet, conjunctive, definition, equivalent, noun class, pronunciation, tone, vowel*. Early attention to metalanguage

helps ensure consistency in editing, and saves labour in the long run, because it sharpens the editorial focus on what aspects of linguistics that are show in the dictionary entries. Every metalanguage term must have an entry, which again means that it must be defined – in the mother tongue.

This means, for most pioneer monolingual lexicographers, taking the plunge into term creation. There is a register of techniques for tem creation, but that is not the subject of this article. My only reminder is that terminology is best established by a measure of consensus in the professional community. It is important to build on what is there already, and to listen to the views of colleagues and potential users.

LEMMA COLLECTION FROM TRIAL ENTRIES

With 2500 words of running text, 4-500 lemmas and most word classes represented, one can start drafting trial entries. A text sample of this size will contain the important function words, show morphological patterns and give a number of content words which is certain to deserve an entry. In my test, every function word form, and all content word forms with high frequency (5 or more occurrences) were also found among the 1000 most frequent word forms in NC, and most of them in the high frequency end.

Trial entries require a format, where a set of categories pertaining to lemma description are organised for easy use and reference, preferably as a computer database. The categories need labels – names – which contributes to the metalanguage list. Each entry requires one or more definitions with accompanying usage examples. As defining progresses, editors should keep an eye on their defining vocabulary. Is every lemma used in the disctionary lemma list? If not, add it. Does it occur in any of the dictionary materials? If not, note it down as undocumented – and get documentation.

SUPPLEMENTARY LEMMA COLLECTION

The initial text collection will need expanding, and the expansion should be planned with a view to covering as wide a vocabulary as possible. Text collection will primarily mean recording and transcribing, in order to get words in context.

A topic list for material collection is essential. The lexicographer needs variety. If the dictionary needs words from vegetable farming, a thoughtful and knowledgeable vegetable farmer is needed as an informant. The ideal is a recorded conducted tour where the farmer talks all the time, followed by a critical review of the transcribed text by informant and registrator. The recording should not be so long that it is exhausting to process (transcribe, standardise, lemmatise).

Thirty minutes of continuous talking can yield 2-2500 words – as much as the initial text sample discussed above, or 7-8 pages of print.

In 1992-93 an extensive collection of oral material was done from the University of Zimbabwe by staff and students from the Department of African Literature and Languages. It was found that older people who have knowledge they are anxious to transmit, make excellent informants, but they do not always want to talk about the planned topic. Some recordings from this collection are without a single question or interruption by the interviewer, and several of them yield unique material. An important factor in the interviews is the atmosphere of mutual trust and liking, based on respect and genuine interest from the interviewer. When planning material collection, these aspects should be taken into consideration, and the interviewers need to be well prepared.

PROBLEM AREAS AND ONE STANDARD SOLUTION

Let me start with the solution. If the digital editing format of the dictionary has a "do not print"-function at entry level, the database may contain entries that are not going to be included, but are useful for future reference. Entries must all the same be written as if they are going to be published, in order to maintain the quality of the dictionary database.

Three common problem areas in lemma collection are derivations, multiword lemmas and lemmas associated with offensive language.

Derivations – predictable or not? All languages have systems for deriving new lemmas from existing lemmas by adding affixes. Derivation can change the word class of a lemma and modify meaning. Some derivation methods generally give highly predictable sense and usage changes. This fact turns derived lemmas into a lexicographical problem. If wholly predictable, do they deserve a separate entry? How to decide? Once again, one has to look at materials showing each lemma in use. Making a general preclusive decision on the treatment of (actual or potential) derived lemmas is not good lexicographical practice.

Lemmas consisting of more than one word form have troubled many dictionary editors. Which quality turns a reduplicated ideophone, or a noun with a postpositioned modifier, into an independent lemma? The test lies in the defining process. If the meaning of a multiword expression cannot be predicted from its parts, and is in fairly common use, it must be considered for inclusion at entry level. Lemmas should be searchable as lemmas. It is poor lexicographical practice to put all compounds into the usage examples under the entry for the first unit in the multiword item, and then comment on each usage example.

Finally a word about the issua of lemmas associated with offensive language. Every language has its swearwords, a vocabulary associated with sex, and

derogative expressions about other people. Some of these words are frequent and well known. The dictionary users (or their parents) may still not want to see them in a school dictionary. These words belong on the edges of standard vocabulary. Excluding them from the first printed dictionary is not a major lexicographical sin. A first monolingual dictionary should not be a storehouse of offences! On the other hand, teaching biology in the mother tongue will be difficult if there is no vocabulary for reproduction. One possibility is to include lemmas with factual definitions, but exclude information about emotive usage.

Whichever solutions are chosen, the dictionary editor must not suppress materials showing these words in use. Their existence, meanings and usage will one day require documentation, and core materials from the collections underlying the first monolingual dictionary will then have pride of place.

CONCLUSION: TAKE HEART AND GET STARTED

Seen from a preliterate linguistic community, the lexicography of the Western world looks daunting. The linguists of poorly documented languages know of lexicography through polished college dictionaries or multi volume national prestige projects. They do not normally encounter the beginnings of European lexicography, or even modern dialect lexicography, which would show models closer to their own position.

Success is not implicit in every beginning – but a good start takes the budding lexicographer a long way. If this article can encourage more linguists to attempt mother tongue documentation through lexicography, it will have achieved its aim.

LITERATURE

ALLEX Project history (1992-2006) http://www.edd.uio.no/allex/

Chimhundu, H. et al (1996). *Duramazwi reChiShona.* ALLEX – College Press, Harare.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics.* Malden, Mass.: Blackwell.

Hogg, J. (1987). *The Private Memoirs and Confessions of a Justified Sinner*. Penguin classics.

Landau, S. (2001): *Dictionaries. The Art and Craft of Lexicography.* 2. ed. Cambridge University Press.

Mpofu, N. (2009): The Shona Adjective as a Prototypical Category. Doctoral thesis. Department of Linguistics and Scandinavian Studies, University of Oslo

NO 2014 Meta Dictionary http://www.edd.uio.no/perl/search/search.cgi?tabid=571&appid=7

NO 2014 Corpus http://no2014.uio.no/korpuset/

Svenssén, Bo (2009): *A handbook of lexicography: the theory and practice of dictionary-making.* Cambridge University Press.