Abstract
Making a dictionary of a minority language, especially, an endangered one, requires solving several problematic issues specific to this kind of lexicographical work. The primary issues to be dealt with in this case are as follows. (1) Identification of the target audience. There are three possible variants: language community, academic community, or both communities. The choice depends on what lexicographical work has already been done for a given language and what are the needs and expectations of the communities. (2) Collecting data and compiling the basic vocabulary. There are several approaches to this issue including a) translation of the list of the most frequent words from a European language, b) extraction of the wordlist from a corpus of texts, and c) thematic elicitation from native speakers in order to obtain basic lexical units belonging to a particular semantic field. (3) Creating dictionary entries. In general, every dictionary entry consists of two main components – a lemma and a commentary. Both components usually have certain peculiarities in dictionaries of minority languages. These may include, for example, orthographic representation of the lemma, hierarchic arrangement of word meanings, presentation of encyclopedic information in the dictionary, and so on.

The present paper provides an account of how these problems were dealt with in the Comprehensive dictionary of Ket, a highly endangered indigenous language spoken in Central Siberia.

Keywords: lexicography; dictionary making; minority languages; Siberian languages; Ket

1 Introduction
The present paper aims, on the one hand, to show how structural and functional peculiarities of a language affect the tasks of dictionary making and, on the other hand, to outline some basic questions to be solved when compiling a dictionary for a minority language through the example of Ket, a highly endangered language spoken in Central Siberia.

It is obvious that though the process of dictionary making for minority languages relies mainly on certain universal lexicographical principles, it inevitably has its own peculiarities. “The main differences between ordinary dictionary projects and those for endangered languages are that the latter are non-profit enterprises with limited resources of time, money and staff. Also, the linguists responsible for the project are not native speakers of the language. The staff working on the dictionary usually consists of a linguist (or several linguists) and a few indigenous people from the endangered speech community. While the linguist does not have a thorough knowledge of the language under investigation, the native speakers are not trained in linguistics” (Mosel 2004: 1). The recently completed project on compiling a comprehensive dictionary of Ket (CDK for short) (Kotorova, Nefedov 2015) is no exception in this respect, and the authors of the dictionary are therefore fully familiar with all the problems that result from this situation.
2 Identifying Target Audience

The first task before starting to make a dictionary is to identify the target audience by deciding who is going to be the prospective user. This decision plays an important role in choosing what should and what should not be included in a dictionary. In fact, there are three possible choices concerning the prospective user, when making a minority language dictionary:

- a dictionary mainly for speech community
- a dictionary mainly for academic community
- a dictionary for both communities

The actual choice depends on what lexicographical work has already been done for this language and what the needs and expectations of speech and academic communities are.

With respect to the Ket language, the situation with the existing dictionaries at the time when the CDK project started was as follows. The first official Ket dictionary was compiled and published by Heinrich Werner in 1993 (Verner 1993). It was bilingual and bidirectional in the sense that it was both Ket-Russian and Russian-Ket. This dictionary was meant, first of all, for teachers and pupils in native Ket schools, therefore it contains only basic vocabulary (about 2800 entries in each part). The entries in this dictionary are compact, consisting of only one- or two-word examples (mainly for verbs). It has no comprehensive usage contexts. The dictionary is based on the Southern Ket dialect, as the one with the largest number of speakers.

In 2001, Zoja Maksunova, a native speaker of Ket, compiled a similar school dictionary based on the Central Ket dialect (Maksunova 2001). Unlike the aforementioned one, this dictionary contains only Ket-Russian entries. The dictionary contains about 2500 lexical units. The structure of the dictionary entries are similar to those in Werner’s school dictionary.

Werner’s “Comparative dictionary of the Yeniseic languages”, published in 2002, represents a different type of dictionary. It is a solid three-volume work containing about 11,000 entries. This dictionary was not intended for native speakers, but for academic circles. Therefore, it is more complex structurally, and it was possible for the author to choose German as a metalanguage. The main concern of the author was to compare words with similar meanings in the Yeniseic languages and, if possible, to reconstruct their Proto-Yeniseic forms. In the body of the entry, he lists word forms from the documented Yeniseic languages: Ket, Yugh, Kott, Arin, Assan and Pumpokol. Still, this fundamental piece of work contains limited information on the semantic content of Yeniseic words, and Ket words in particular. The author is more interested in the etymology of the Ket word and its connection with words in the related languages.

As can be seen, despite a number of available dictionaries, a comprehensive dictionary of Ket that would contain as much documented vocabulary as possible was still lacking. This is why the CDK project came into being.

