OUDB Glossing Conventions FLEx8 - Preliminaries

1. Start a new project

1.1. Set-Up: Writing system and alphabetic order

In the task bar – Go to:
File > Project Management > Fieldworks Project Properties > Writing System > Modify > Sorting > Custom ICU rules
Copy & Paste into the box:

```
& A < E < I < O < U < B < D < F < G < H < j < K < L < M < N < P
< R < S < T < V < W < Z& A < a <<< A < aː <<< Aː < ɐ < ɐː < ɑ < ɑː <
ɒ < ɒː < æ < æː& E < e <<< E < eː <<< Eː < ə <<< əː& I < i <<< I < iː
<<< Iː < ɯː& O < o <<< O < oː <<< Oː < oa̯ <<<ɔ<<< ɔː <<< ɵ
<<< ø <<< øæ̯& U < u <<< U < uː <<< Uː < ʉ <<< ʉː <<< y& B < b
<<< B& D < d <<< D& F < f <<< F& G < g <<< G < ɣ& H < h <<< H
< x <<< X < xʷ <<< Xʷ& K < k <<< K < kʰ <<< Kʰ < kʷ <<< Kʷ < q
<<< Q& L < l <<< L < lʲ <<< Lʲ < ɬ <<< ɬʲ& M < m <<< M& N < n
<<< N < nʲ <<< Nʲ < ŋ <<< Ŋ& P < p <<< P& R < r <<< R& S < s
<<< S < sʲ <<< Sʲ < ʃ& T < t <<< T < tʲ <<< Tʲ < tsʲ <<< Tsʲ < tʃ& V
< v <<< V& W < w <<< W < β& Z < z <<< Z
```

2. Language

2.1. The meta language for glossing is English,

therefore the glosses are in English as well as the translations in the lexicon entries. Translations into other languages (German, Russian, Hungarian) are optional.

2.2. In the lexicon, every translation has to be put correctly into the corresponding translation field,

i.e. English translations are to be written into the field for English translations, Russian into the Russian field, etc.

Figure 1 Sample entry with several translations

2.3. It is important that the language of the translations matches with the language preset in the corresponding fields.

2.4. In texts & words, the language preset in the taskbar on top of the screen must match the language of the translation line in the interlinear layers, too.

If only one language has been chosen for translation, the taskbar field remains inactive and has not to be edited. If there is more than one language of translation, this has to be checked thoroughly as there will be major errors in the export of the texts if there is any mismatch.

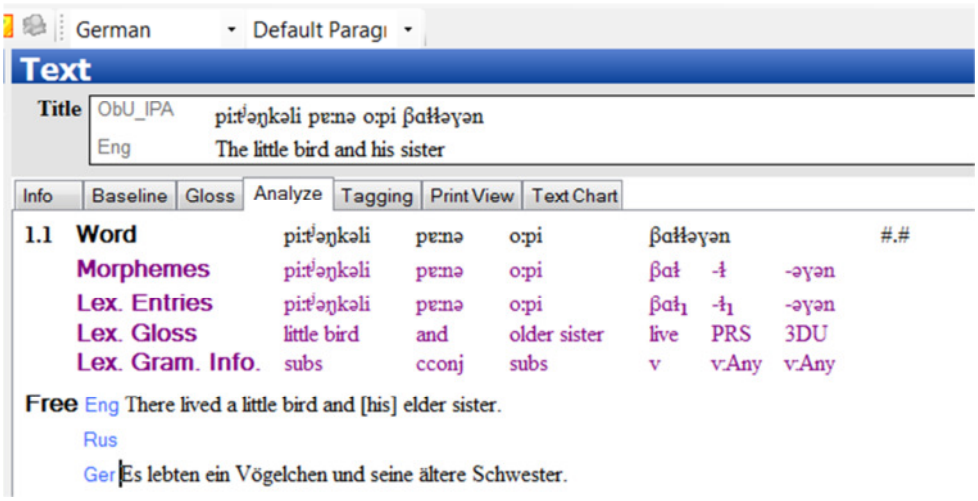


Figure 2 Matching languages in taskbar and translation line

To alter the language in the taskbar field if there is a mismatch, mark the whole line first and then adapt he language in the taskbar; sometimes it is necessary to do this with all lines of translations, even if only one line has a mismatch.

