Aggregating dictionaries into the language portal Sõnaveeb: issues with and without solutions

Kristina Koppel, Arvi Tavast, Margit Langemets, Jelena Kallas

issues

- quick overview of Ekilex and Sõnaveeb
- issues with and without solutions:
  - consistency of information and avoiding duplicates when unifying the dictionaries
  - **turning dictionary-specific information into customizations of the central service**
  - deciding on deliberate ambiguities
  - parsing data fields containing more than one data element, including textual condensation
  - moving from annotating form (e.g. italics) to annotating content (e.g. a citation)
  - moving from (near) duplicates to sensible information fragments
  - deciding between an app and a responsive web page
  - possible legal problems regarding the authorship of the new central resource
  - **NEW: concerns about the quantity and quality of information**

why?

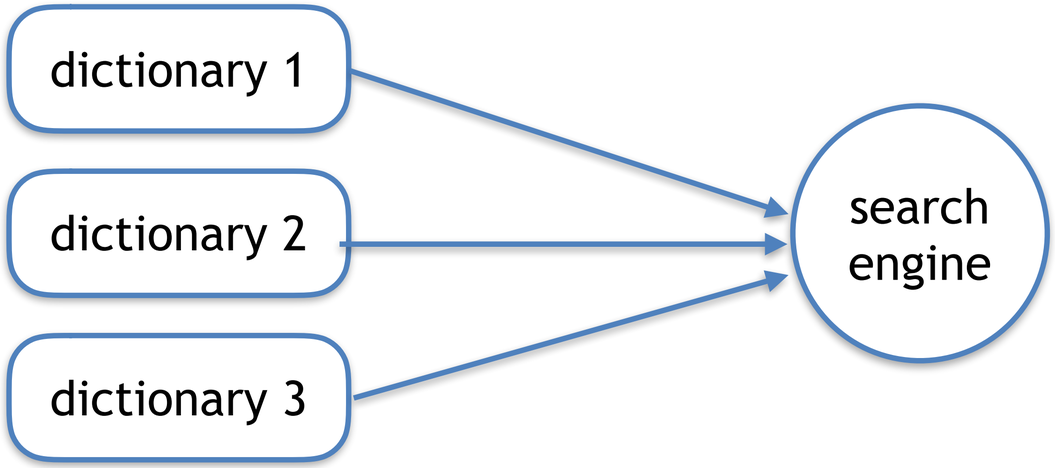
assumption: users look for information about the language, not about dictionaries