In many cases, the most obvious choice is to make a dictionary that would serve both the interests of native speakers and linguists (as a number of dictionaries for Native American languages, for example, the Cherokee-English dictionary (Feeling, Pulte 1975), where an entry is represented both in practical orthography and in the Cherokee syllabary). Unfortunately, in spite of the fact that 21.9% of ethnic Kets report they can speak their native language, only 2.8% are able to read and write it. A somewhat larger percent, 10.5%, can only read it (Krivonogov 2003: 86). However, according to our fieldwork experience, even these sad numbers did not reflect the real situation at that time, not to mention that it has become much worse today (Nefedov 2015: 5). Therefore, in the case of CDK, the linguist community was chosen to be the primary target audience.
As will be shown further, this choice had influence on decisions concerning various practical questions while compiling the dictionary. However, in order to provide the native speakers with an opportunity to somehow use the dictionary, it was decided to choose the Russian language as the meta language in CDK (though translations of all lemmata into German and English were provided as well), since all Kets have a good command of Russian.

The dramatic socio-linguistic situation also forced the CDK team to concentrate on making a unidirectional dictionary (i.e. Ket-Russian) only. There is no sense in making a Russian-to-Ket part, when there is no real need for it in the native community. Such work would be necessary only in case of active bilingualism, where native speakers believe that a dictionary would be useful to them in using the major language and improving their knowledge of it (s. Lichtenberk 2003: 390, 398). For the academic community it seems sufficient to provide an index (in our case, Russian-Ket, English-Ket and German-Ket) in order to make word searches less complicated.

3 Gathering Material and Compiling the Basic Vocabulary

The next important task to be done when making a dictionary is to gather language material and compile the basic vocabulary. There are several ways of making the initial wordlist. Linguists experienced in compiling dictionaries of minority languages suggest the following methods:

1) Translation of the list of the most frequent words from a European language (very often English). The size of such a list varies depending on the size of the prospective dictionary. This method is simple and easy enough, but it has a crucial drawback, as such a wordlist will not be representative of the lexicon of the indigenous language and will miss all cultural-specific concepts, some of which may also be basic. On the other hand, the list may contain words, which do not have a translation equivalent in the indigenous language (Mosel 2004: 3).

2) Extraction of the wordlist from a corpus of texts of the described language. The advantage here is that a representative corpus will cover almost everything. However, it involves a great deal of work to extract lemmata from the corpus data. Moreover, it can be applied only to languages where there is such a corpus, which is not the case with most indigenous languages.

3) Thematic elicitation from native speakers to find words for narrowly defined subject areas (for example, colour terms, housekeeping, etc.). This method also helps reveal the basic and culturally specific words of the target language. However, it produces fragmentary results and should be used only as an auxiliary method.

As for CDK, it was a great advantage that there exists a hand-written card file dictionary at the Laboratory of Indigenous Languages of Siberia in Tomsk. It contains a large number of entries extracted from the field notes ever made by Tomsk Ketologists.

In the first stage of the CDK project (dedicated to all parts of speech except the verb), we extracted all the relevant data from the card files and converted them into an electronic format. However, as our experience has shown, this card file does not cover the entire lexicon, therefore the CDK wordlists were updated with the data from Werner’s comparative dictionary. Nevertheless, even after that there were some gaps left, for example, recent loanwords in Ket.

4 Entry Description

The next significant problem in lexicographical work is making dictionary entries. The composition of each dictionary entry is determined by the user’s potential goals. In translation lexicography, the
most important and primary question that the dictionary user seeks to answer involves the totality of meanings and usage possibilities of the foreign vocabulary item. However, the semantic information in the entry of a bilingual dictionary is traditionally accompanied by another sort of information that permits recognition of a number of required and optional parameters. It should be noted though that any dictionary entry includes two elements of prime importance: the citation form and the commentary to it. Both components have certain characteristics (cf. Kotorova 2003: 140).

**HEAD WORD:**

<table>
<thead>
<tr>
<th>OBLIGATORY COMPONENTS</th>
<th>OPTIONAL COMPONENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Spelling</td>
<td>1. Grammatical features</td>
</tr>
<tr>
<td>2. Pronunciation (usually coincides with 1 in unwritten languages)</td>
<td>2. Stylistic remarks</td>
</tr>
</tbody>
</table>

**COMMENTARY TO THE HEAD WORD:**

<table>
<thead>
<tr>
<th>OBLIGATORY COMPONENTS</th>
<th>OPTIONAL COMPONENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Translation of main meaning(s)</td>
<td>1. Contexts of usage</td>
</tr>
<tr>
<td></td>
<td>2. Idiomatic usage</td>
</tr>
<tr>
<td></td>
<td>3. Cultural-historic commentary</td>
</tr>
<tr>
<td></td>
<td>4. Dialectal variants</td>
</tr>
<tr>
<td></td>
<td>5. References to other dictionary entries, etc.</td>
</tr>
</tbody>
</table>

Table 1: Dictionary entry components.