2.5. Words of foreign origin in corpus texts (code-switching)

If there is a word of foreign origin in the text (e.g. a Russian term is used instead of the correct Khanty or Mansi term), it will be glossed and translated, too. This applies mainly to conjunctions or similar fillers and is often found in recorded texts. Adapted foreign words or loanwords (e.g. words which are found in Khanty or Mansi dictionaries) are glossed as usual.

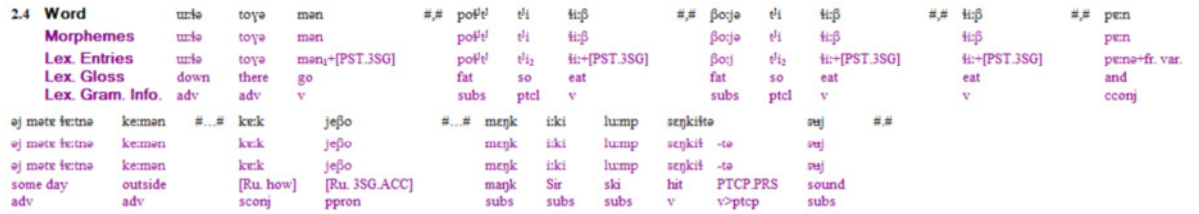


Figure 3 Words of foreign origin in Surgut text sample

In consequence, the foreign word gets an entry in the lexicon, too. Every translation is put into square brackets including the information of the foreign language. Part of speech information is according to the language of origin.

boːtʃkɐ	boːtʃkɐ	[Ru. barrel]	Substantive	[Ru. бочка]	[Ru. Fass]	[Ru. hordó]
jəʃʃo	jəʃʃo	[Ru. still]	Particle	[Ru. ещё]	[Ru. noch]	[Ru. még]
nɐː	nɐː	[Ru. take]	Particle	[Ru. на]	[Ru. nimm]	[Ru. tessék]
ʃtobɯː	ʃtobɯː	[Ru. so that]	Subordinating conjunction	[Ru. чтобы]	[Ru. um zu]	[Ru. azért, hogy]
tʲeːbe	tʲeːbe	[Ru. 2SG.DAT]	Personal pronoun	[Ru. 2SG.DAT]	[Ru. 2SG.DAT]	[Ru. 2SG.DAT]
βoːt	βoːt	[Ru. here]	Particle	[Ru. вот]	[Ru. nun]	[Ru. nos]
βsjo	βsjo	[Ru. that's all]	Particle	[Ru. всё]	[Ru. das ist alles]	[Ru. ez minden]

Figure 4 Lexicon entries of words of foreign origin, SK

The same applies for other registers of speech that are notable, e.g. child language: put a note behind the word in square brackets, e.g.

Eng fish [child lang.]

Ger Fisch [Kindersprache]

N.B.: we do not mark Code-Switching in the translations, since it is clearly marked in the glossing. If there is whole sentences in a foreign language within the text, however, we mark them in the translation.

2.6. Unknown senses

If we cannot find out the sense of a word, it will be glossed with [n.n.] in the lexicon, part of speech is an assumption derived from its role in the sentence. In the translation, we mark the unknown word with [n.n.], too.

Entries

Headword	Lexeme Form	Glosses	Grammatical Info...	Glosses (Rus)	Glosses (Ger)	Glosses (Hun)
Show All	Show All	n.	Show All	Show All	Show All	Show All
tʲiːntɬit	tʲiːntɬit	[n.n.]	Adjective	[n.n.]	[n.n.]	[n.n.]

Figure 5 Lexicon entry with unknown sense, PA

3. Corpus-based working

In order to work as corpus-based as possible, it must be guaranteed, that only those forms are part of the lexicon that occur in the texts. Therefore please pay attention to the following:

- Only note those senses in the lexicon that are used in the texts. If more senses for one lemma can be found in the dictionary, do not enter them in our lexicon if they are not used in the texts.
- The same applies to words that are not used in the texts (so-called empty head entries). We only create empty head entries if we e.g. deal with complex word forms (such as prefix verbs). In these cases it might be necessary to enter words which do not occur in the texts individually in order to state all components of a complex word form. It might also be necessary to create empty head entries for variants. If a verb only occurs e.g. in the third person singular past in the texts, and this form is a variant, we have to create an entry for the basic verb form, too, even it is not used in this form in the texts. Try to keep the amount of empty head entries small and use them only if necessary.

