Vilnius University

Artificial learners of morphology

Ignas Rudaitis · Academia Salensis XV

Vilnius University

Artificial learners of morphology

a non-technical survey

Ignas Rudaitis · Academia Salensis XV

methodological template

theory of morphology

▶

theory of morphology + theory of acquisition

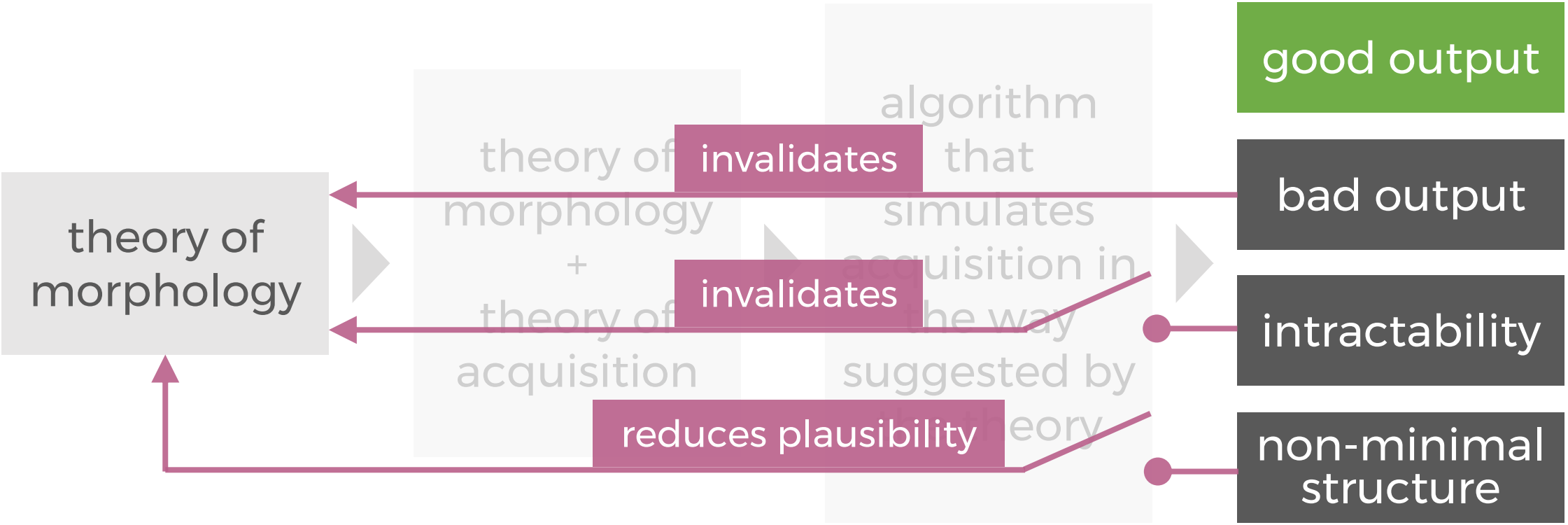
▶

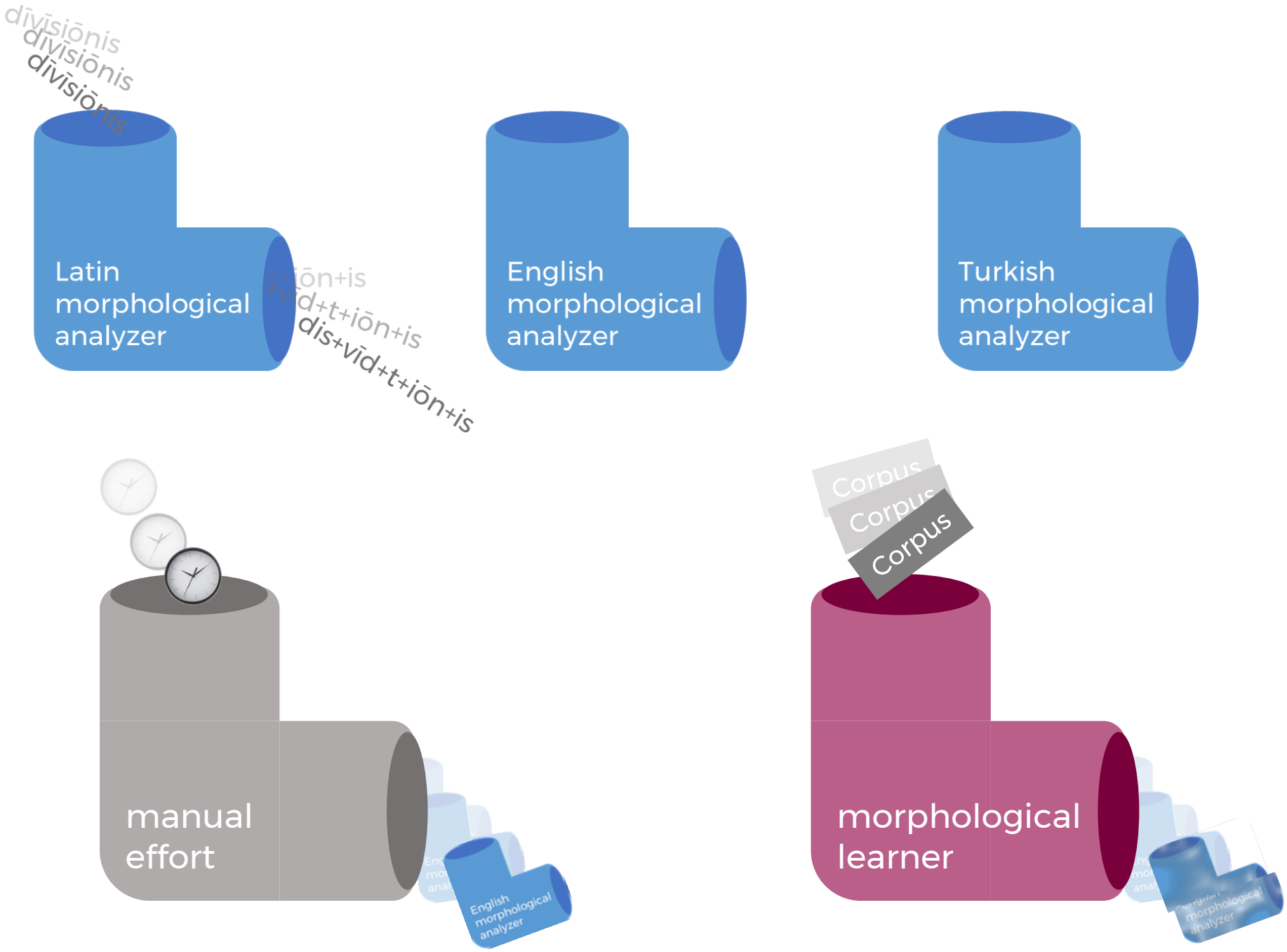
algorithm that simulates acquisition in the way suggested by the theory