avoid duplication and conflicts

make dictionary compilation realistic

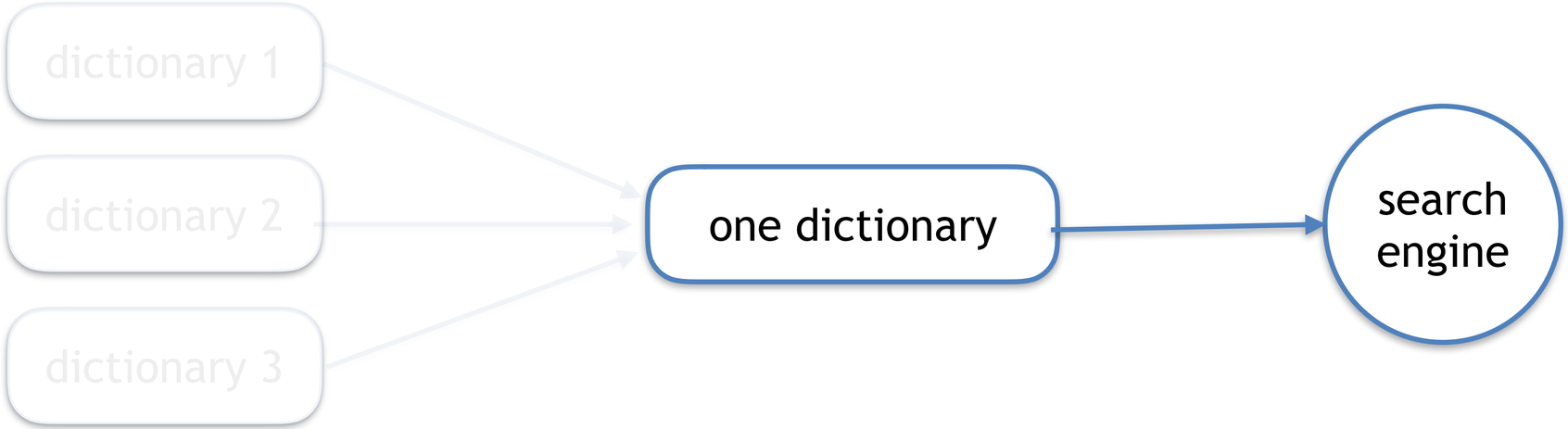
reduce confusion for users and sponsors

aggregated search



vs

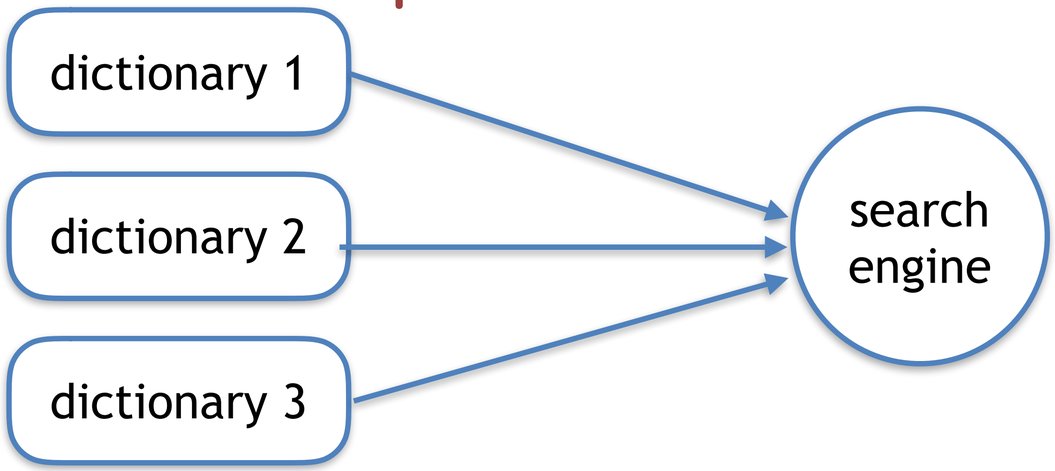
one dictionary



one dictionary

present differences to the user

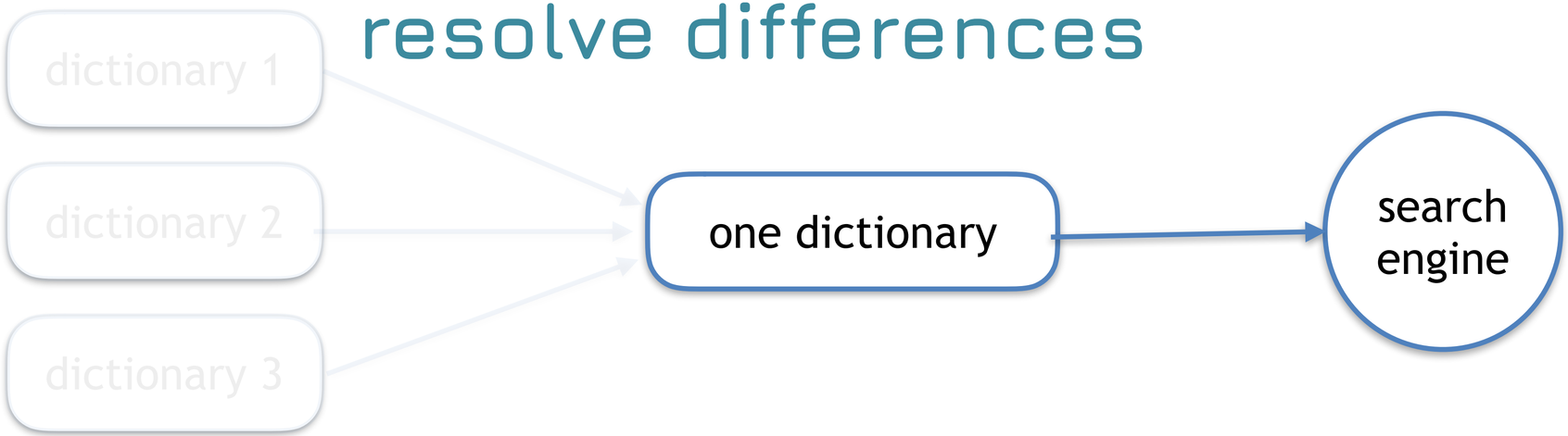
aggregated search



vs

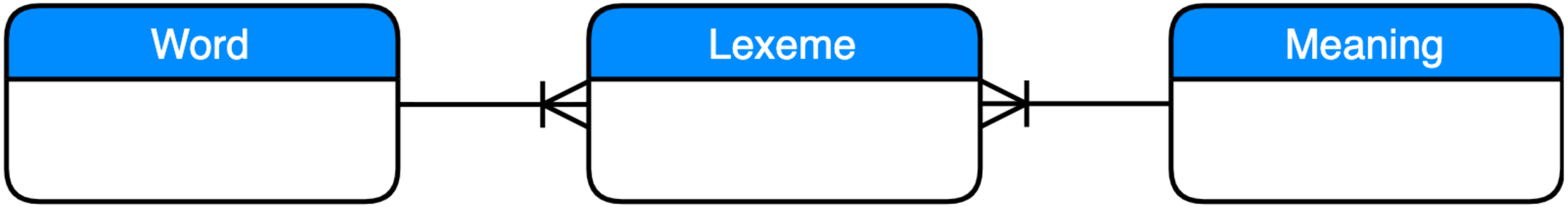