4. **There are several layers of language analysis in FLEx 8:**
   - Word (automatically filled with the input on the baseline)
   - Morphemes (segmentation of words into morphemes)
   - Lex. Entries (the lexemes according to their entry in the lexicon, e.g. indication of spelling variants)
   - Lex. Gloss (the actual glossing, generally rendered by abbreviated grammatical category labels)
   - Lex. Gram. Info. (information on part of speech)
   - Free (Translations)

1.1	Word	e:kʷa piɣrisʲ	a:kʷentəl			o:lcəɣ			#.#
	Morphemes	e:kʷa piɣrisʲ	a:kʷ	-cn	-təl	o:l	-c	-əɣ	
	Lex. Entries	e:kʷa piɣrisʲ	a:kʷ	-cn$_2$	-əl	o:l$_2$	-c:ɣ	-ɣ	
	Lex. Gloss	E:kʷa Piɣrisʲ	aunt	SG<3DU	INST	live	PRS	3DU	
	Lex. Gram. Info.	nprop	subs	n:Any	n:Any	v	v:Any	v:Any	

Free Eng E.P.'s been spending life with his aunt.

Ger Es lebten E:kʷa piɣrisʲ und seine Tante.

Figure 6 Example FLEx 8 layers of analysis (NM)

5. **The glossing rules are based on the Leipzig Glossing Rules**

(version 08-02-05 http://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf) which are modified corresponding to the needs of Ob-Ugric languages and the data handling in FLEx 8.

6. **For the glossing abbreviations capital letters are used (SG, DLAT …).**

A list of abbreviations is given at the end of this document.

If the preset abbreviations do not suit our abbreviations are missing, please set up new ones and discuss them with the coordinators of all dialects; it is always possible, to add abbreviations so you don't have to set up all before you start to work. This keeps our glossings as corpus-based as possible.

7. **All in all, glossing in FLEx follows a word - by -word alignment**

with a morpheme-by-morpheme correspondence (there must be exactly the same number of segments in the example and in the gloss).

8. **Glossing Characters and formal conventions**

Generally, we should analyze and separate by hyphens as many morphemes as possible. Ideally, there should be a 1:1 correspondence of segmented morpheme and grammatical information.

8.1. The break characters for the analysis which are integrated in FLEx are

- the hyphen - [Unicode 002D] for all kind of affixes (prefix, suffix, circumfix, infix) and
- the equal sign = [Unicode 003D] for all kind of clitics (enclitic, proclitic)

The break character is always grouped with the bound morpheme/clitic, not with the stem. Please apply this rule in dependence from the status as prefix/proclitic resp. as suffix/enclitic:

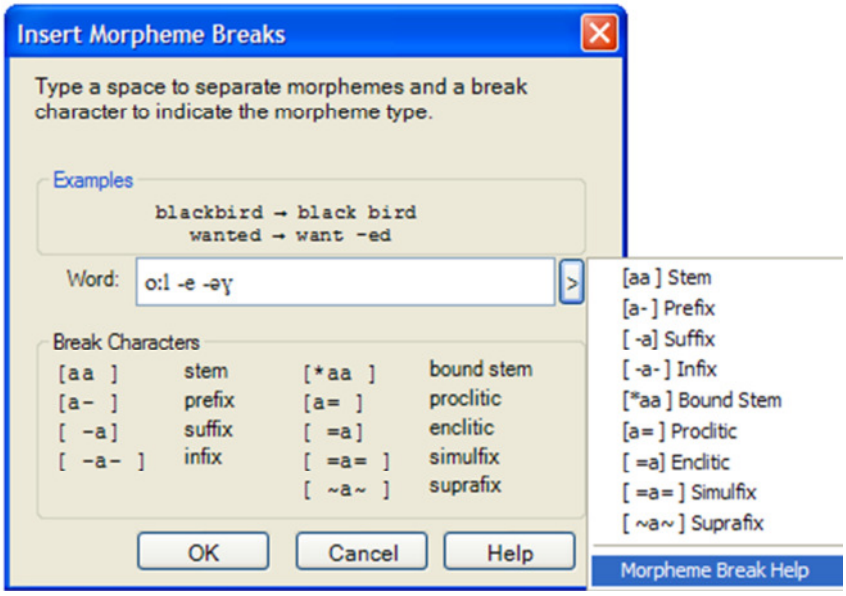


Figure 7 Morpheme break box FLEx 8 with possibilities for separating morphemes

8.2. Person and number are always separated by hyphens within the gloss if possible.

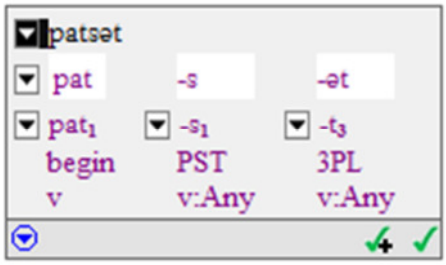


Figure 8 Person and number glossing, NM

8.3. If a single object-language element represents a combination of several meta language categories, these will be separated by a full stop.