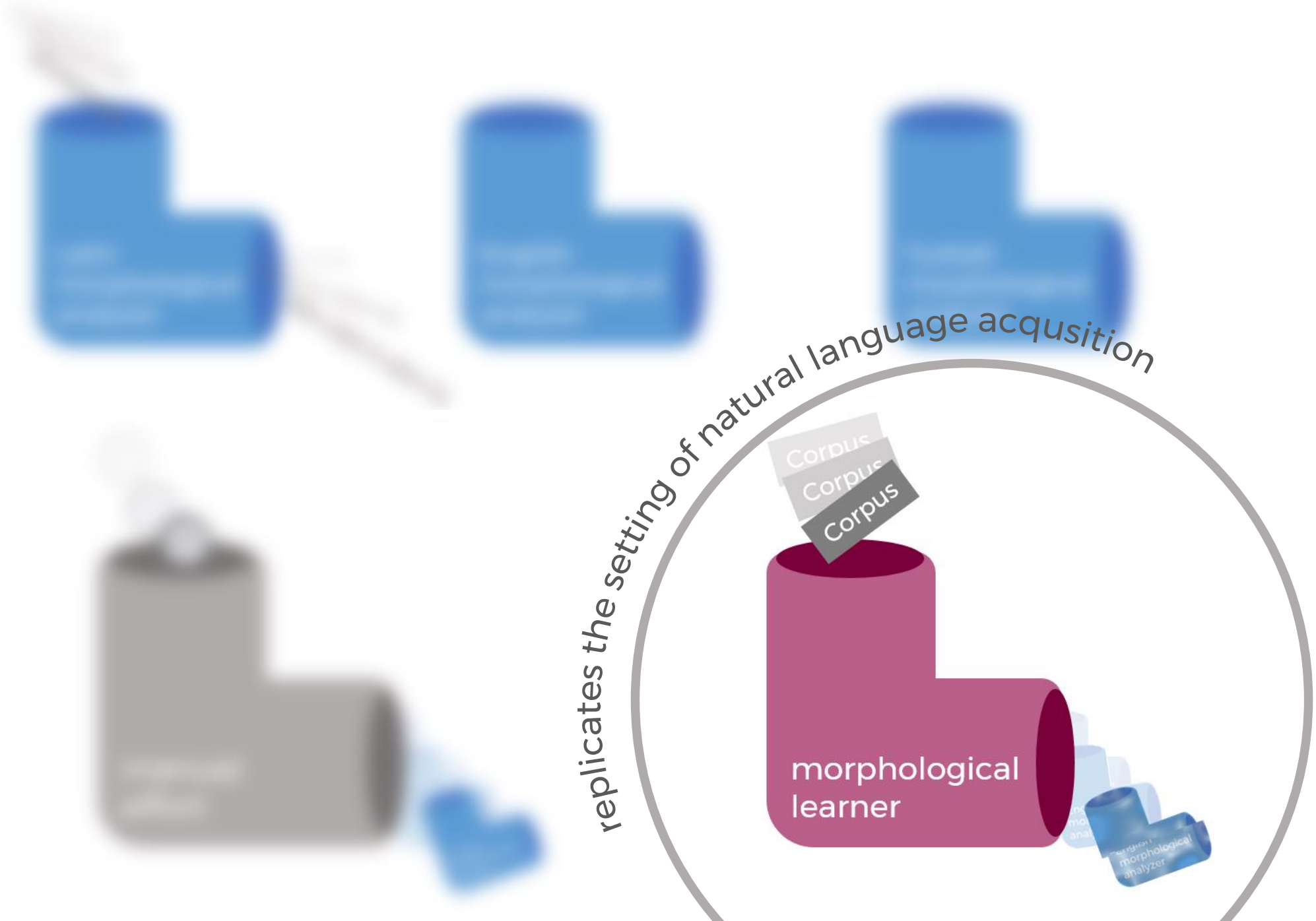
▶

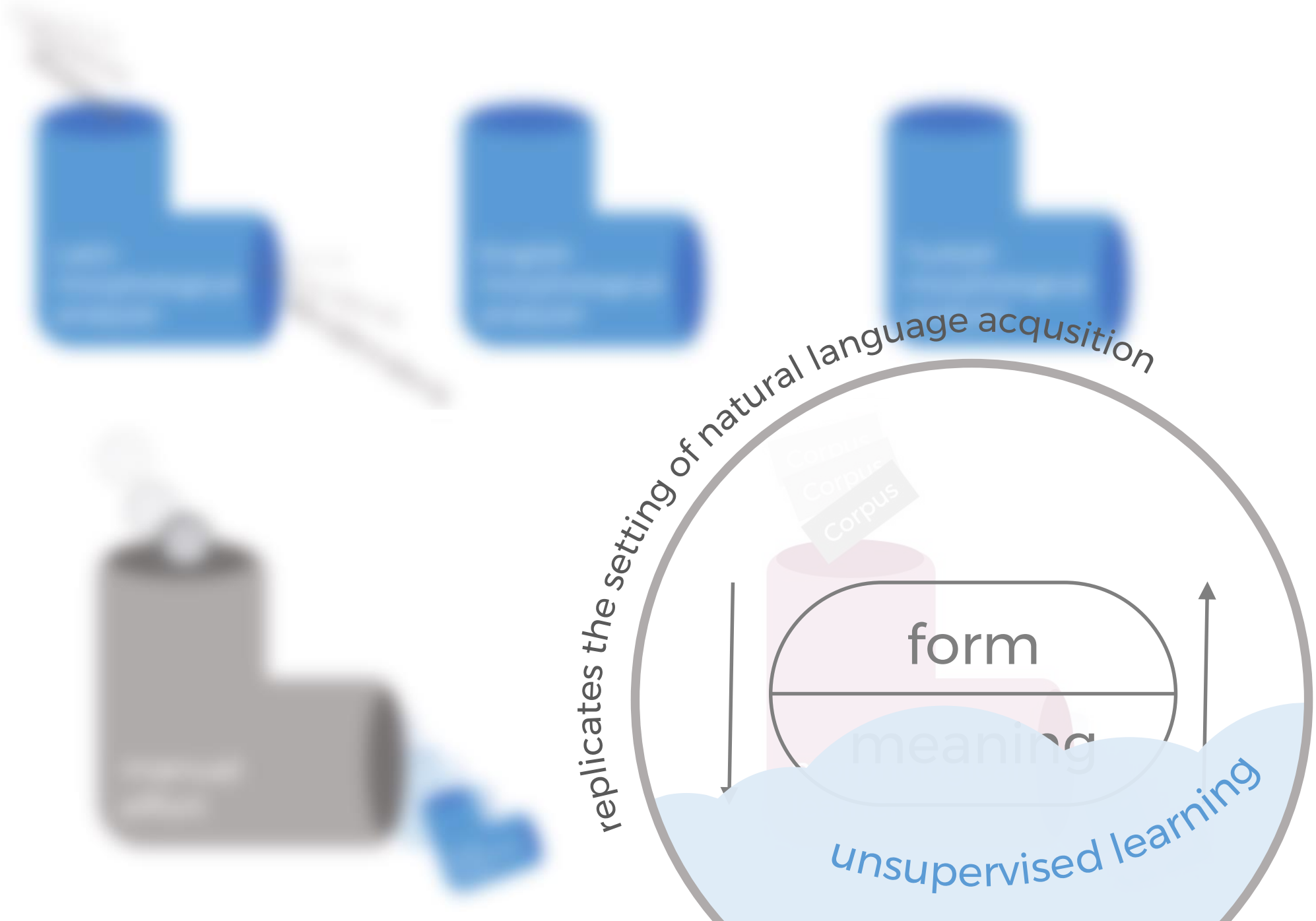
- good output
- bad output
- intractability
- non-minimal structure