one dictionary

resolve differences

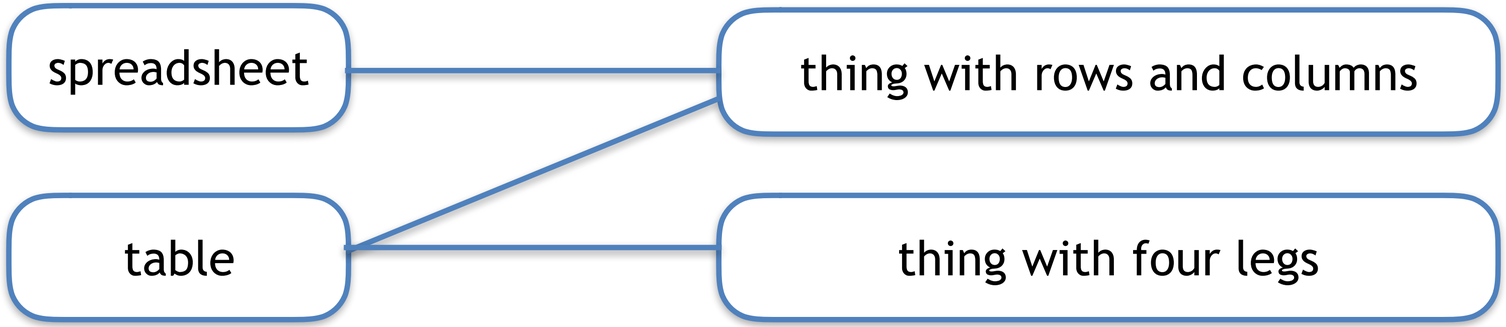


one dictionary

Ekilex data model

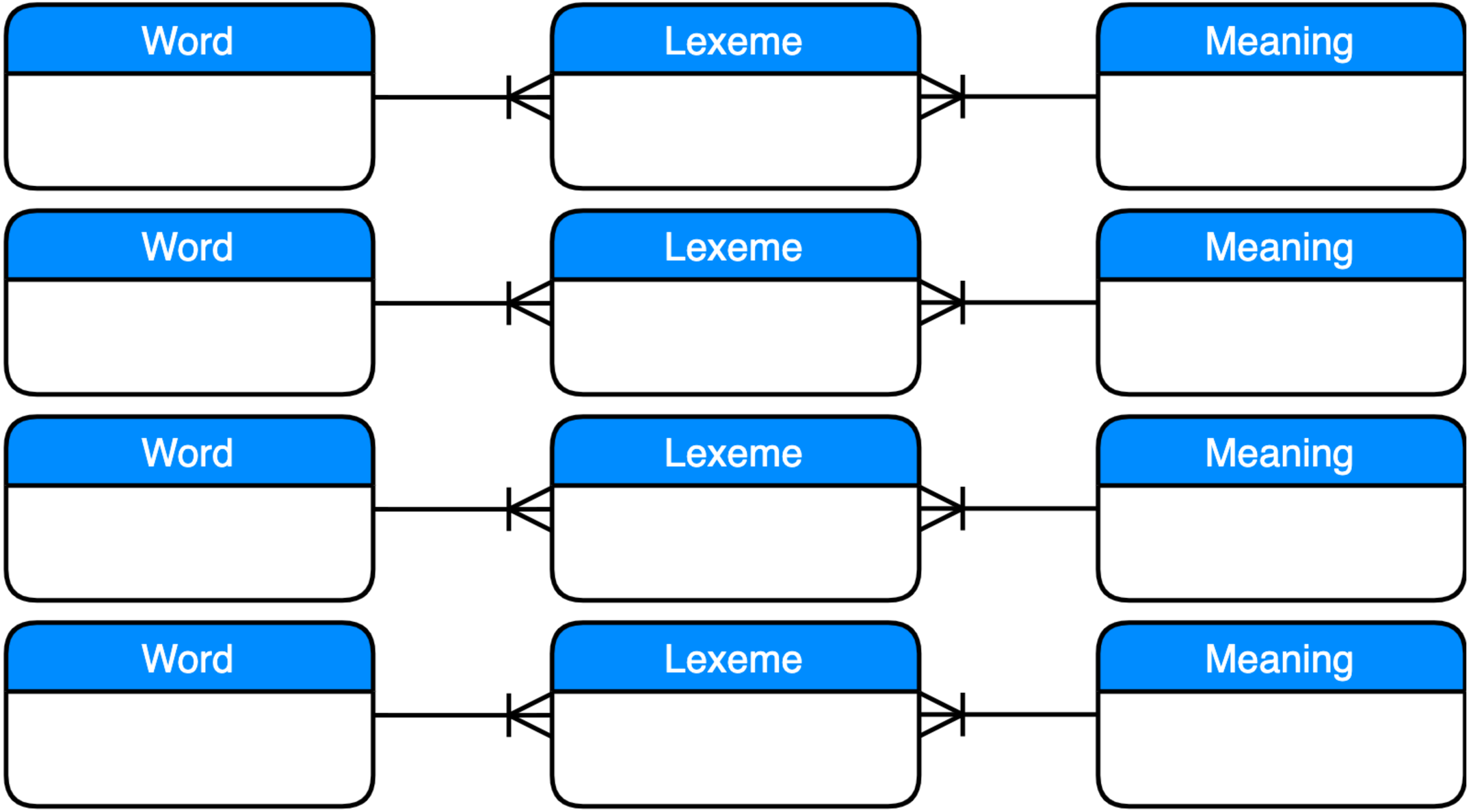


lexeme:
this word
in this meaning
as described in this dictionary



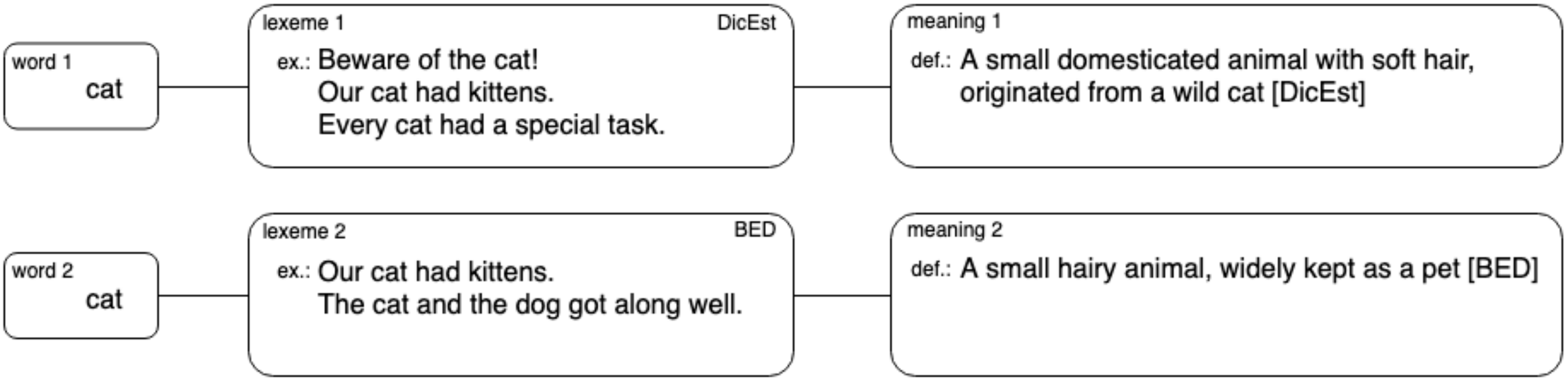
one dictionary

Previously: many dictionaries



one dictionary

one dictionary



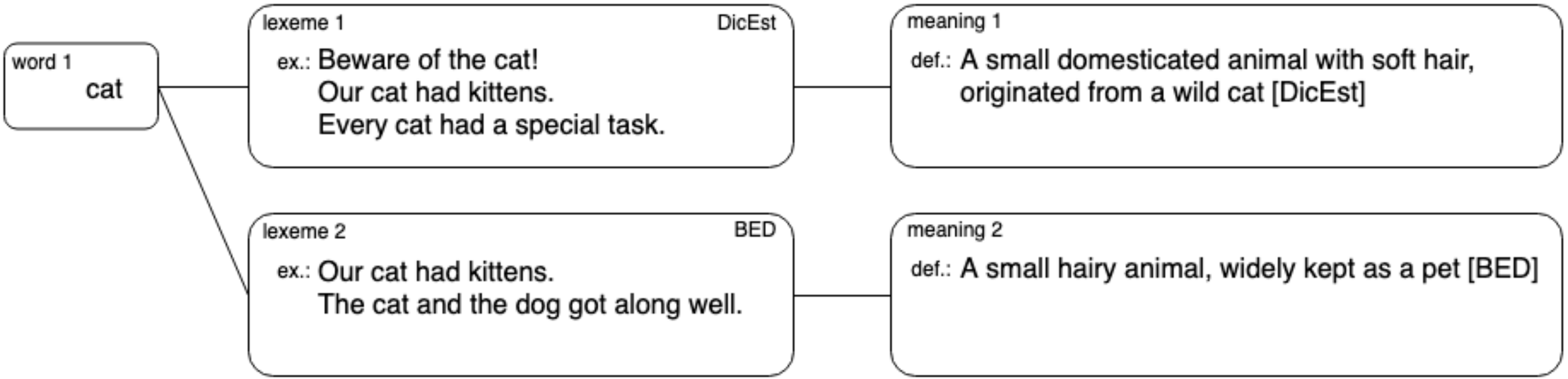
initial state:

2 words

2 meanings

2 dictionaries

one dictionary

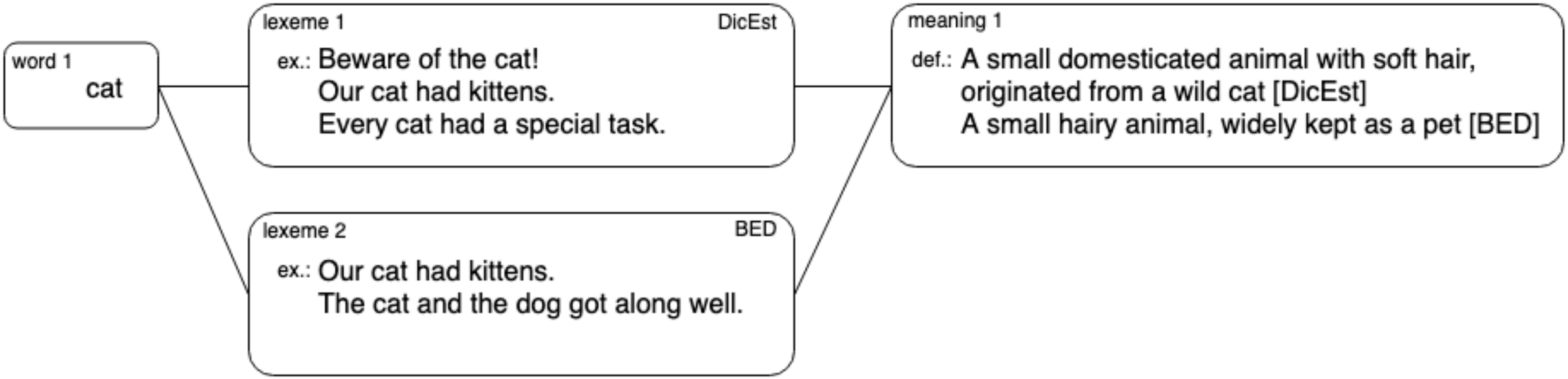


words combined: 1 word

2 meanings

2 dictionaries

one dictionary



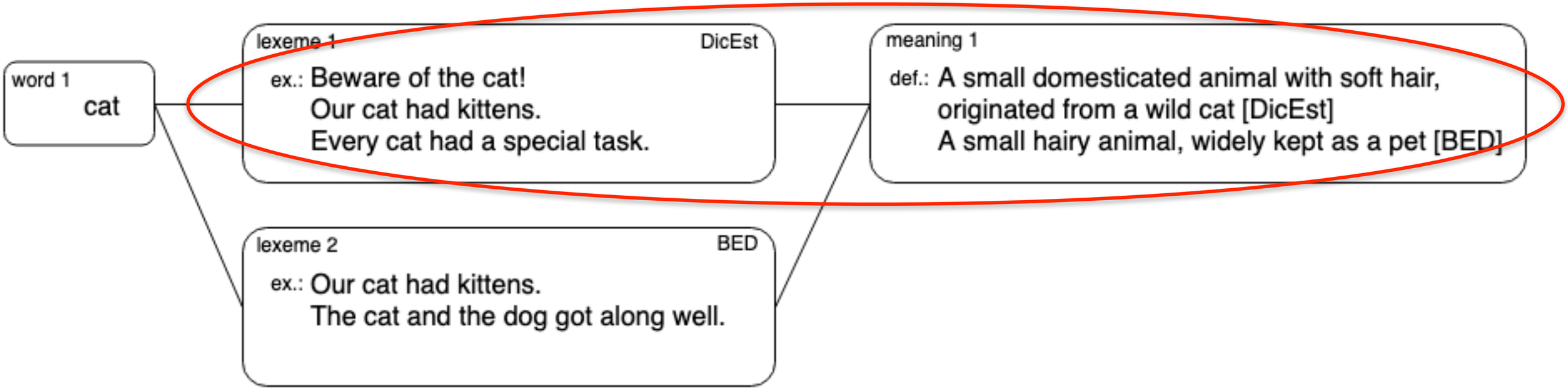
words and meanings combined:

1 word

1 meaning

2 dictionaries

one dictionary



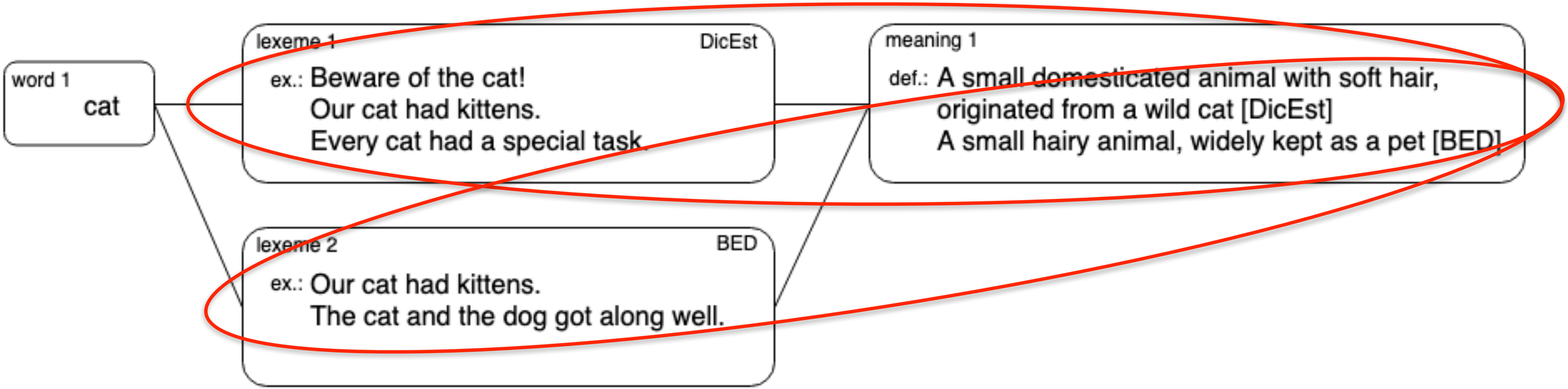
words and meanings combined:

1 word

1 meaning

2 dictionaries

one dictionary



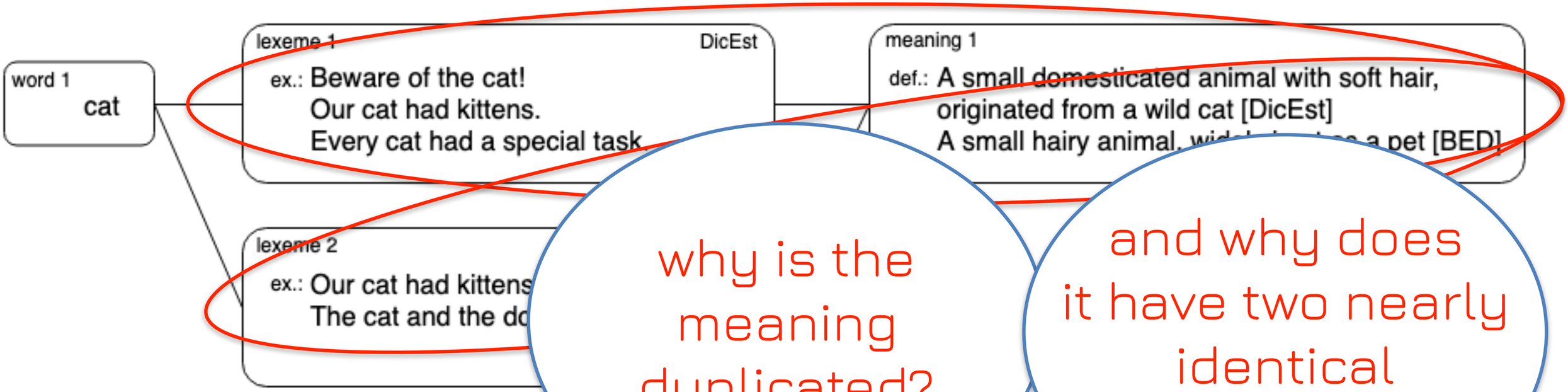
words and meanings combined:

1 word

1 meaning

2 dictionaries

one dictionary



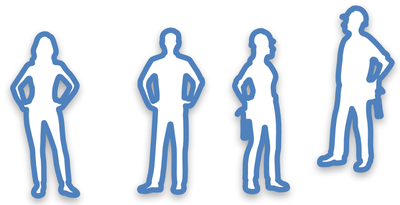
why is the meaning duplicated?

and why does it have two nearly identical definitions?

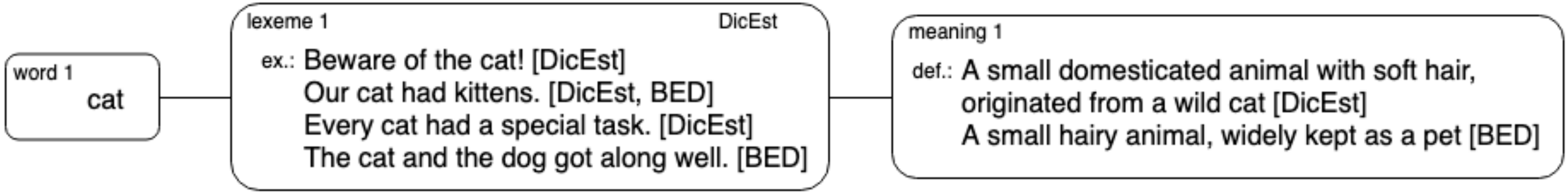
words and meanings combined:

1 meaning

2 dictionaries



one dictionary



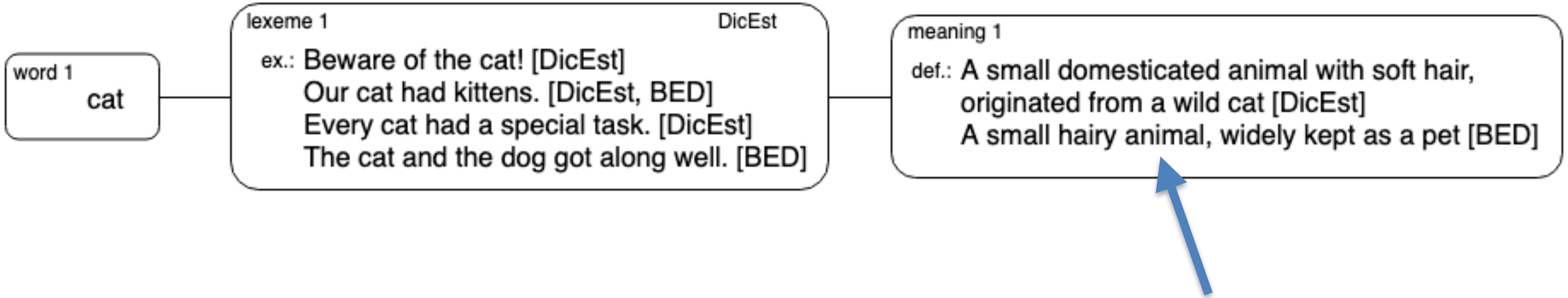
all combined:

1 word

1 meaning

1 dictionary

one dictionary



resolve these manually

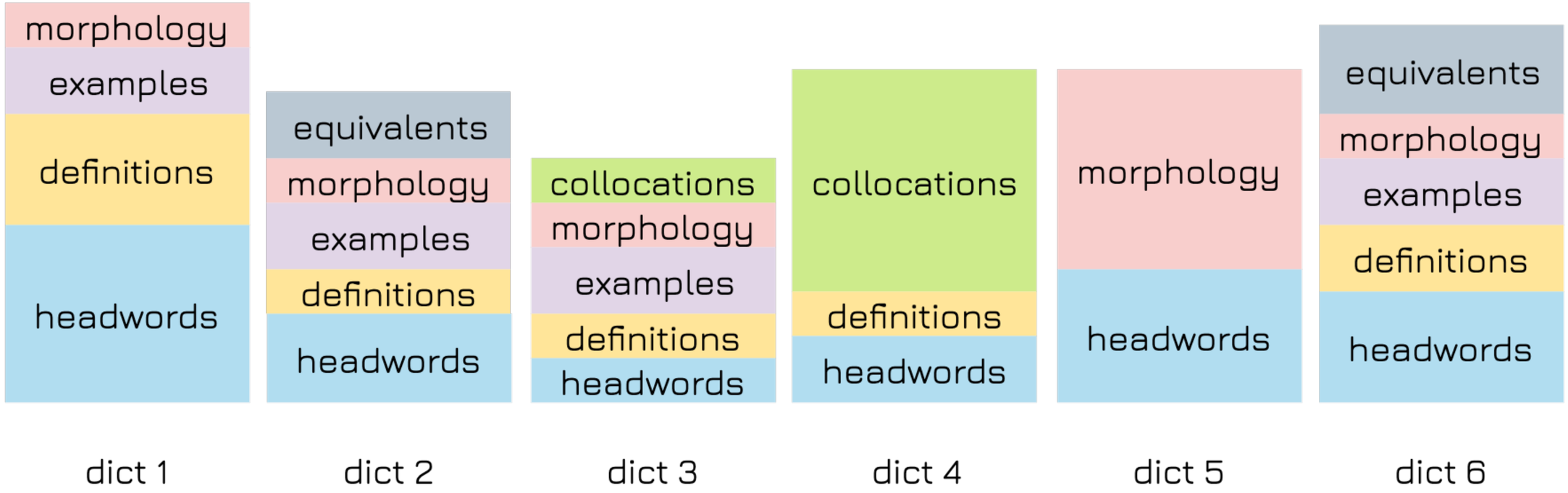
all combined:

1 word

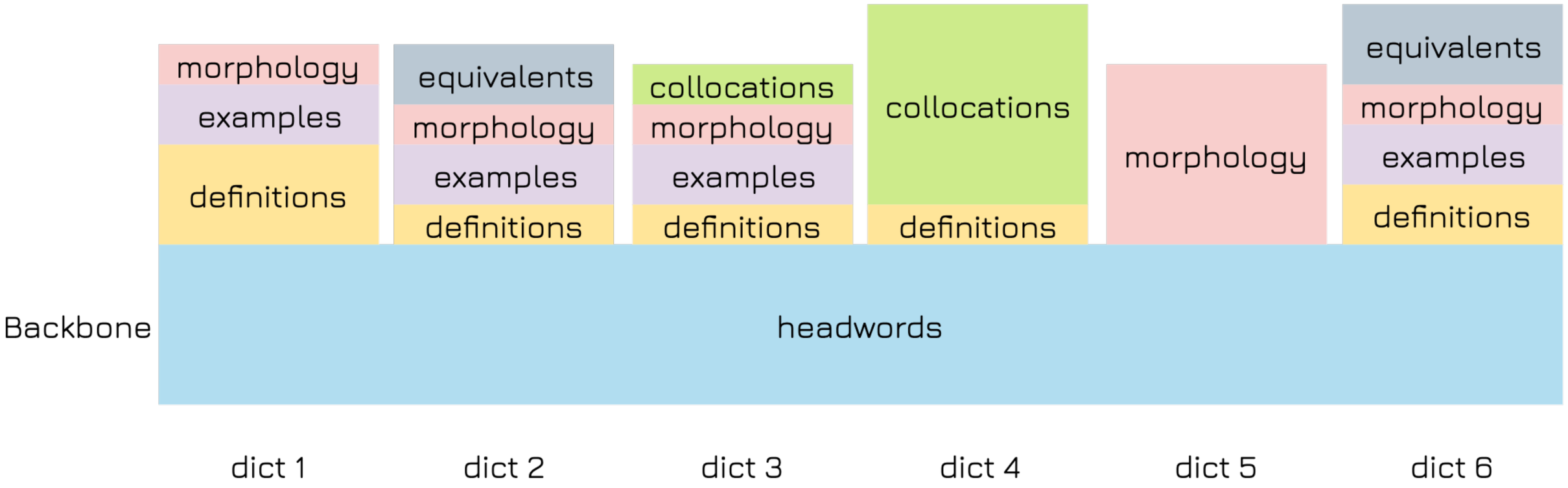
1 meaning

1 dictionary

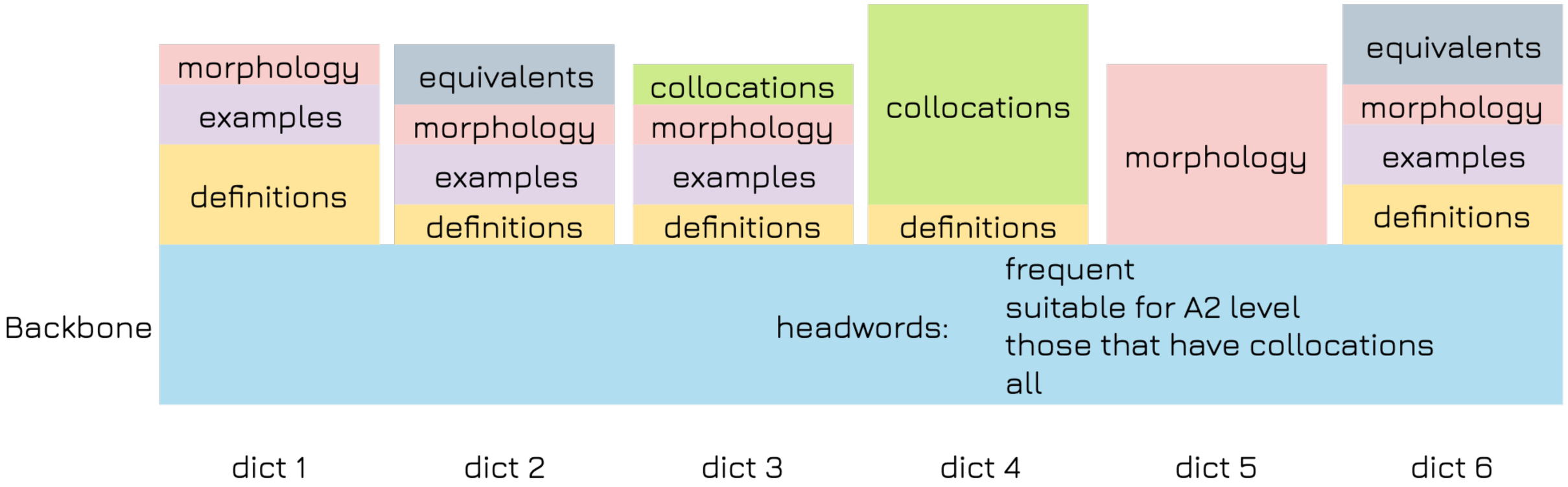
initial situation



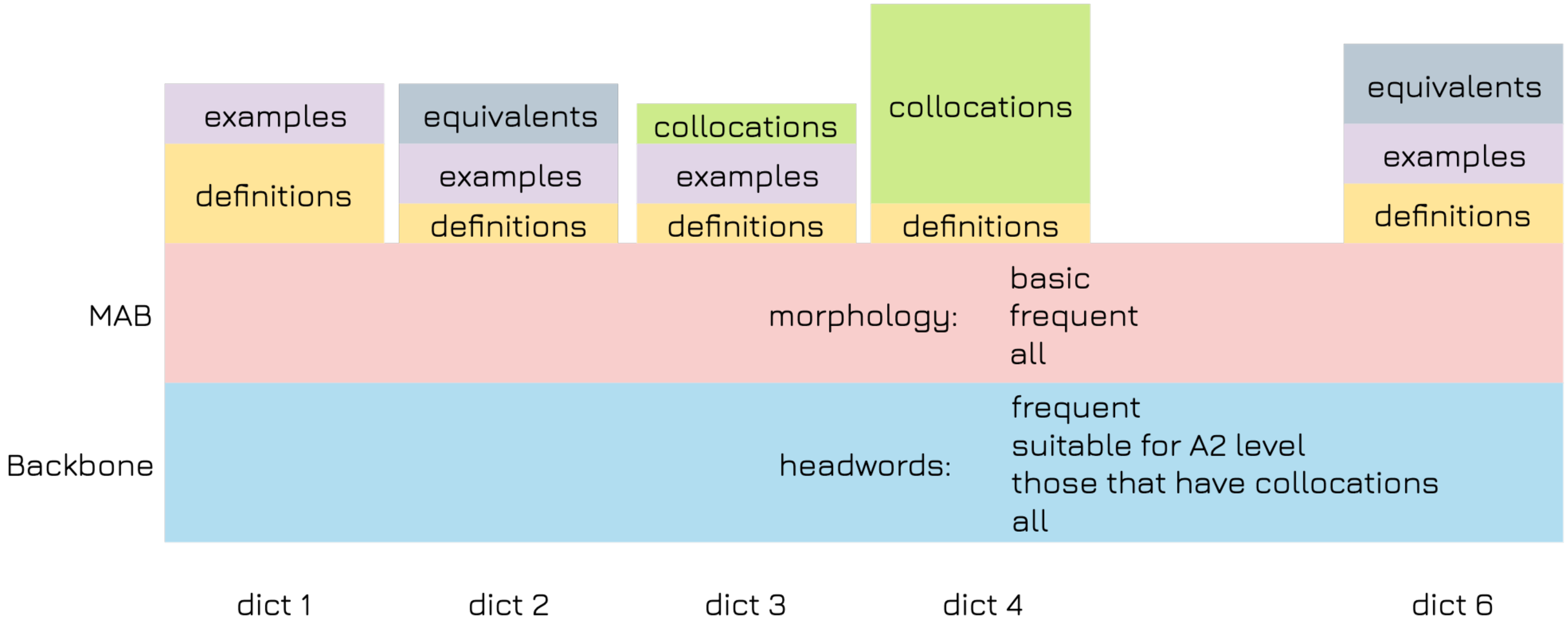
headwords aggregated



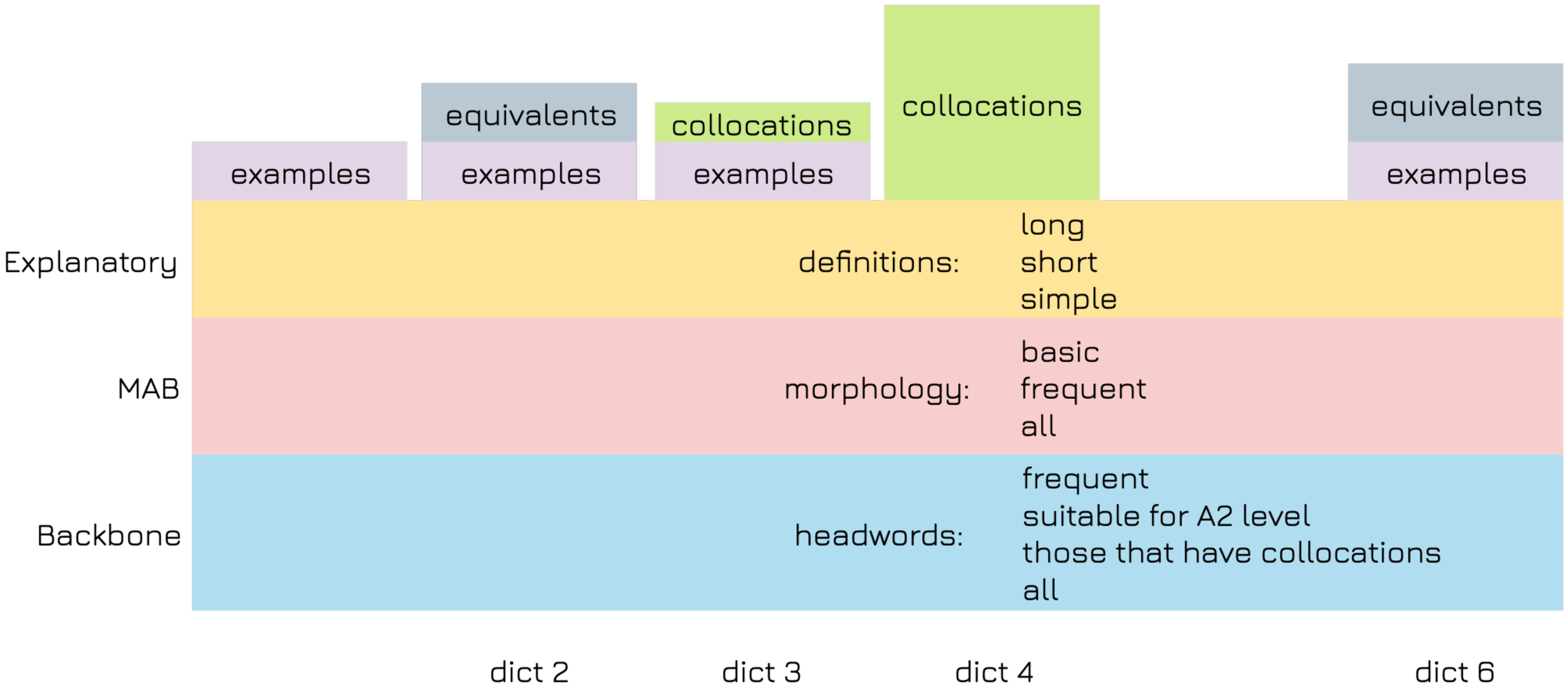
headwords aggregated



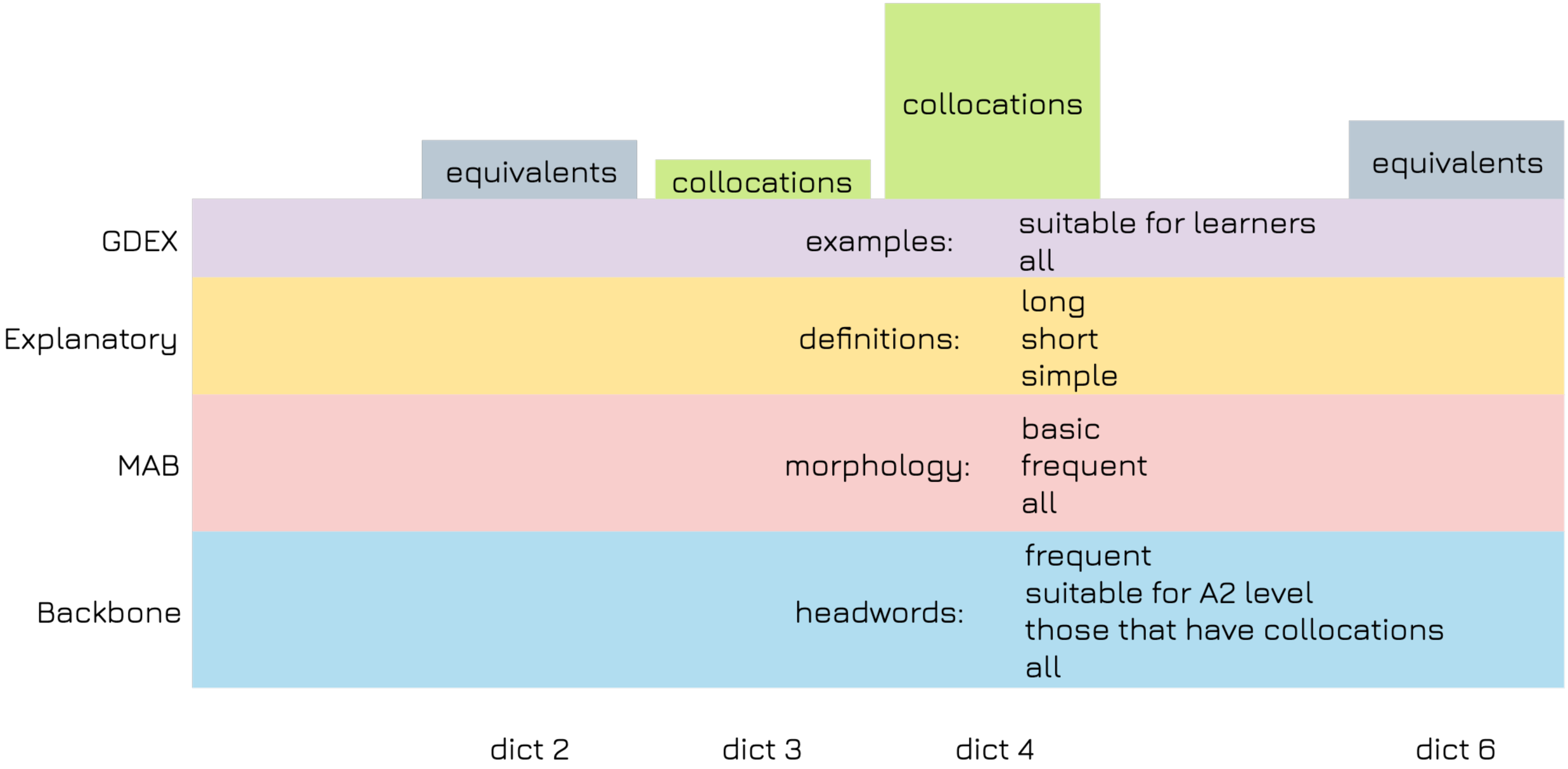
morphology centralised



meanings aggregated



examples aggregated

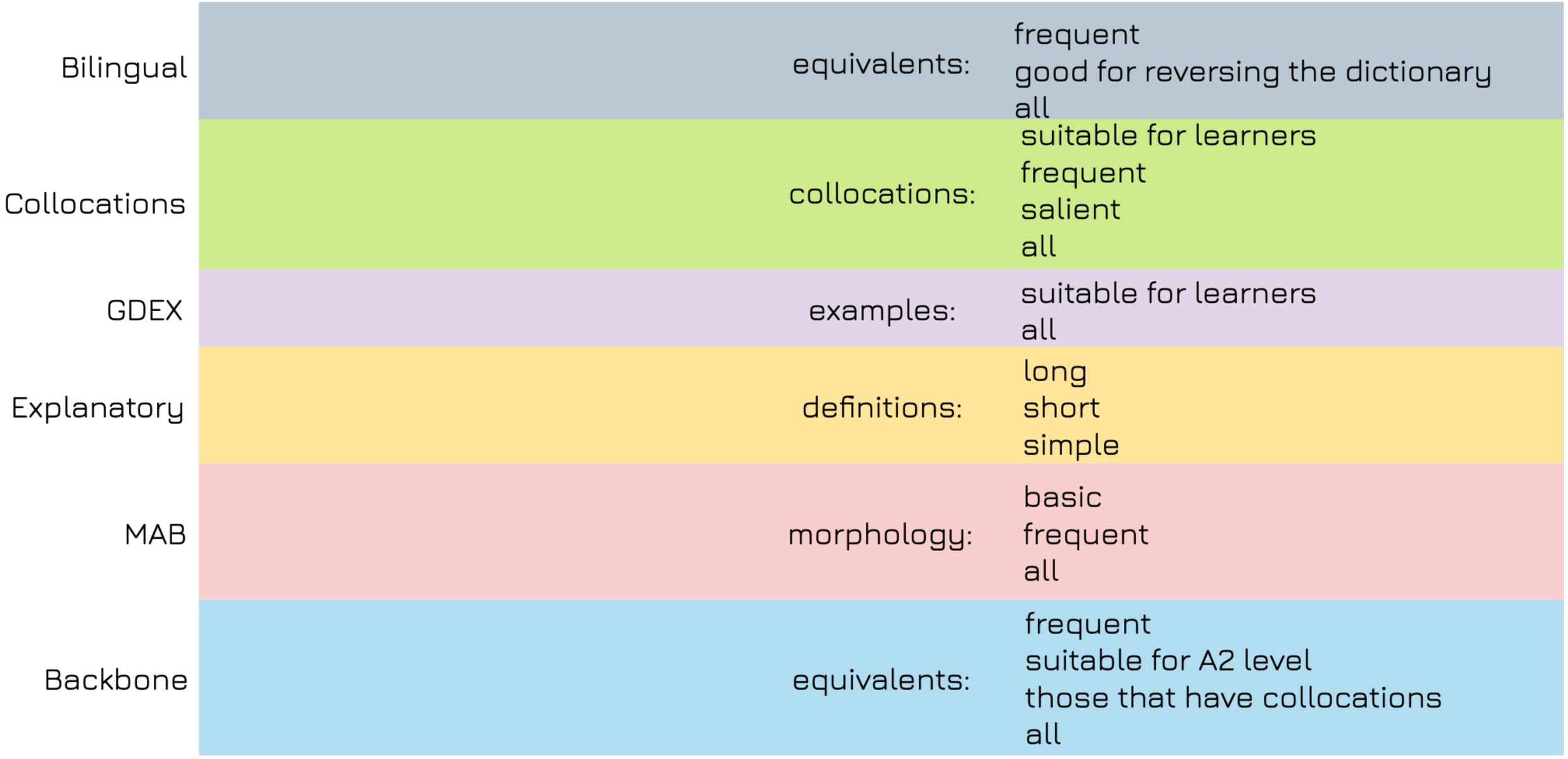


collocations aggregated (we are here now)

	equivalents		equivalents
Collocations	collocations:	suitable for learners frequent salient all	
GDEX	examples:	suitable for learners all	
Explanatory	definitions:	long short simple	
MAB	morphology:	basic frequent all	
Backbone	headwords:	frequent suitable for A2 level those that have collocations all	

dict 2

dit 6



one dictionary

corpus-based

number of headwords?

60000 as in the popular dictionary

150000 as in the largest dictionary

1 million most frequent lemmas

6 million as in the current corpus

x million as in the next corpus

corpus-based

minimum amount of information?

corpus example and frequency

definition

equivalents

manually picked examples

translations of examples

corpus-based

minimum quality of information?

errors of tokenising, tagging etc

ungrammatical

politically incorrect

not approved by language planning

noise of distributional semantics

thank you

Kristina Koppel
Arvi Tavast
Margit Langemets
Jelena Kallas

arvi@tavast.ee

This work has been partially supported by receiving funding from the European Regional Development Fund.