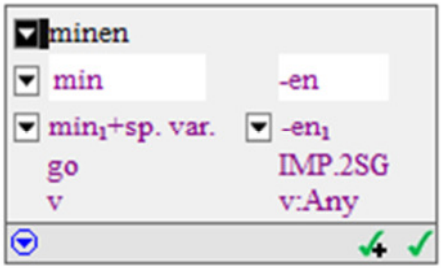


Figure 9 Glossing of several meta language categories, NM

The order of the abbreviations corresponds to the general order of categories/suffixes within the dialects in cases when the affixes can be separated transparently. Participles, for instance, are glossed: PTCP.PST and PTCP.PRS

8.4. The symbol "<" is used in the glossing of verbal personal markers encoding simultaneously subject and direct object (objective conjugation) and of nominal personal markers encoding two referents (possessive suffixes).

The encoding of both referents is asymmetric:

- person and number of subject and number of direct object
- person and number of the referent generally referred to as *possessor* and number of the referent generally referred to as *possessum*;
- The order of the glossing components is direct object < subject, resp. *possessum* < *possessor*

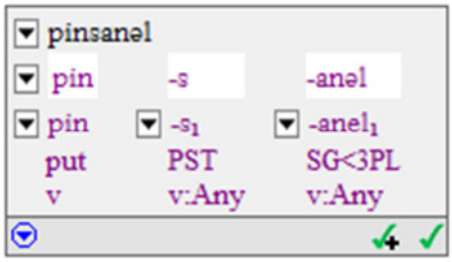


Figure 10 Glossing of objective conjugation, NM

8.5. If an object-language element cannot be segmented formally or semantically, and the meta language lacks a single-word equivalent, use several words, separated by blanks.

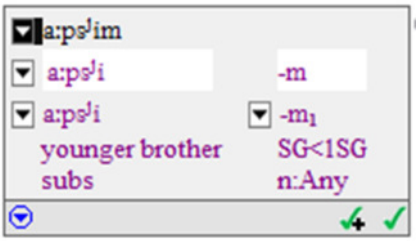


Figure 11 Glossing of translations consisting of several lexemes, NM

8.6. Grammatical information which is not overtly expressed (zero morphemes), is displayed in square brackets [ ], i.e. information expressed by zero is bracketed; the bracket is attached to the preceding gloss.

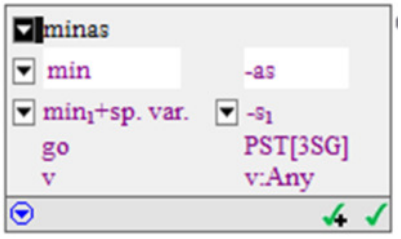


Figure 12 Glossing of zero morphemes, NM

8.7. Punctuation characters used in Baseline

Punctuation character	Unicode (hex)	To use in following cases
.	002E	Ad: if you need the triple point at the end of a sentence, please type this character triply
,	002C	
:	003A	WARNING: don't mix up with ː
;	003B	
-	002D	WARNING: don't mix up with –
–	2013	please use it when text is interrupted
!	0021	
?	003F	
…	2026	WARNING: use it only in the middle of a sentence, never at the end!
#	0023	
(	0028	Do not use brackets in texts, as FLEx cannot proceed them
)	0029	
[	005B	
]	005D	
“	201C	please use it at the beginning of the passage
”	201D	please use it at the end of the passage
«	00AB	please use it only in case of text in text at the beginning of the passage
»	00BB	please use it only in case of text in text at the end of the passage
|	007C	If a poetry text is rather structured in lines than in sentences please put the following combination of symbols at the end of each line: #|#, as a marker for line break.