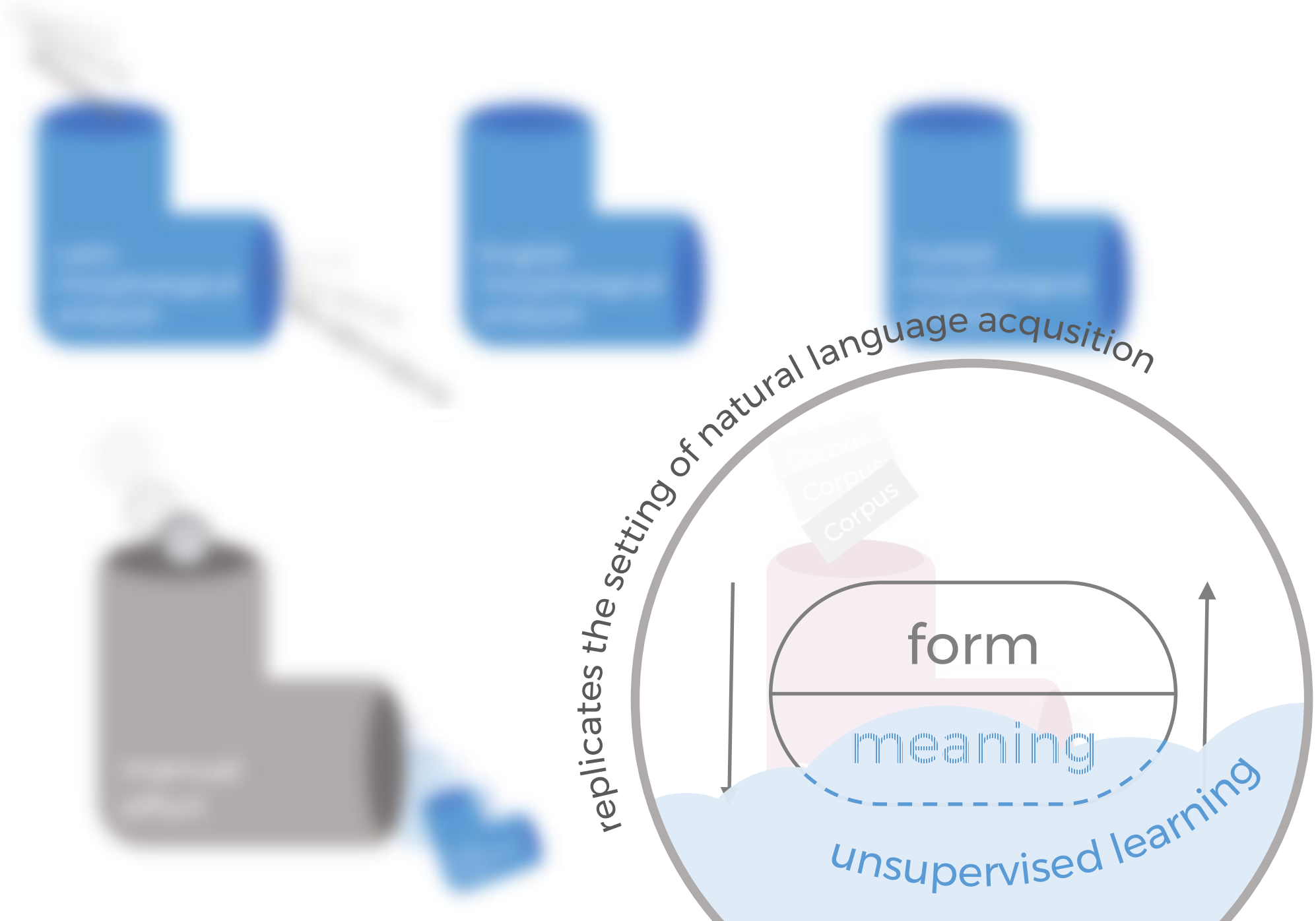
methodological template

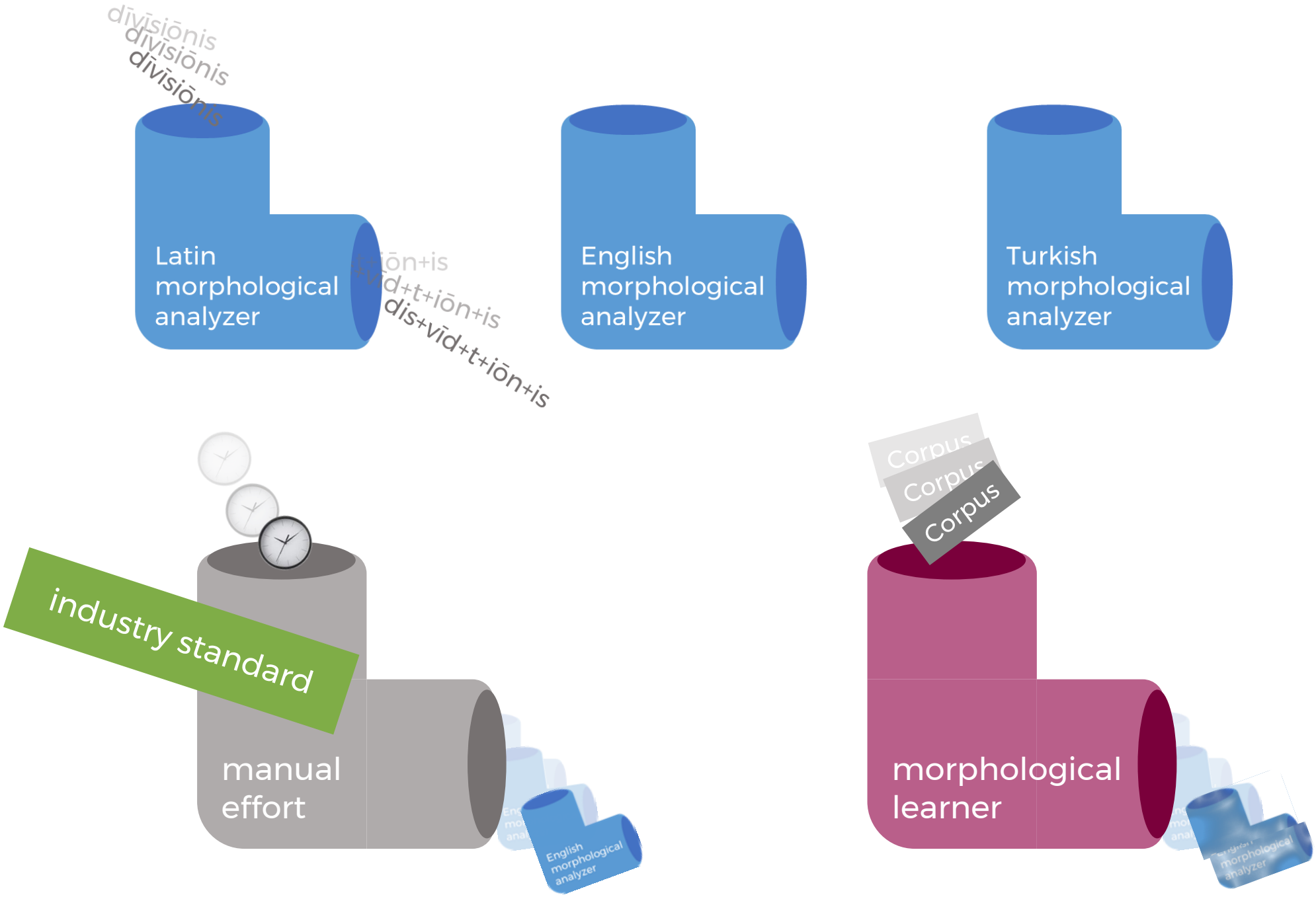


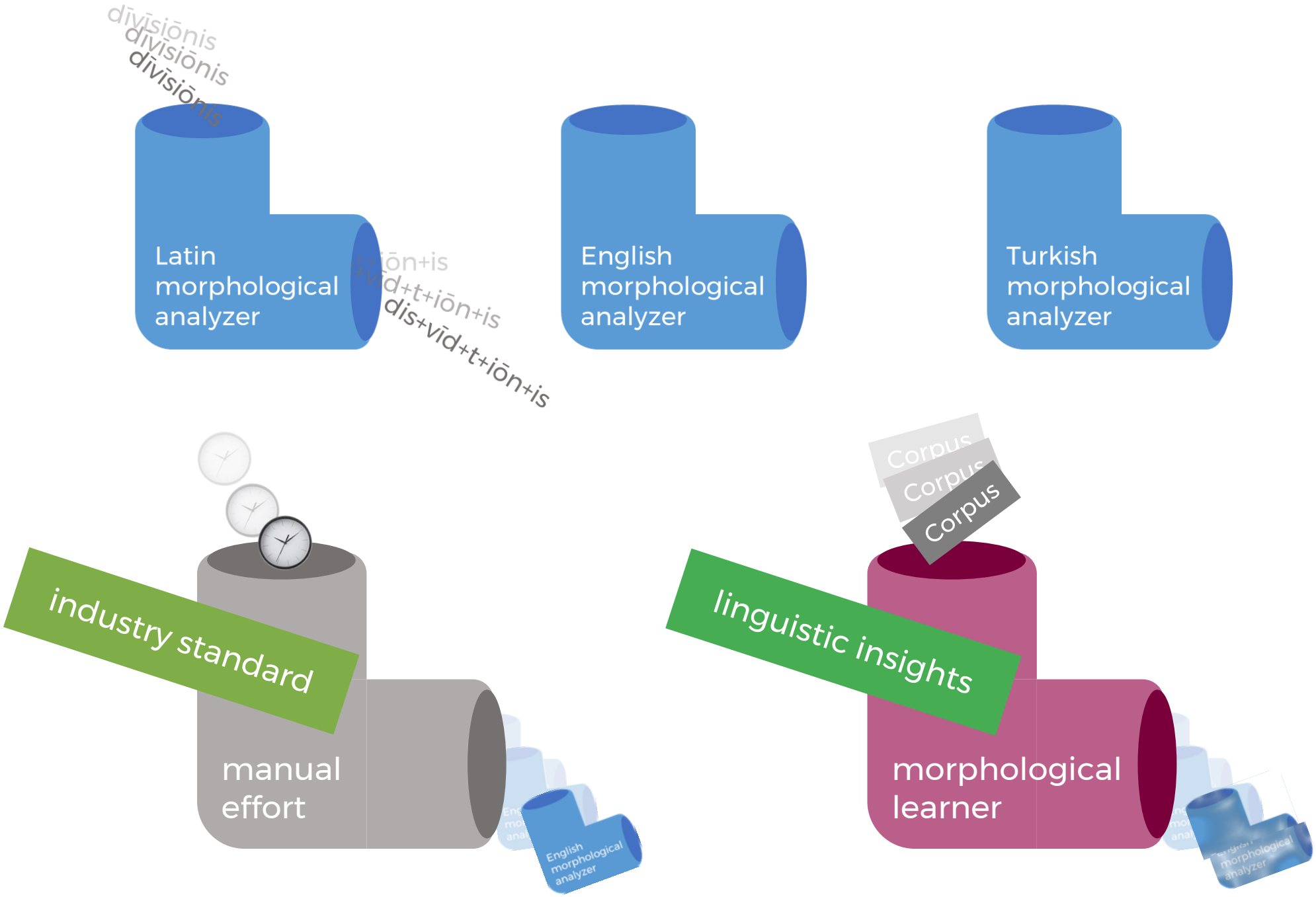

dīvisiōnis
divisiōnis
dīvisiōnis
Latin morphological analyzer
iōn+is
vid+t+iōn+is
dis+vīd+t+iōn+is
English morphological analyzer
Turkish morphological analyzer
manual effort
English morphological analyzer
Corpus
Corpus
Corpus
morphological learner

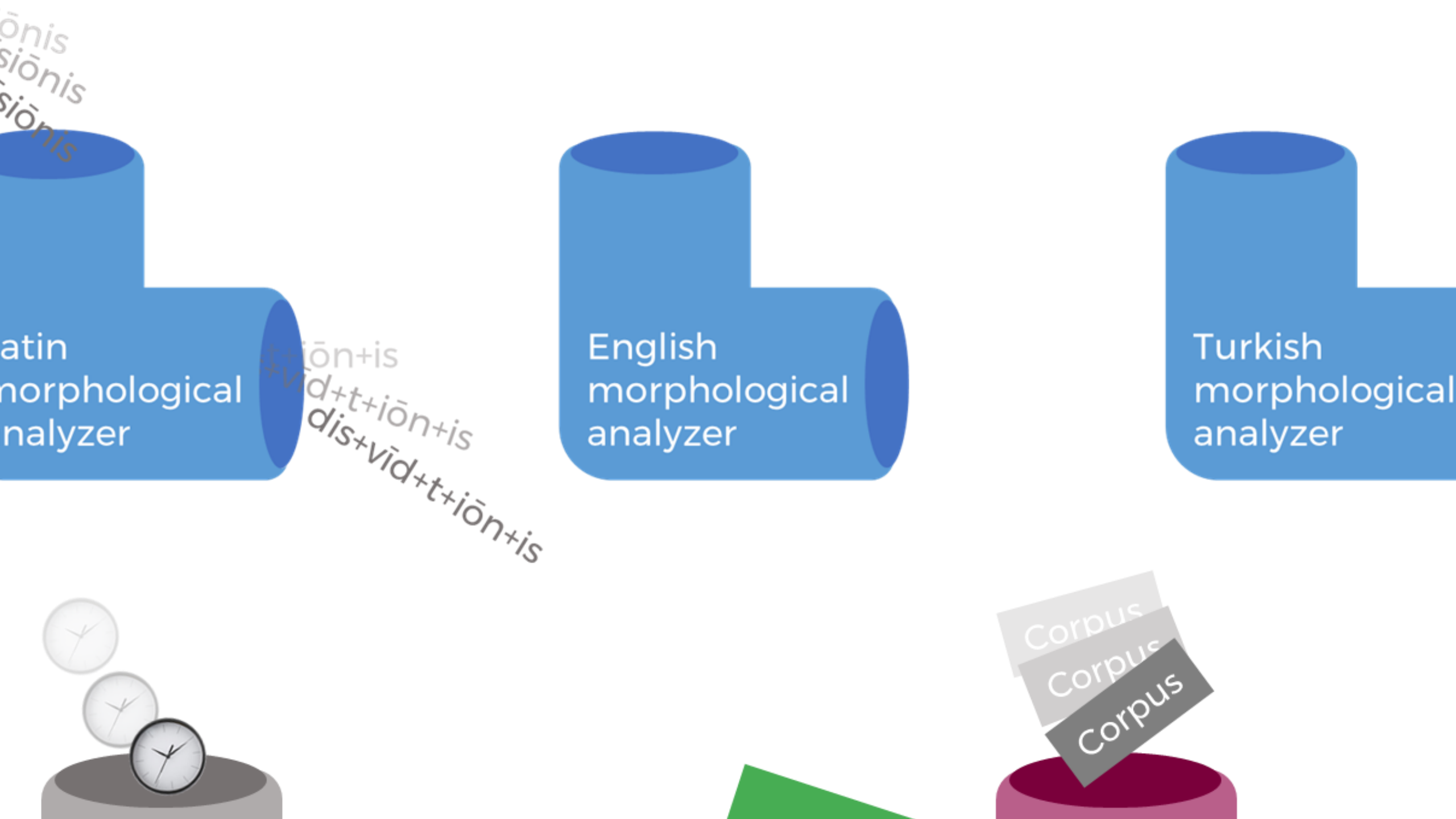

replicates the setting of natural language acqusition
Corpus
Corpus
Corpus
morphological
learner


replicates the setting of natural language acqusition
form
meaning
unsupervised learning


replicates the setting of natural language acqusition
Corpus
Corpus
Corpus
form
meaning
unsupervised learning


dīvīsiōnis
dīvīsiōnis
dīvīsiōnis
Latin morphological analyzer
iōn+is
d+t+iōn+is
dis+vīd+t+iōn+is
English morphological analyzer
Turkish morphological analyzer
industry standard
manual effort
English morphological analyzer
Corpus
Corpus
Corpus
morphological learner


dīvīsiōnis
dīvīsiōnis
dīvīsiōnis
Latin morphological analyzer
iōn+is
vīd+t+iōn+is
dis+vīd+t+iōn+is
English morphological analyzer
Turkish morphological analyzer
industry standard
manual effort
English morphological analyzer
Corpus
Corpus
Corpus
linguistic insights
morphological learner


ōnis
siōnis
siōnis
atin
morphological
nalyzer
t+iōn+is
+vīd+t+iōn+is
dis+vīd+t+iōn+is
English
morphological
analyzer
Turkish
morphological
analyzer
Corpus
Corpus
Corpus

Latin
morphological
analyzer

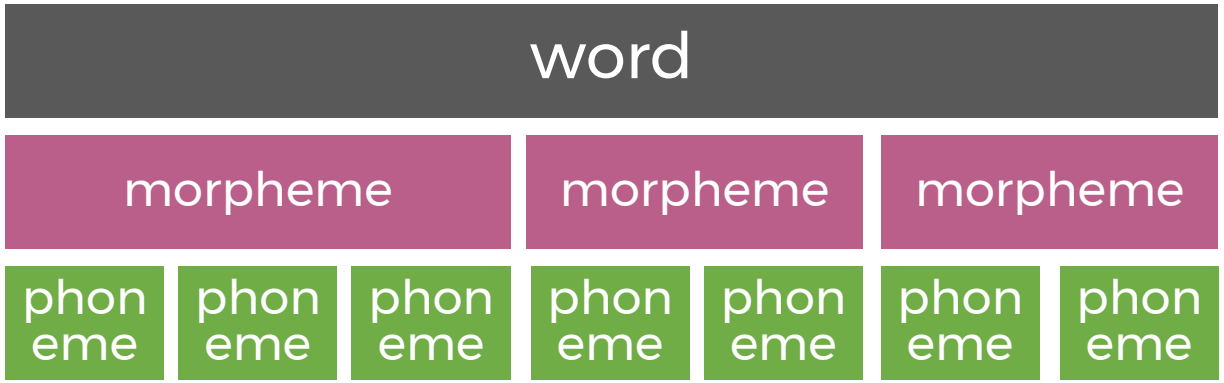
English
morphologica
analyzer

root	
suffix	inflectional suffix
prefix	inflectional prefix
infix	inflectional infix

root	
suffix	inflectional suffix
prefix	inflectional prefix
infix	inflectional infix

invīsibilis

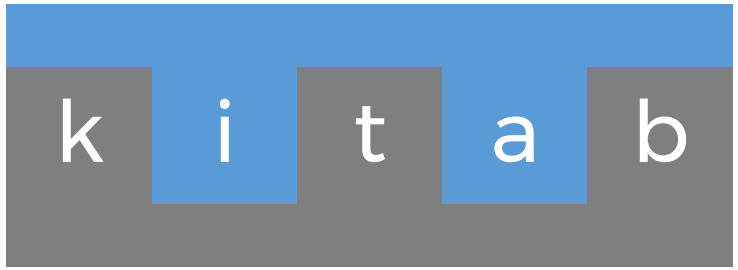
root	
suffix	inflectional suffix
prefix	inflectional prefix
infix	inflectional infix



invīsibilis


AND NOW FOR
SOMETHING
COMPLETELY
DIFFERENT

Arabic

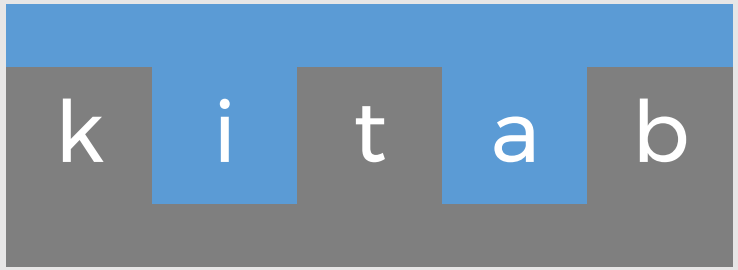


‘letter, book’

‘it/he was written’

‘it/he was caused to write’

Arabic



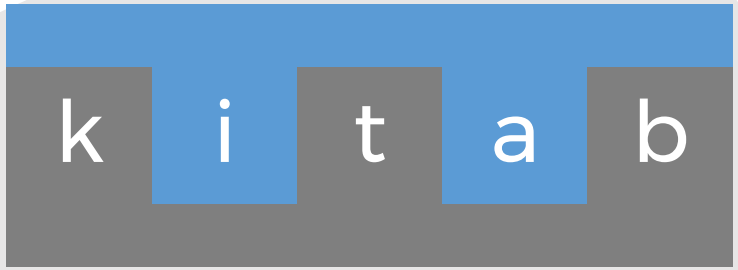
‘letter, book’

the same root

‘it/he was written’