While bilingual dictionaries of languages with rich written and literary traditions include, as a rule, all of the components listed in Table 1, dictionaries of unwritten languages have their own peculiarities. The reason for this is the lack of a generally established norm (either written or oral). It imposes upon the dictionary compiler a complex and serious task of choosing which attested forms are the most frequently used and therefore the most appropriate for use as citation forms. In this way, dictionary compilers are faced with significant normative linguistic problems in addition to tasks of a purely lexicographic nature. At the same time, the field notes and other materials at the lexicographer’s disposal do not always contain the forms needed for a definitive solution to such problems.

In what follows, I will outline each of the components of an entry in CDK.

### 4.1 Head Word

Characterization of the citation form in dictionaries of unwritten languages tends to be limited to phonetic transcription and rudimentary grammatical information. Remarks on stylistic usage are generally absent due to the weak differentiation of the linguistic material into individual functional styles. Etymological comments are possible, but this was not the task of CDK.

Also crucial is the question of whether such a dictionary should contain all of the dialectal variants recorded in the field, or whether it, like other Ket dictionaries, should be based upon a single dialect. The easiest way would be to take the Southern Ket dialect to represent headwords, since the majority of speakers left are Southern Kets. The dialectal variants, if there are any, would be provided after the headword. However, in the card file dictionary we have found a lot of contexts in the Northern Ket dialect. Including these examples could create a sort of disbalance between a headword and its contexts of usage. At the same time, there is no sense in compiling a dictionary on the basis of Northern Ket, because there are only a couple of fluent speakers left.

Alongside the problem of dialect choice, we have faced a problem of field notes unification. All the field notes at the disposal of the Tomsk Laboratory as well as almost all published Ket texts are
represented in a special Cyrillic-based phonetic transcription. In CDK, we decided to use an IPA-based notation format, since the target audience is the academic community. Therefore, the first problem here was to convert the Cyrillic transcription into IPA. Then, we needed to decide how far such unification should go. Should it be strictly phonological or not? The phonological system of Ket was rather well described in the literature (see e.g. Vall/Kanakin 1990, Werner 1997, Vajda 2000, Georg 2007). However, even in these works, as well as in the published dictionaries, notation format preserved some features of phonetic realization. At the same time, making an overall phonological unification through the whole dictionary is a very complicated and time-consuming task. Besides, it will not allow the user to see how the phonological system of Ket is realized and will level the important differences between the Ket dialects. Therefore, we came to the following decision: headwords are given in phonological transcription and illustrative examples are represented in a uniform phonetic transcription. Besides, a headword is provided with its dialectal variants, if their phonetic representation differs much from the phonological one, as in the example below.

(1) èd m, edn; (sket. ɛ̀rʲ, cket. ɛ̀də, nket. ɛ̀rʲì) соболь//Zobel//sable; sur. qɔˀk èd один соболь, sur. ëtн kurtəp связка (шкурок) соболей, sur. kisɛ́ŋ ëtн знаŋ здесь соболей много, bak. hitl qasɛŋ ʰllа ètн внизу (там) соболята, kel. kirʲ qārʲ erdaŋalʲ это шерсть от соболя, kel. erda iːlʲalŋ пища соболя, kur. ëtнna iːlŋ bɔŋlɛs мышь, которой питается соболь, pak. ëdda tam ʰks dugdaptaŋ, bū (t)qagdәqona соболь что-то тащит, он стал гонять его ɪ̀d de dakuŋ dil (d)butɔlut, ëddaŋks da-taŋajal잘 соболь из норы вылез, черкан его ударил (КСД: 51)

Representation of the grammatical information in a dictionary of a minority language has its peculiarities as well. As a rule, lexicographers working on compiling a dictionary for a major language have in mind that prospective users of the dictionary possess certain knowledge of grammar of the language. The situation with indigenous languages is quite different. Here it is important to provide an entry with sufficient grammatical information because otherwise working with the dictionary may become rather complicated. At the same time, since our dictionary is comprehensive, it should in the first place fulfill the task of revealing the content of a lexical item. Thus, it is undesirable to overload entries with grammatical data. Our solution was to provide CDK with a grammatical sketch and to supply entries with a system of grammatical references. It provides the user with all the necessary information while working with the dictionary.