8.8. Phrases consisting of several independent words

For a phrase, first click on the chain symbol which is displayed on the upper left side of the box in order to connect the two elements:

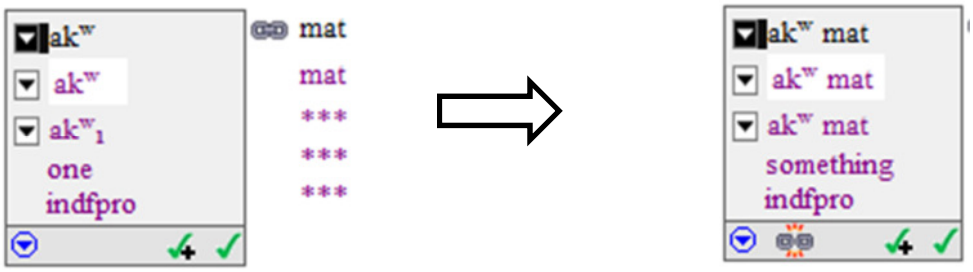


Figure 13 Connecting two words into one phrase, NM

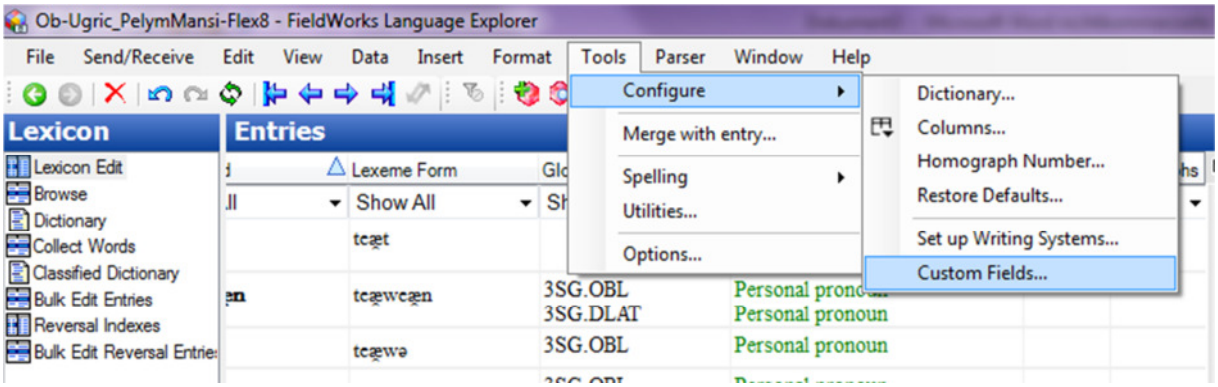
Then insert the phrase as a lexicon entry and specify the morpheme type as "phrase".

work with the texts it turned out to be helpful to take notes about where to find an entry in one of the dictionaries we are working with. Since this is valuable information, we decided to create a custom field “dictionary” which will be part of our corpus-based dictionary on the website, too.

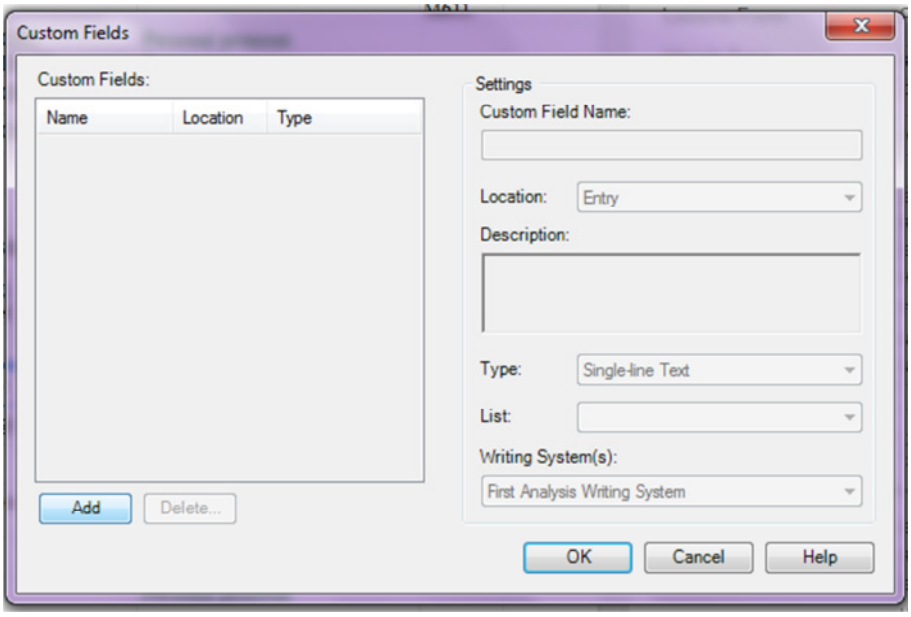
8.9. How to create a new field in FLEx Lexicon:

In the menue, click on *Tools*;

Choose *Configure* and then select *Custom Fields*:



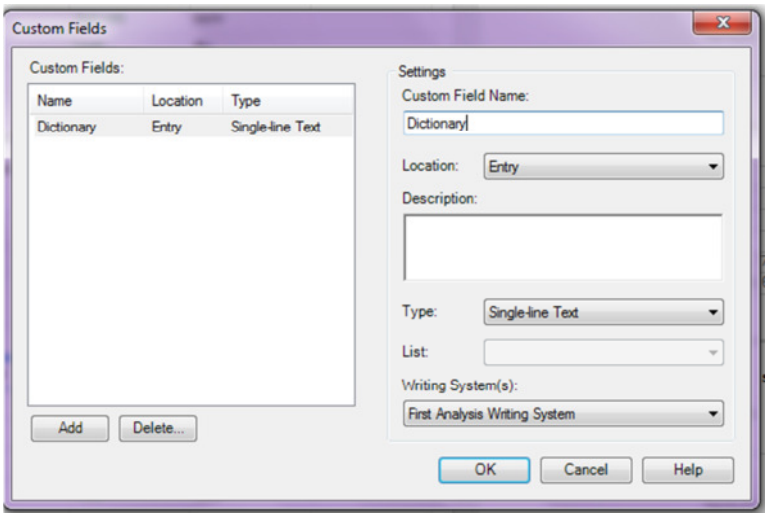
A box *Custom Fields* opens:



Click *Add* on the bottom, a new entry appears on the left side of the box.

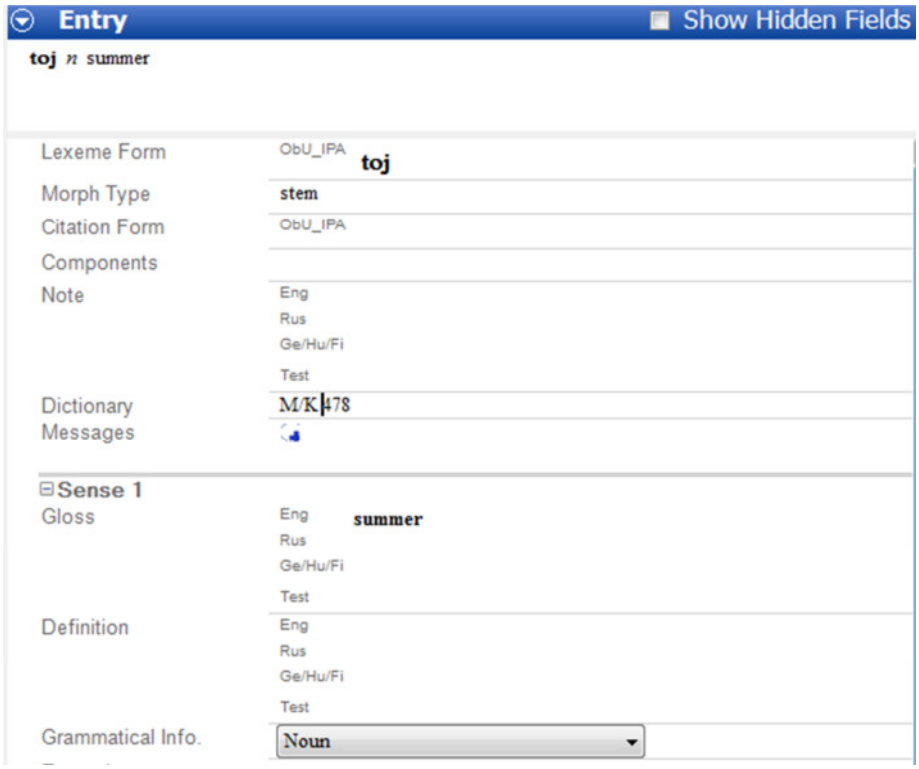
Now you can edit the new entry on the right side; fill in the name of the new field: **Dictionary**

Note: Make sure that *Entry* is selected in the field *Location*, otherwise change it with clicking on the arrow!