‘it/he was caused to write’

Arabic



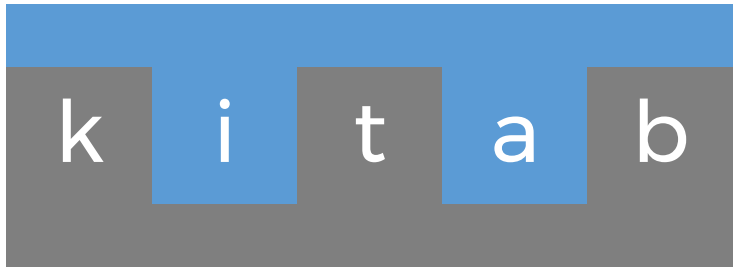
‘letter, book’

the same binyan

‘it/he was written’

‘it/he was caused to write’

Arabic



‘letter, book’

the same affix

‘it/he was written’

‘it/he was caused to write’

a morpheme is

a minimal meaning-bearing **part** of a word

a formal **operation** on a word,
accompanied by a semantic alteration

a morpheme is

a minimal meaning-bearing **part** of a word

a formal **operation** on a word,
accompanied by a semantic alteration

a morpheme is

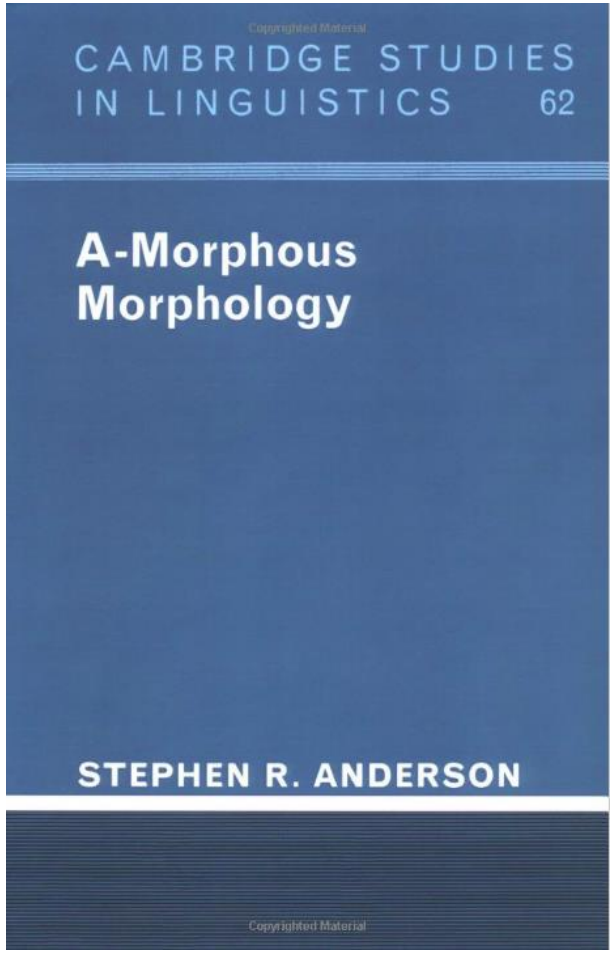
a minimal meaning-bearing **part** of a word

a formal **operation** on a word,
accompanied by a semantic alteration

a morpheme is

a minimal meaning-bearing **part** of a word

a formal **operation** on a word,
accompanied by a semantic alteration



non-exhaustive list

the diversitarian morpheme

root	
suffixation	English *bring* → *bring**ing***
prefixation	Latin *amīcus* → ***in**imīcus*
infixation	Tagalog *sulat* → *s**um**ulat*
transfixation	Arabic √*ktb* → *k**i**t**a**b*
reduplication	Lakota [hã**ska**] → [hã**ska**ska]
ablaut	Proto-IE **mentis* → **mnteis*
prosodic shift	English *insúlt* → *ínsult*
truncation	French */grãd/* → */grã/*

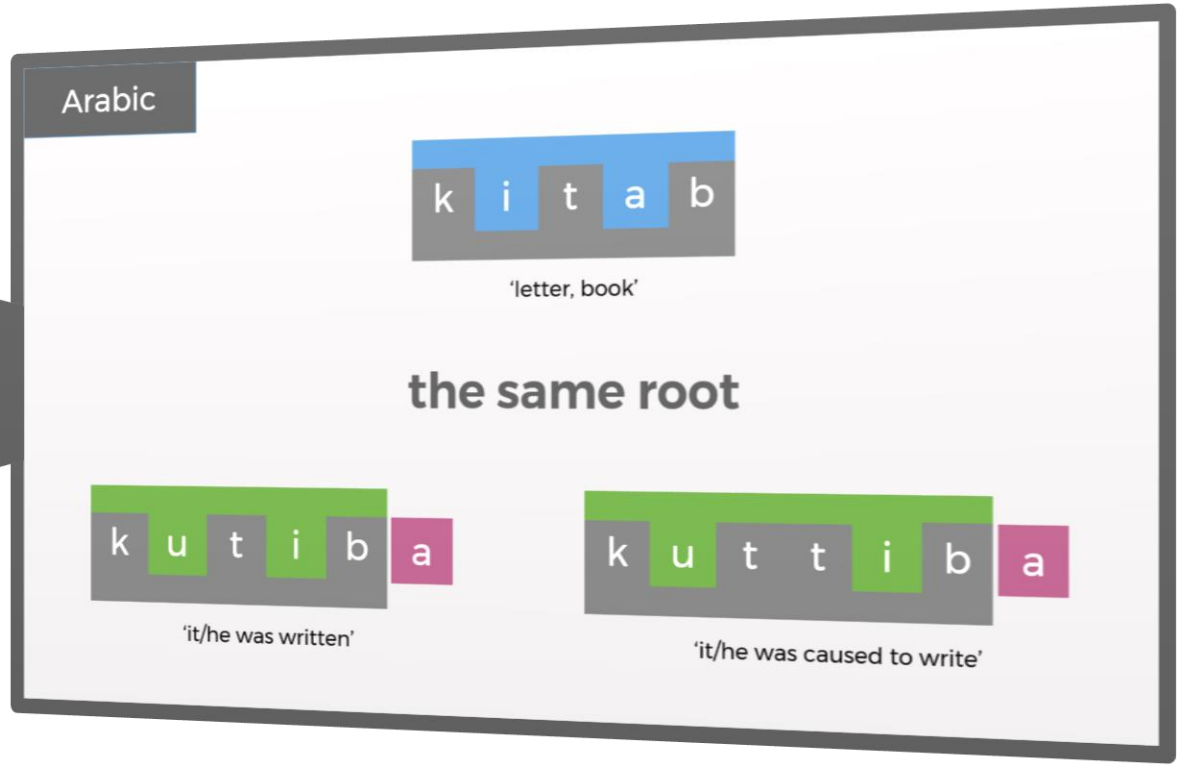
do not forget sandhi either

it can get quite heavy

Finnish example from Anderson's *A-Morphous Morphology*:

/karahka + i + ta/	"stick" + "plural" + "partitive"
karahko + i + ta	(a → o before i)
karahko + i + a	(t → θ after a weak syllable before a short vowel)
[karahkoja]	(glide formation)

digression to warm up with computer-scientific concepts



how could a machine find the longest cluster* of letters that two words have in common?

*not necessarily contiguous

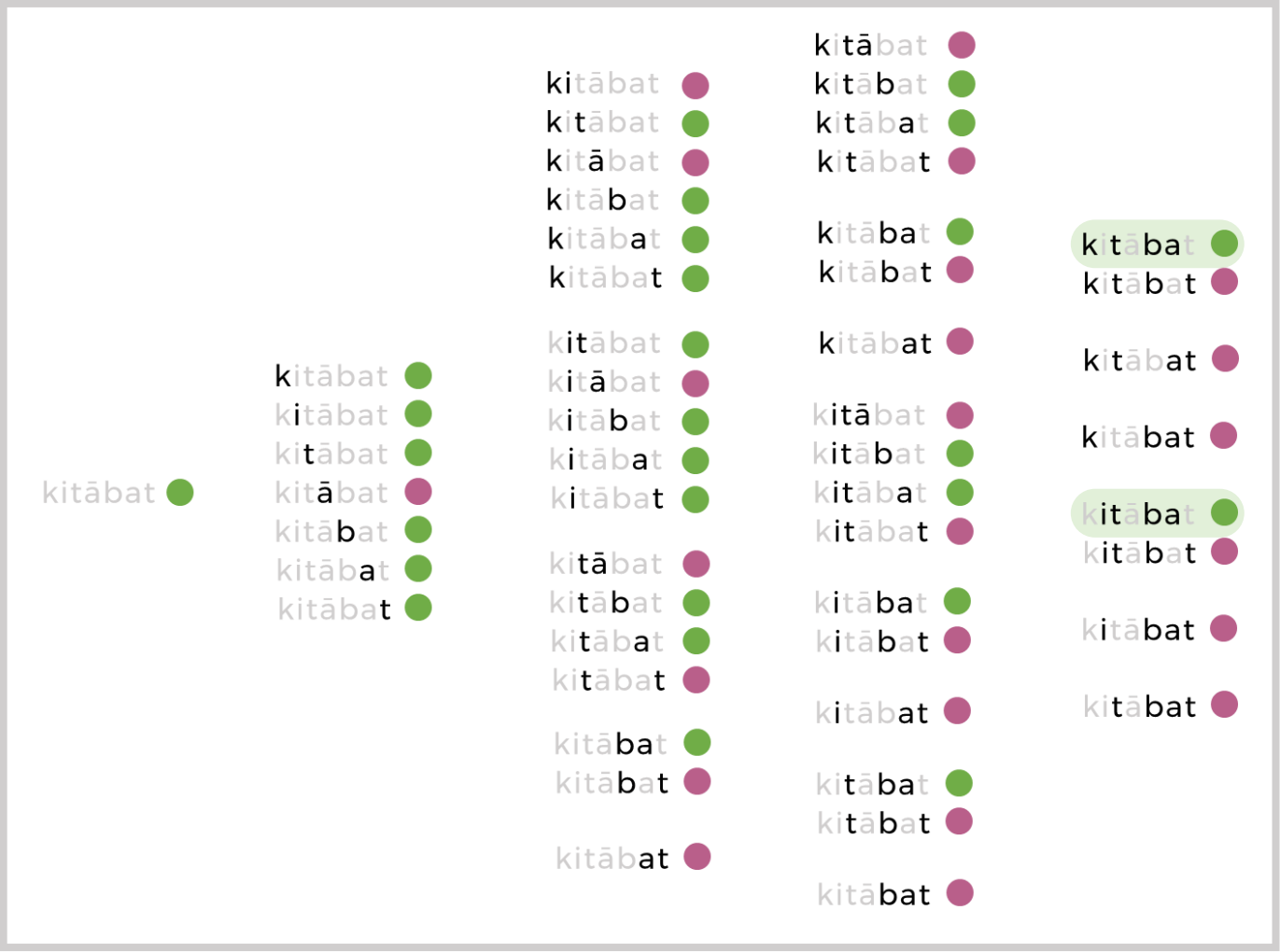
= longest common subsequence (**LCS**) problem

LCS(“istaktaba”, “kitābat”) = “ktba” **or** “itba”

A series of decisions leads to a candidate solution:

(D_1) pick letter 1 of LCS,
(D_2) pick letter 2 of LCS,
⋮
(D_M) pick letter M of LCS

The choice at D_1 shapes the option range of D_2 and so on.