In CDK, the problem of grammatical representation mostly concerns the verb. The nouns are provided with usual grammatical labels, such as gender and number, and the rest of the words have labels, describing them only as such and such part of speech.

There are certain peculiarities in the structure and functioning of Ket verbs and action nominals (traditionally called ‘infinitives’) that make it impossible to use the latter as the citation form for verbs in the dictionary. They are as follows. First, lexical components of an infinitive may be different from that of the corresponding verb form, and it is often difficult to generate one from the other. But what is more crucial in this respect is that for some verbs there are no corresponding infinitive forms recorded and it is not possible to elicit them from native speakers. That is why in CDK we use a special abstract formula as lemma for verbs. The abstract formula contains all the positions of the given verb lexeme except those for proper agreement markers; the sign [ ] marks an affix that presents either past or non-past tense of the given verb lexeme.

Figure 1 below illustrates the components of a verbal entry in CDK.
Figure 1: Ket verb entry components.

The components indicated in Figure 1 are as follows:

1 – lemma
2 – configuration type
3 – derivational model
4 – translation
5 – contextual examples

The grammatical information in the verb entry includes the following components: 1) configuration type – a reference to the type of the configuration of the given verb lexeme; the abbreviation “vt” stands for transitive verbs, the abbreviation v stands for intransitive verbs, while “vk” is for verbs with without verb-internal agreement marking; the number refers to the number of the configuration type; 2) derivational model is a reference to the type of the derivational model of the given verb lexeme.

In addition, the verbal entry contains 1) translation – a phrasal representation of the lexical meaning of the given verb lexeme; and 2) contextual examples – an illustration of the contextual use of the given verb lexeme elicited in the field; the sign \( \equiv \) marks contextual examples taken from the published Ket texts.

### 4.2 Commentary to the Head Word

An obligatory and very important component of the commentary to the citation form in a bilingual dictionary is, as mentioned above, a certain hierarchic arrangement of word meanings reflected in the corresponding translations. Dictionary compilers rely on the totality of contexts in which the given word has been found; and, if they are native speakers, upon their own intuition. Lexicographers compiling dictionaries of unwritten languages are generally not native speakers, and therefore must rely solely on examples in context to determine the various meanings of a given word. Therefore we tried to confirm each meaning of a word with proper contexts. The corpus of contexts is based on following sources: 1) the hand-written card file dictionary at the Laboratory of Indigenous Languages of Siberia (Tomsk), 2) materials from the field work trips organized during the work on CDK, 3) the hand-written field work volumes at the Laboratory of Indigenous Languages of Siberia (Tomsk) 4) the hand-written archive of the renowned Ketologist Erukhim Krejnovič (Sakhalin), 5) published Ket texts and dictionaries, 6) unpublished dissertations. In the case of the latter four, the contextual examples were introduced with a special symbol `\( \equiv \)`, as exemplified below.

(2) \( \text{in nom I} \) \text{стояние (о живых существах)}/Stellen, Stehen (von Lebewesen)/putting, standing (said of living beings); \( \text{kel. дива усозднта дан винуут} \) \text{стояние ребёнка в углу закончилось}, \( \text{kel. ллл дяжна \text{in binut} на улице} \text{стояние людей закончилось II} \) ставящий, стоящий (о живых
Taking into consideration that the target audience of the dictionary will be mostly academia, we decided to supply headwords with translation into German and English, which makes the dictionary useful for scientists who does not possess a good command of Russian. At the same time, we decided not to translate illustrated examples, as it would complicate the work and increase the volume of the dictionary enormously.

It is important to note that the traditional division of dictionaries into monolingual and bilingual ones does not need to be strictly observed in dictionaries for endangered languages, as they are not primarily used for translation. In fact, for many headwords a translation into a European language is not sufficient, because it will not capture the concept of the indigenous language. In these cases, a translation should be accompanied by a definition. Moreover, some encyclopedic information is also included, especially information that deals with traditional social, cultural and religious concepts. Such comments make the dictionary a resource for further linguistic and anthropological research.

However, we have to admit that it was sometimes very difficult to elicit usage contexts for such ethnocultural lexical units, since modern Kets have already lost knowledge of the myths, beliefs, customs and traditions of their folk. Nevertheless, we include such words into the dictionary, and try to give all the necessary comments, e.g.