Click *OK*;

Every lemma entry should now show the new field:



Format

Here are the dictionaries used for reference for Mansi, e.g.:

K = Kannisto, Artturi (2013): Wogulisches Wörterbuch. Gesammelt und geordnet von Artturi Kannisto. Bearbeitet von Vuokko Eiras. Herausgegeben von Arto Moisio Lexica Societatis Fenno-Ugricae XXXV Kotimaisten kielten keskuksen julkaisuja 173. Helsinki: Suomalais-Ugrilainen Seura / Korimaisten Kielten Keskus.

M/K = Munkácsi, Bernát – Kálmán, Béla (1986): Wogulisches Wörterbuch Budapest: Akadémiai Kiadó

B/V = Баландин, А. Н. - Вахрушева, М.П. [Balandin, A. N. – Vahruševa, M. P.] (1957): *Мансийский язык. Учебное пособие для педагогических училищ* Ленинград: Учпедгиз

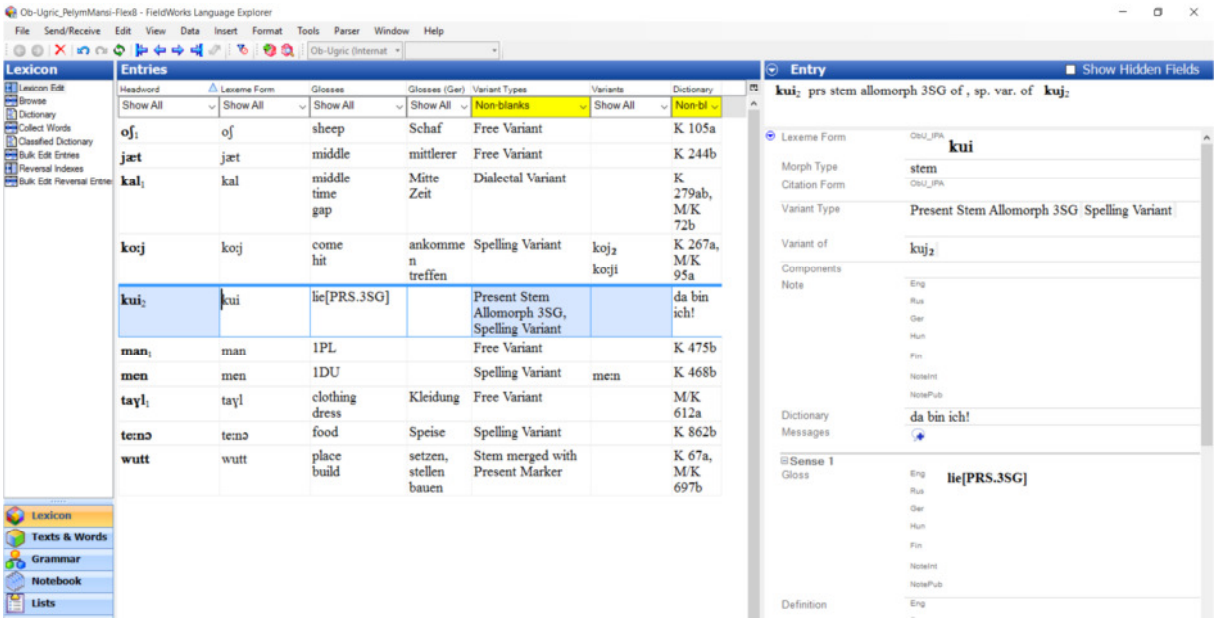
Ká = Kálmán, Béla (1963): *Chrestomathia Vogulica*. Budapest: Tankönyvkiadó, 124

Ká2 = Kálmán, Béla (1976): *Wogulische Texte mit einem Glossar*. Gesammelt und bearbeitet. Aus dem Ungarischen übersetzt von H. Krüger-Tokody und P. Kocsány. Budapest: Akadémiai Kiadó

R = Riese, Timothy (2001): *Vogul*. Languages of the World/Materials, 158 . München – New Castle: LINCOM EUROPA

The abbreviation is at the beginning of the note:	M/K
Just type the page, without any abbreviations :	478
If there is several columns on one page, use additional small letters:	M/K 478a
If you refer two more than one dictionary, use a comma between them:	M/K 478a, K 552b

If you wish to display the dictionary field in the columns on the left side, click on the tiny icon in the upper right edge of the columns.



A box *Configure Columns* opens. Choose *Dictionary* on the right side, click „ADD“ in the center and it appears on the left side of the box (the selected columns). Click OK.