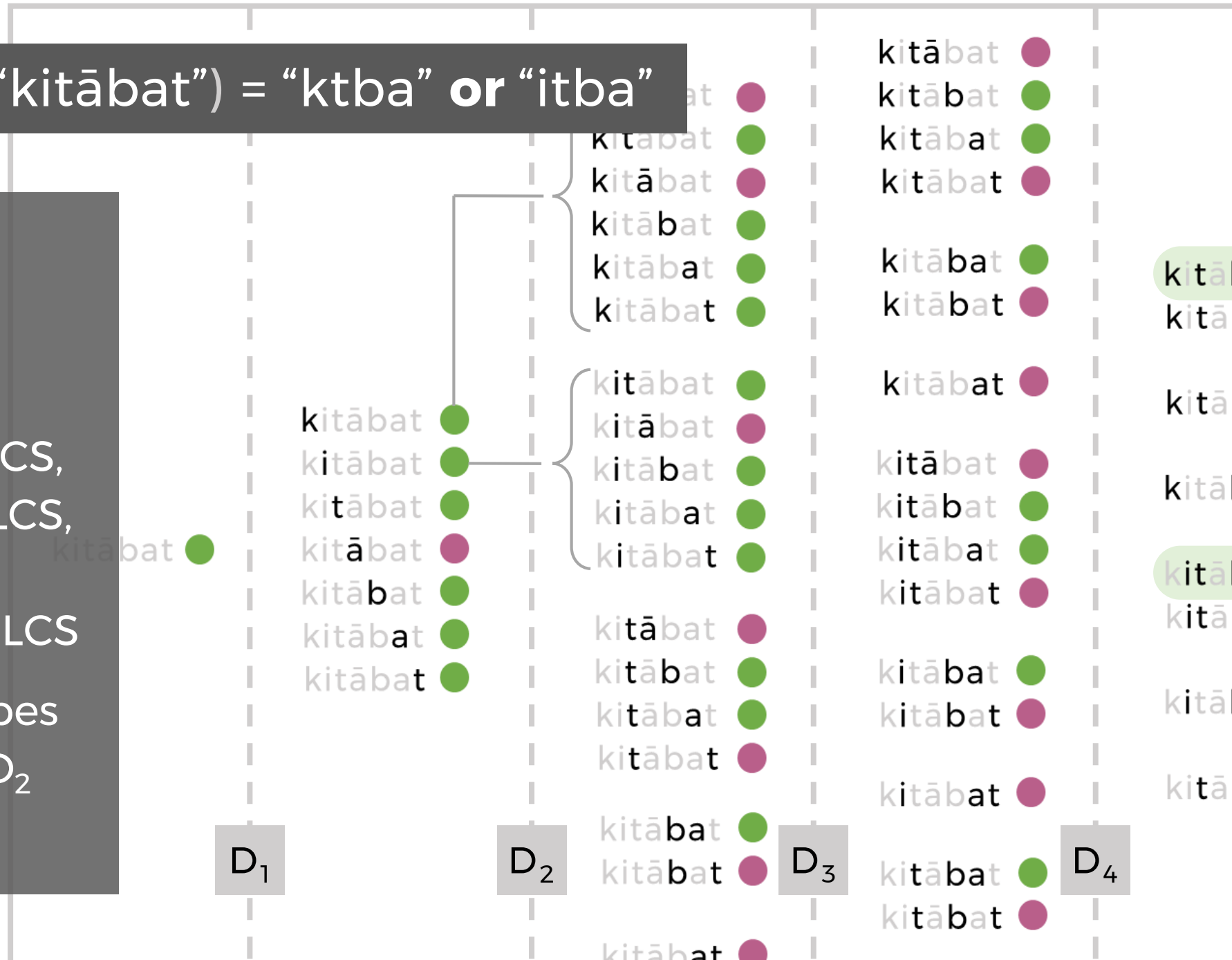


LCS(“istaktaba”, “kitābat”) = “ktba” **or** “itba”

A series of decisions leads to a candidate solution:

(D$_1$) pick letter 1 of LCS,
(D$_2$) pick letter 2 of LCS,
⋮
(D$_M$) pick letter *M* of LCS

The choice at D$_1$ shapes the option range of D$_2$ and so on.

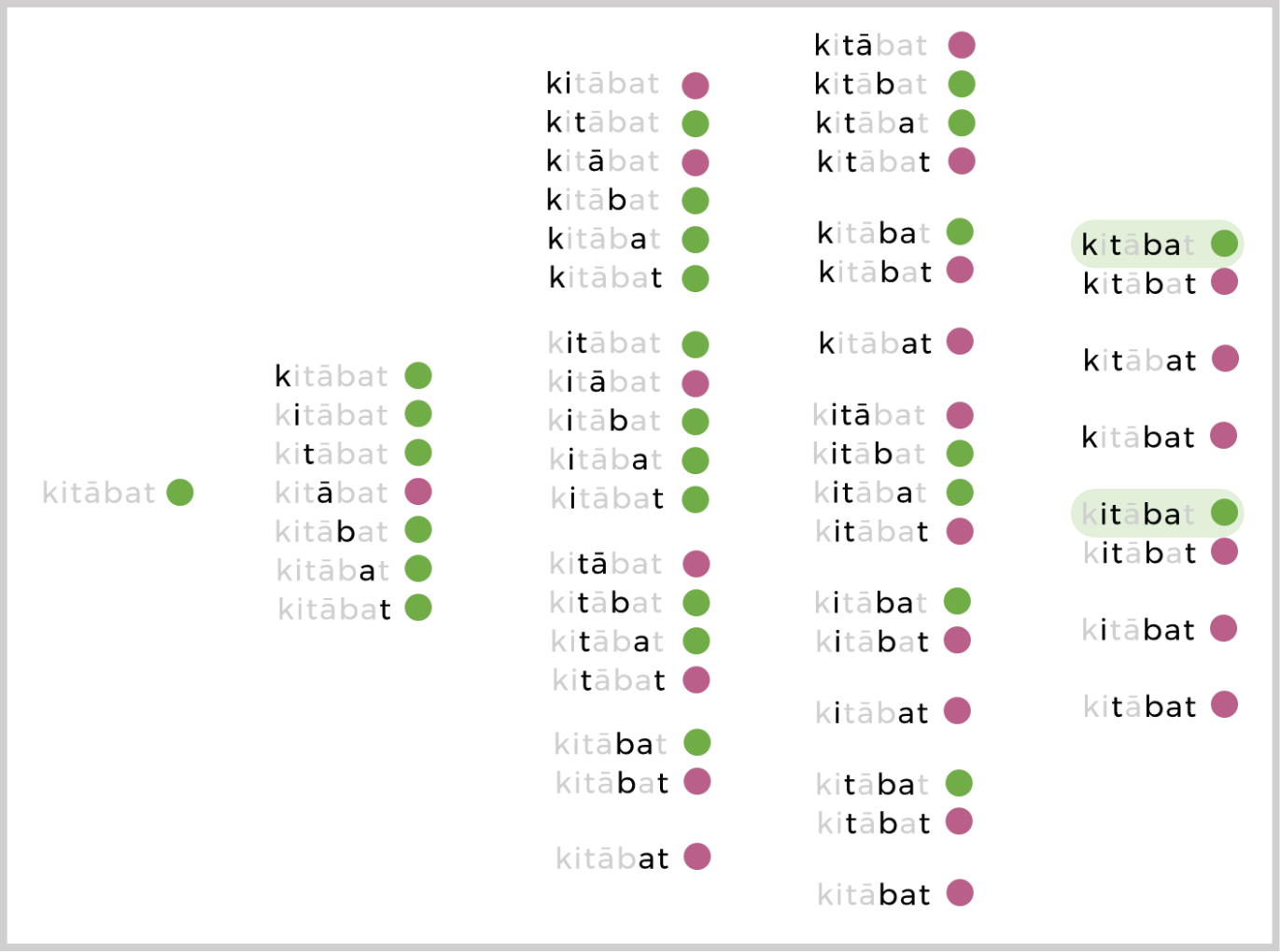


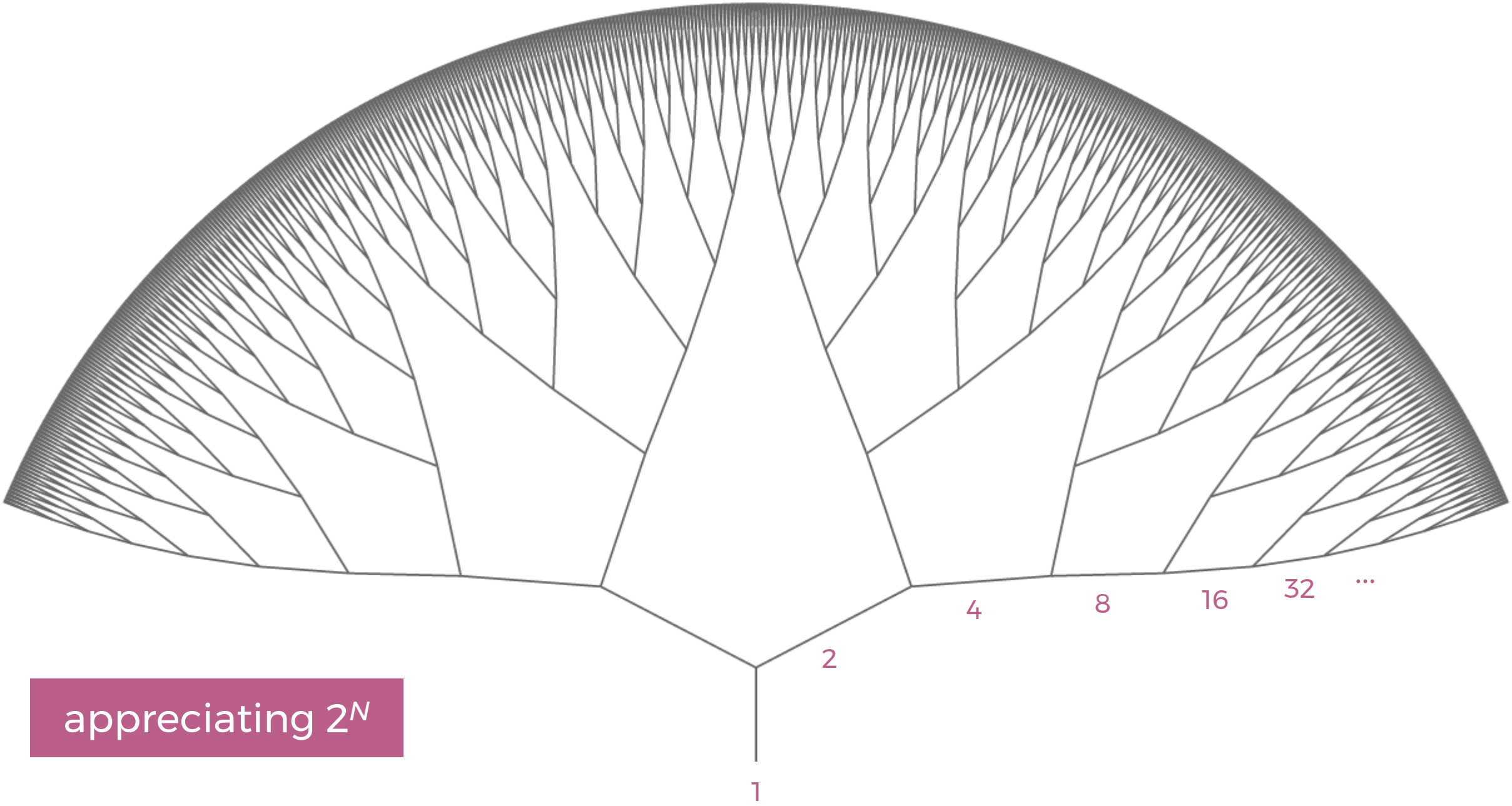
LCS(“istaktaba”, “kitābat”) = “ktba” **or** “itba”