(3) **imil n, imilaŋ; имиль (кетское лакомство – ореховая масса из желудка белки) // ketischer Leckerbissen (Nussmasse aus dem Magen eines Eichhörnchens) // Ket delicacy (nut paste from a squirrel’s stomach);** mad. inam (t)daqqimmen imilaŋ saːnna ранее жарили желудки белки с орехами, kel. āt saqd imilʲ (d)bilʲ я имиль белки съел, kel. imilʲ ĭnʲ tɔblaʁut, biːlaq имиль долго лежал, сгнил, kel. āp imilʲ daqqɔɣɔlʲda отец имиль жарил

(4) **hosedam f, no pl; Хоседам (властительница подземного мира) // Hosedam (Herrscherin der Unterwelt) // Hosedam (mistress of the underworld); bak. āt bɔɣɔt hɔsɛdam-baːmdɛŋa я пойду к старухе Хоседам, dɔɣ hɔsʲɛdamd ɛ̀j kεˀt dilaq askʌt сказка о том, как Дог ходил убивать Хоседам (СНСС72: 92), hɔsʲɛdam da-iːmbɛsʲ-qaɣa, bɨlʲda baŋga qɔnʲijɔ́bɔŋ если (когда) приходит Хоседам, на всей земле темно становится (ПМБ: 200)

Idiomatic expressions were also included in the dictionary after the special symbol. However, their number is quite small, as is often the case with other indigenous languages. This is the reason why we also tried to include riddles (together with their answers), since they are rather idiomatic, though it is usually not the case in the practice of dictionary making, ex.:

(5) **kūb n, küun; 1. конец // Ende // end; [...] 2. рот, губы (две губы вместе) // Mund, Lippen // mouth, lips; sur. kedda kūp губы [рот] человека, kel. tǐp kūp dubbɔ собака губы облизывает, kel. tajya āp kūp təlimna я на морозе губы застудил, kel. sɛːlda kūp kubɔːlʲ da-hatabuksibεt олень губы куболем она зажимает ⊗ **sum. kūp bąmbesʲ dugdaṭajŋ тот, кто что-то вынюхивает, может украсть [губу по земле волочит] [...]**

(6) **t⁵-a⁴-[l²]-qot⁰ v3 ITER кто-л прибегает, появляется // jmd kommt gelaufen, erscheint // smn arrives running, suddenly appears; kel. āt datuŋ ke’t aːndɛnjaɫ dbutɔlɔt я вижу человека, который из
леса появился дbatəʁɔt я прибегаю (WER2: 227), dbutɔlɔtn [dbutɔlɔtn] они прибегали (WER2: 227), sul. hɔːm tundan' sestĩl (t)butɔʁɔt Хом тогда из речки выскоцил (КФТ: 53), imb. d̪əgbandija įgde re-buíutut на мыс к краю реки [вниз] вышла (КСб: 177) ∙ kul'te nimil'tet dinas tbutɔ́lut (q̥̂j-bाàt) уголёк нырнул, на другой год вынырнул (загадка; отгадка – медведь) (КСб: 229)

5 Conclusion

Peculiarities in lexicographical work with respect to minority languages are in general the result of certain differences in their status compared to that of major languages. In the first place, these differences can be observed in the sociolinguistic situation which is as a rule quite stable for major languages, while for minority ones it varies greatly depending on the state of their endangeredness. This factor plays an important role in setting goals for the authors compiling a dictionary.

Another important difference is that major languages have a stable written and spoken standard, while in minority languages such standard is usually lacking. Therefore, in addition to tasks of purely lexicographical nature, dictionary compilers working with these languages have to deal with language standardization issues.

Finally, it is often the case that minority languages are underdescribed from the grammatical point of view and there are various controversial approaches to certain grammatical phenomena (for example, with respect to defining parts-of-speech in a language). In such situations, dictionary compilers have either to argue for one of the existing points of view or to make up their own one before they can start their lexicographical work.

Therefore, making a minority language dictionary often goes beyond pure lexicographical work, becoming a theoretical and practical scientific enterprise as well as a major means of preservation of an indigenous language and culture.

6 References

Dictionary for a Minority Language: the Case of Ket


**Acknowledgements**

Our thanks go to Professor Bernard Comrie, the Director of the Department of Linguistics of Max Planck Institute for Evolutionary Anthropology (Leipzig, Germany) and the staff of the Institute for providing all-round help and support to the project, which made it possible to complete the work on the Comprehensive Dictionary of Ket.