An **exhaustive search** algorithm will consider all combinations of choices, up to a total of

$$2^N,$$

where *N* is the length of the shorter word.





appreciating 2^N

$2^{10} = 1\,024$ — **a second**

$2^{20} = 1\,048\,576$

$2^{30} = 1\,073\,741\,824$ — **a year**

$2^{40} = 1\,099\,511\,627\,776$

$2^{50} = 1\,125\,899\,906\,842\,624$ — **a billion years**

$2^{60} = 1\,152\,921\,504\,606\,846\,976$ — **no life on Earth**

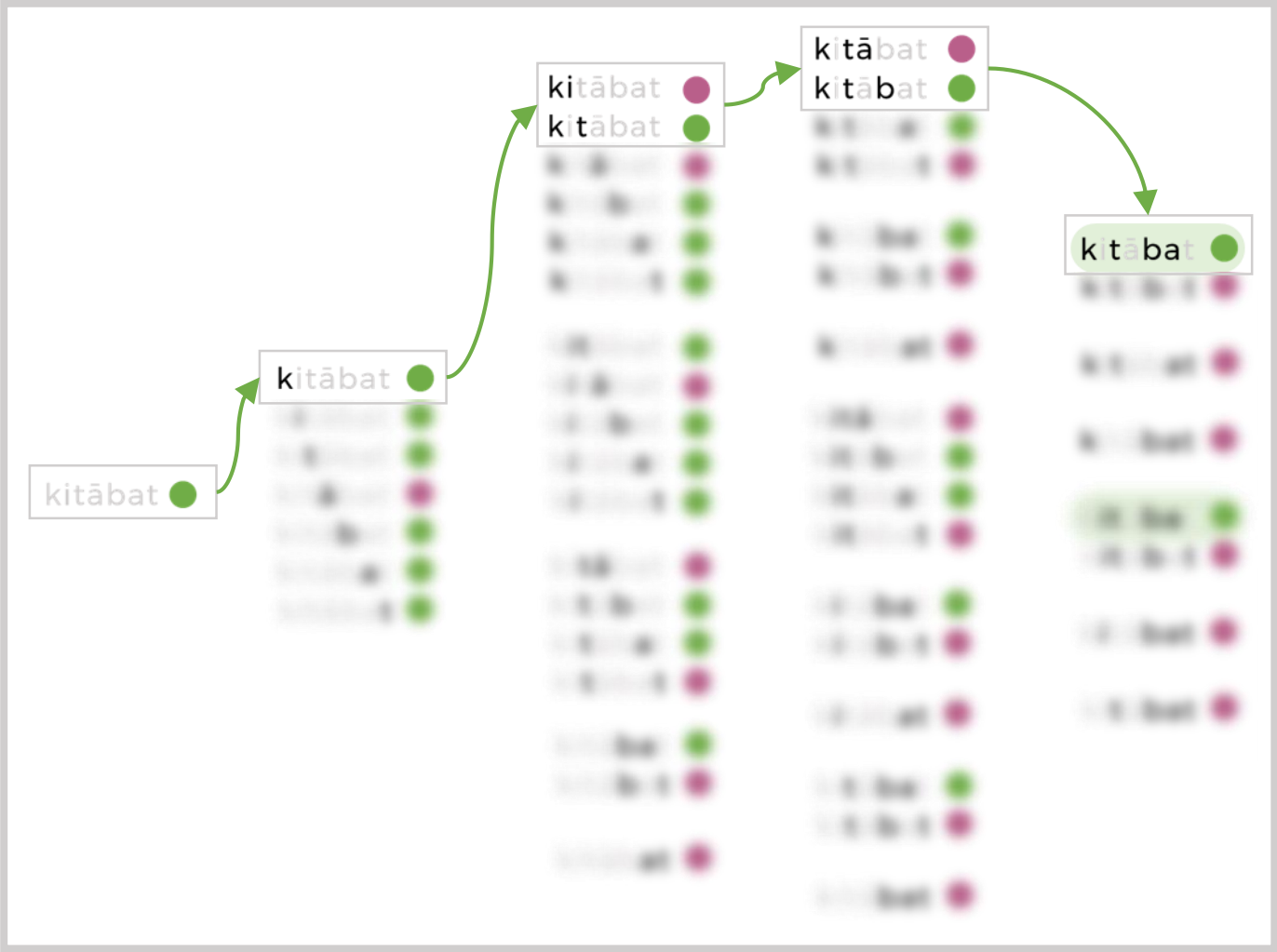
$2^{70} = 1\,180\,591\,620\,717\,411\,303\,424$

$2^{80} = 1\,208\,925\,819\,614\,629\,174\,706\,176$

$2^{90} = 1\,237\,940\,039\,285\,380\,274\,899\,124\,224$

Upping the length of our words from **10** to **70** is enough to prolong the running time of the algorithm from **a second** to the **lifetime of our planet**.

LCS(“istaktaba”, “kitābat”) = “ktba” **or** “itba”



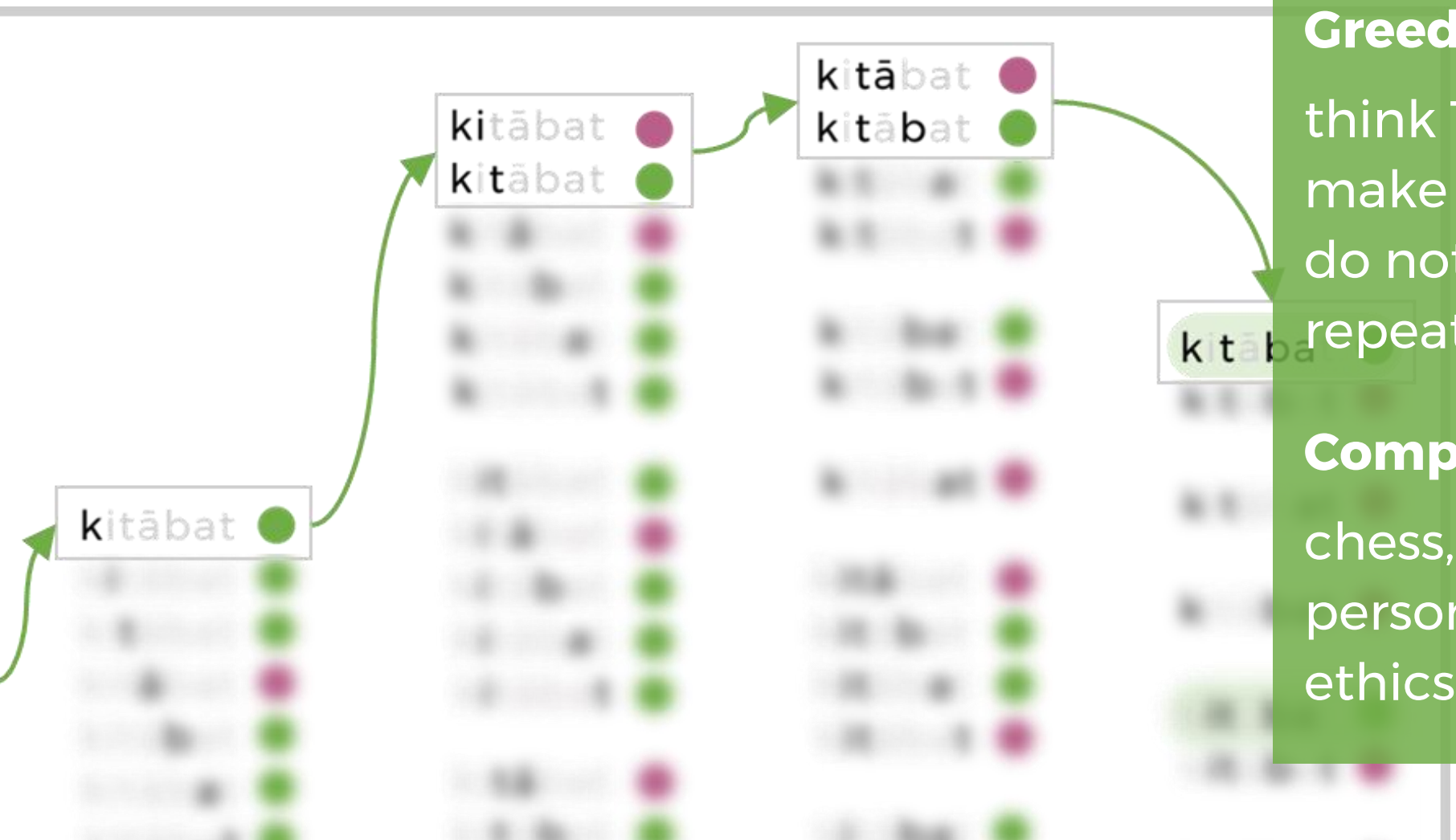
Greedy strategy:

think 1 move ahead,
make your choice,
do not look back,
repeat.

Compare:

chess,
personal utilitarian
ethics.

LCS(“istaktaba”, “kitābat”) = “ktba” **or** “itba”



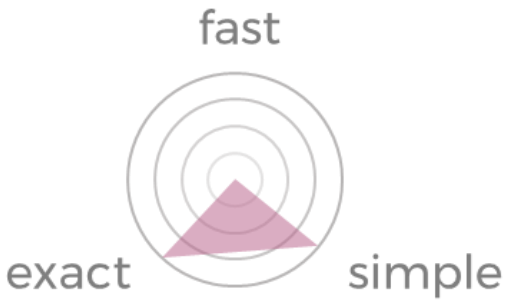
Greedy strategy:

think 1 move ahead,
make your choice,
do not look back,
repeat.

Compare:

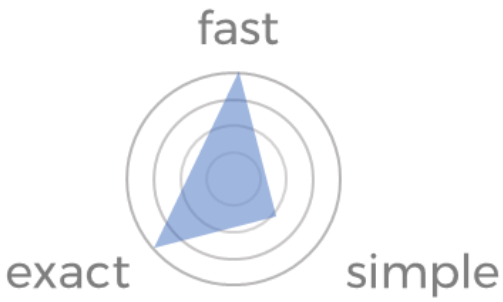
chess,
personal utilitarian
ethics.

Typical trade-offs in multiple (alternate) algorithms that all solve the same problem



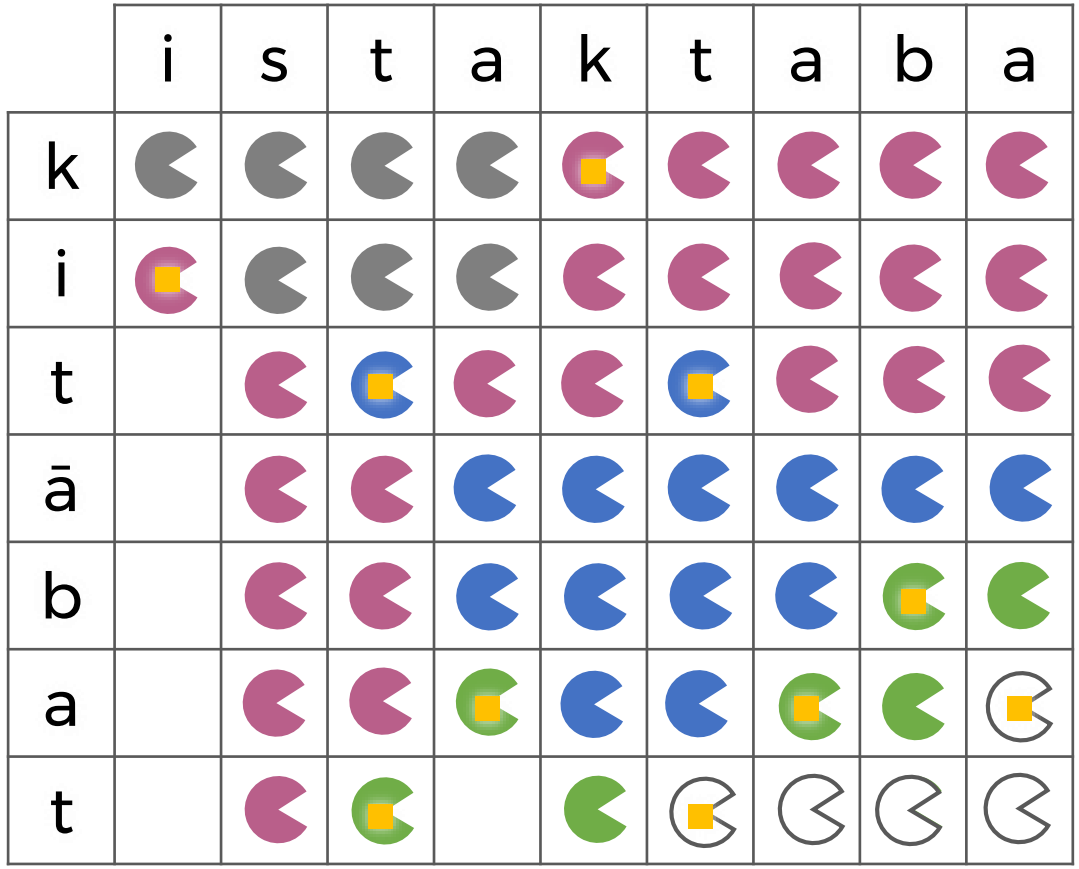
Exhaustive search

Greedy strategies

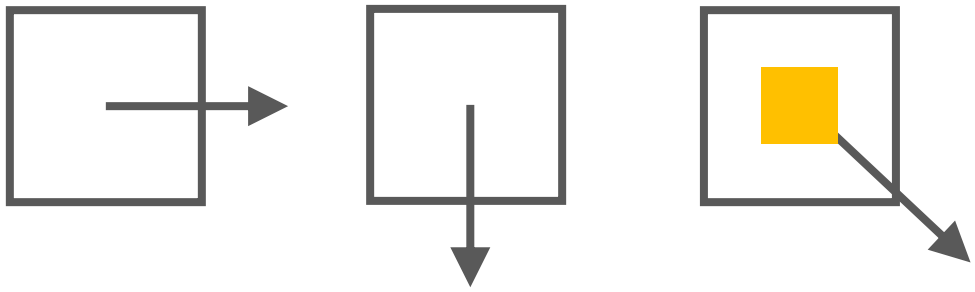


Non-standard solutions

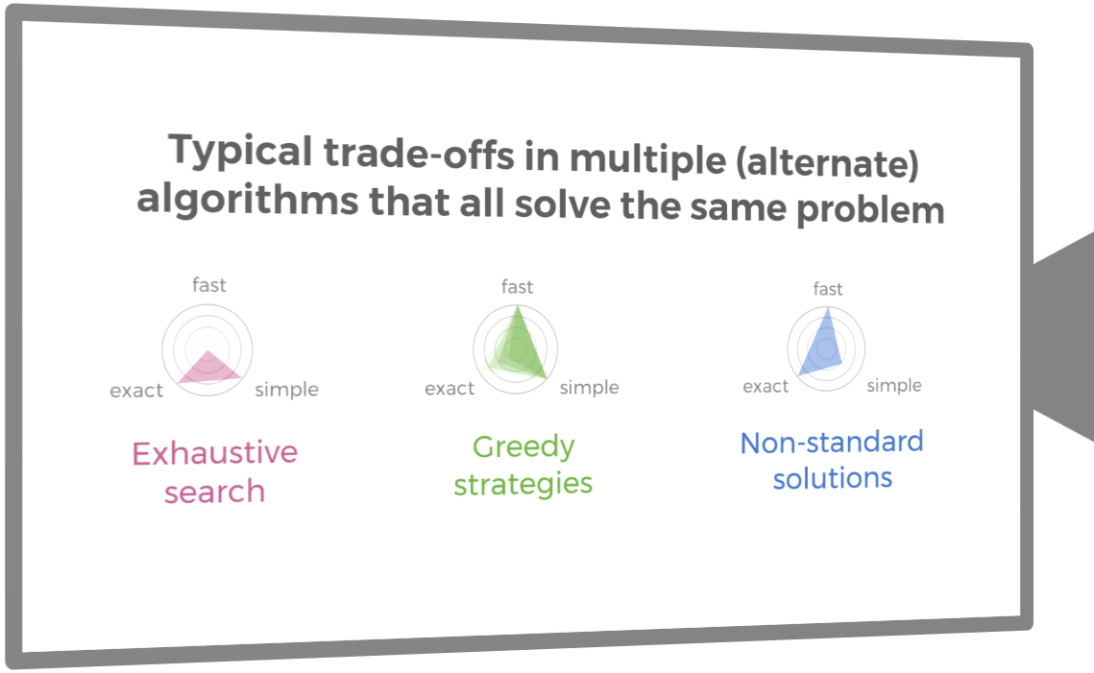
LCS(“istaktaba”, “kitābat”) = “ktba” **or** “itba”



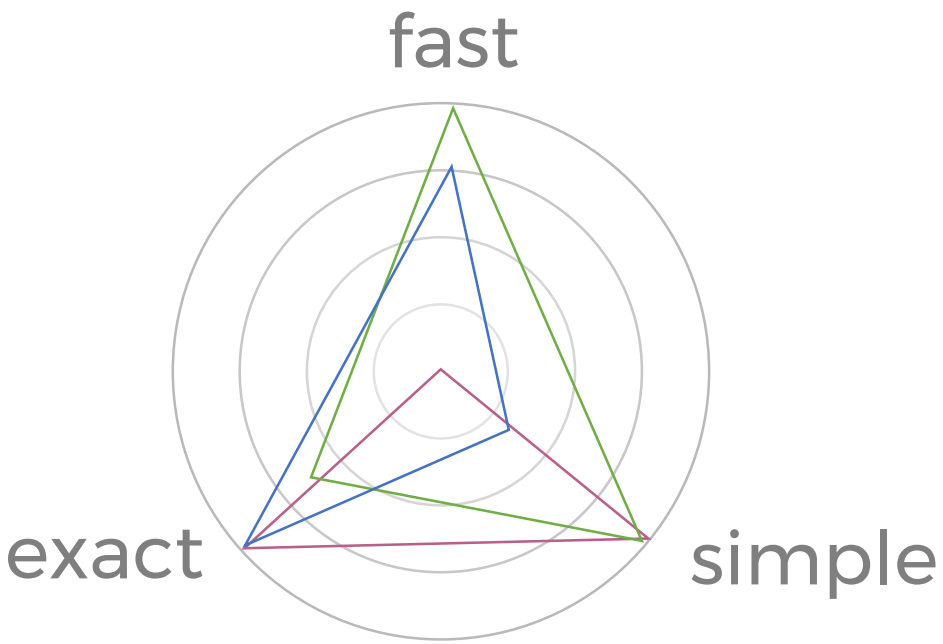
The Needleman-Wunsch algorithm for LCS is very much a non-trivial invention. It is as exact as the 2^N algorithm, but consumes just N^2 units of time instead.

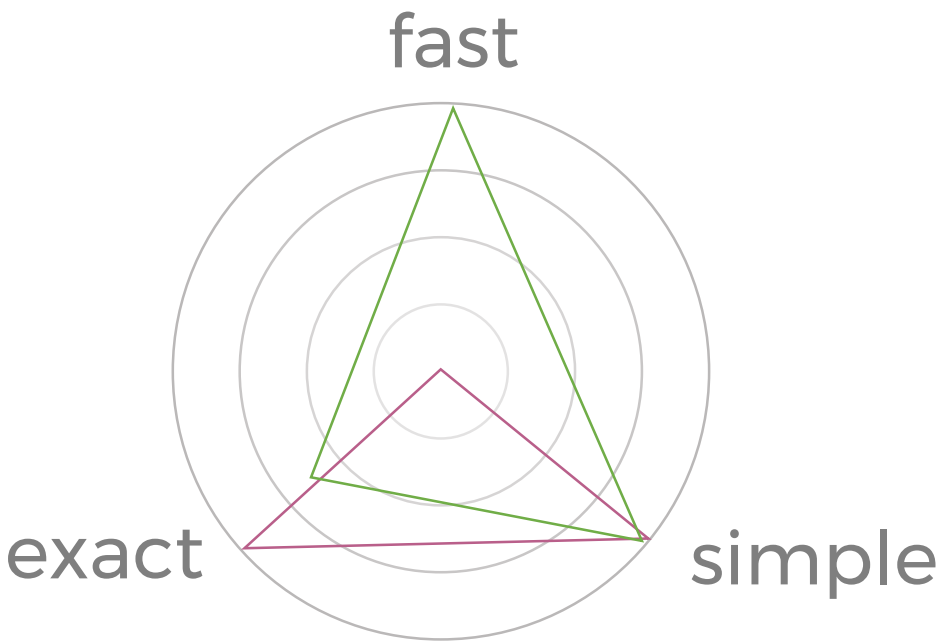


Trade-off structure for the longest common subsequence (LCS) problem



a **tractable** problem





an **intractable** problem

one that is believed to have **no** exact solution of appreciable efficiency

The artificial learners await us now.

c a t t u s m ū r ī l e g u s m ū r ē s l e g i t

c+a t t u s m ū r ī l e g u s m ū r ē s l e g i t

c a+t t u s m ū r ī l e g u s m ū r ē s l e g i t

c+a+t t u s m ū r ī l e g u s m ū r ē s l e g i t

c a t+t u s m ū r ī l e g u s m ū r ē s l e g i t

c a t t+u s m ū r+ī+l e g+u s m ū r+ē s l e g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e g+i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i+t

2^N

c a t t u s m ū r ī l e g u s m ū r ē s l e g i t

c+a t t u s m ū r ī l e g u s m ū r ē s l e g i t

c a+t t u s m ū r ī l e g u s m ū r ē s l e g i t

c+a+t t u s m ū r ī l e g u s m ū r ē s l e g i t

c a t+t u s m ū r ī l e g u s m ū r ē s l e g i t

⋮

c a t t+u s m ū r+ī+l e g+u s m ū r+ē s l e g+i t

⋮

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e g+i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i+t

“alternate” morpho-logies	c a t t u s	m ū r ī l e g u s	m ū r ē s	l e g i t
	c+a t t u s	m ū r ī l e g u s	m ū r ē s	l e g i t
	c a+t t u s	m ū r ī l e g u s	m ū r ē s	l e g i t
	c+a+t t u s	m ū r ī l e g u s	m ū r ē s	l e g i t
	c a t+t u s	m ū r ī l e g u s	m ū r ē s	l e g i t
	⋮			
	c a t t+u s	m ū r+ī+l e g+u s	m ū r+ē s	l e g+i t
	⋮			
“alternate” morpho-logies	c+a+t+t+u+s	m+ū+r+ī+l+e+g+u+s	m+ū+r+ē+s	l+e g+i+t
	c+a+t+t+u+s	m+ū+r+ī+l+e+g+u+s	m+ū+r+ē+s	l+e+g i t
	c+a+t+t+u+s	m+ū+r+ī+l+e+g+u+s	m+ū+r+ē+s	l+e+g i+t
	c+a+t+t+u+s	m+ū+r+ī+l+e+g+u+s	m+ū+r+ē+s	l+e+g+i t
	c+a+t+t+u+s	m+ū+r+ī+l+e+g+u+s	m+ū+r+ē+s	l+e+g+i+t

the right morphology

what makes the right morphology stand out among the others?

the proposed lexical items **recur**

c a t t+u s m ū r+ī+l e g+u s m ū r+ē s l e g+i t

⋮

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e g+i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i+t

what makes the right morphology stand out among the others?

the proposed lexical items **recur**

c a t t+u s m ū r+ī+l e g+u s m ū r+ē s l e g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e g+i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i+t

what makes the right morphology stand out among the others?

the proposed lexical items **recur**

c a t t+u s m ū r+ī+l e g+u s m ū r+ē s l e g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e g+i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i+t

what makes the right morphology stand out among the others?

the proposed lexical items **recur**

c a t t+u s m ū r+ī+l e g+u s m ū r+ē s l e g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e g+i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g i+t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i t

c+a+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i+t

what makes the right morphology stand out among the others?

the proposed lexical items **recur**

c a t t+u s m ū r+ī+l e g+u s m ū r+ē s l e g+i t

+0 +0 +0 +1 +1 +1

hypothesis: score a point for each re-use of a lexical item; adopt the top-scoring lexicon

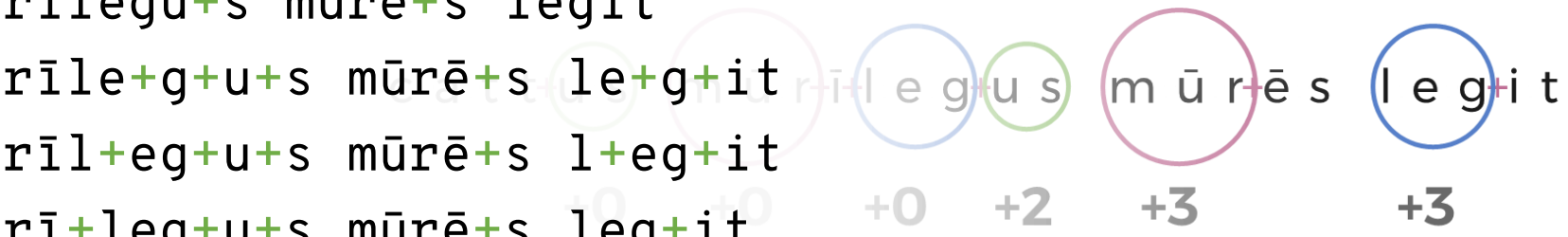
simulation does **not** support our hypothesis

Score

1 cattus mūrīlegu+s mūrē+s legit
2 cattus mūrīle+g+u+s mūrē+s le+g+it
3 cattus mūrīl+e+g+u+s mūrē+s l+e+g+it
4 cattus mūrī+l+e+g+u+s mūrē+s l+e+g+it
5 cattus mūr+ī+l+e+g+u+s mūr+ē+s l+e+g+it
6 cattus mū+r+ī+l+e+g+u+s mū+r+ē+s l+e+g+it
7 cattus m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+it
8 cattu+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+it
9 catt+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+it
10 cat+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i+t
11 ca+t+t+u+s m+ū+r+ī+l+e+g+u+s m+ū+r+ē+s l+e+g+i+t

we perform much better, though, by making longer repetitions score more points

Score	
1	cattus mūrīlegu+s mūrē+s legit
2	cattus mūrīle+g+u+s mūrē+s le+g+it
3	cattus mūrīl+eg+u+s mūrē+s l+eg+it
4	cattus mūrī+leg+u+s mūrē+s leg+it
5	cattus mūr+īle+g+u+s mūr+ē+s le+g+it
6	cattus mūr+īl+eg+u+s mūr+ē+s l+eg+it
7	cattus mūr+ī+leg+u+s mūr+ē+s leg+it
8	cattu+s mūr+ī+leg+u+s mūr+ē+s leg+it
9	catt+u+s mūr+ī+leg+u+s mūr+ē+s leg+it
10	cat+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t
11	ca+t+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t



Unsettling interim observations

1. Still operating within an 2^N-time exhaustive search framework. No tractability guarantee.
2. Still considering only prefixes and suffixes.
3. No regard for well-formedness of the analyses, e.g. suffixes can also appear as prefixes.

no regard for well-formedness of the analyses, e.g., suffixes can also appear as prefixes

```
1 cattus mūrīlegu+s mūrē+s legit
2 cattus mūrīle+g+u+s mūrē+s le+g+it
3 cattus mūrīl+eg+u+s mūrē+s l+eg+it
4 cattus mūrī+leg+u+s mūrē+s leg+it
5 cattus mūr+īle+g+u+s mūr+ē+s le+g+it
6 cattus mūr+īl+eg+u+s mūr+ē+s l+eg+it
7 cattus mūr+ī+leg+u+s mūr+ē+s leg+it
8 cattu+s mūr+ī+leg+u+s mūr+ē+s leg+it
9 catt+u+s mūr+ī+leg+u+s mūr+ē+s leg+it
10 cat+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t
11 ca+t+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t
```

no regard for well-formedness of the analyses, e.g., suffixes can also appear as prefixes

iterated suffixes (like +*t*+*t*) are unlikely

```
 2 cattus mūrīle+g+u+s mūrē+s le+g+it
 3 cattus mūrīl+eg+u+s mūrē+s l+eg+it
 4 cattus mūrī+leg+u+s mūrē+s leg+it
 5 cattus mūr+īle+g+u+s mūr+ē+s le+g+it
 6 cattus mūr+īl+eg+u+s mūr+ē+s l+eg+it
 7 cattus mūr+ī+leg+u+s mūr+ē+s leg+it
 8 cattu+s mūr+ī+leg+u+s mūr+ē+s leg+it
 9 catt+u+s mūr+ī+leg+u+s mūr+ē+s leg+it
10 cat+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t
11 ca+t+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t
```

no regard for well-formedness of the analyses, e.g., suffixes can also appear as prefixes

iterated suffixes (like +*t*+*t*) are unlikely

word-finality and -initiality is not captured

5 cattus mūr+īle+g+u+s mūr+ē+s le+g+it
6 cattus mūr+īl+eg+u+s mūr+ē+s l+eg+it
7 cattus mūr+ī+leg+u+s mūr+ē+s leg+it
8 cattu+s mūr+ī+leg+u+s mūr+ē+s leg+it
9 catt+u+s mūr+ī+leg+u+s mūr+ē+s leg+it
10 cat+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t
11 ca+t+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t

no regard for well-formedness of the analyses, e.g., suffixes can also appear as prefixes

iterated suffixes (like +*t*+*t*) are unlikely

word-finality and -initiality is not captured

incur markedness penalties through arbitrary rules?

```
                 g+it
                 s mūr+ē+s le+g+it
                 s mūr+ē+s l+eg+it
                 s mūr+ē+s leg+it
                 8 cattu+s mūr+ī+leg+u+s mūr+ē+s leg+it
                 9 catt+u+s mūr+ī+leg+u+s mūr+ē+s leg+it
-penalty1 + 10 cat+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t
-penalty2 + 11 ca+t+t+u+s mūr+ī+leg+u+s mūr+ē+s leg+i+t
```

*NoRoot (disqualify?)

*TwoRoots

*TwiceInAWord

*IsBothSuffixAndPrefix

*InconsistentlyFinalOrInitial

Did we just let UG in?

uniform reward scheme $\approx$ weighted reward scheme $\approx$ **MDL**


uniform reward scheme
≈
weighted reward scheme
≈
MDL
MDL family

minimum description length (**MDL**) proper

```
DL = 28 cattus mūrīlegus mūrēs legit
DL = 29 catt1 mūrīleg1 mūrēs legit/us
DL = 29 catt1 2īleg1 2ēs legit/us,mūr
DL = 29 catt1 2ī31 2ēs 3it/us,mūr,leg
```

reduce "us", of length 2, to a single digit, twice, **but** include "us" in the lexicon, once

lexicon

minimization was not fruitful in a corpus this small

superficially, a very much distinct principle

but can be also conceived of as just another scoring scheme

MDL as just another reward scheme

catt	+	us	#	mūr	+	ī	+	leg	+	us	#	mūr	+	ēs	#	leg	+	it
0		0		0		0		0		**1**		**1**		0		**1**		0

uniform reward scheme

catt	+	us	#	mūr	+	ī	+	leg	+	us	#	mūr	+	ēs	#	leg	+	it
0		0		0		0		0		**2**		**3**		0		**3**		0

weighted reward scheme

	catt	+	us	#	mūr	+	ī	+	leg	+	us	#	mūr	+	ēs	#	leg	+	it
	-4		-2		-3		-1		-3		-1		-1		-2		-1		-2
+	4		2		3		1		3		2		3		2		3		2
=	0		0		0		0		0		**1**		**2**		0		**2**		0

MDL, as presented here

add lengths

different flavors of **MDL** unequivocally dominate the artificial morphological learner scene

different flavors of **MDL** unequivocally dominate the artificial morphological learner scene

the **Goldsmith et al.** papers and *Linguistica*:

2000	Linguistica: An Automatic Morphological Analyzer
2001	Unsupervised Learning of the Morphology of a Natural Language
2002	Using eigenvectors of the bigram graph to infer morpheme identity
2004	An algorithm for the unsupervised learning of morphology
	From Signatures to Finite State Automata
2005	A heuristic for morpheme discovery based on string edit distance
	Using morphology and syntax together in unsupervised learning
	The SED heuristic for morpheme discovery: a look at Swahili
2006	Exploring variant definitions of pointer length in MDL.

different flavors of **MDL** unequivocally dominate the artificial morphological learner scene

the **Creutz et al.** papers and *Morfessor*:

2002 Unsupervised Discovery of Morphemes

2004 Induction of a Simple Morphology for Highly-Inflecting Languages

2005 Inducing the Morphological Lexicon of a Natural Language from Unannotated Text

Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0

Latent Linguistic Codes for Morphemes Using Independent Component Analysis

2006 Morfessor in the Morpho Challenge

non-exhaustive list

Morpho Challenge 2010 by Creutz et al.

AUTHOR	METHOD	PRECISION	RECALL	F-MEASURE	TYPE
Kohonen	Morfessor S+W	65.62%	69.28%	67.40%	S
Kohonen	Morfessor S+W+L	67.87%	66.43%	67.14%	S
Lignos	Base Inference	80.77%	53.76%	64.55%	U
Lignos	Iterative Compounding	80.27%	52.76%	63.67%	U
Spiegler	DEAP MDL-NOCAT	51.44%	80.95%	62.91%	S
Spiegler	DEAP PROB-NOCAT	58.52%	63.06%	60.70%	S
Lignos	Aggressive Compounding	71.45%	52.31%	60.40%	U
Kohonen	Morfessor U+W	60.33%	59.55%	59.94%	P
Spiegler	DEAP MDL-CAT	49.83%	75.08%	59.90%	S
Nicolas	MorphAcq	67.83%	53.43%	59.78%	U
Spiegler	DEAP PROB-CAT	55.00%	56.61%	55.79%	S
-	Morfessor Baseline	81.39%	41.70%	55.14%	U
Spiegler	Promodes	39.59%	64.72%	49.13%	P
Spiegler	Promodes-E	49.69%	46.22%	47.90%	P
-	Morfessor CatMAP	86.84%	30.03%	44.63%	U
Golenia	MAGIP	29.88%	70.65%	42.00%	S
Spiegler	Promodes-H	26.84%	63.69%	37.77%	P
-	Letters	4.41%	99.88%	8.	

English

AUTHOR	METHOD	PRECISION	RECALL	F-MEASURE	TYPE
Spiegler	DEAP MDL-NOCAT	56.03%	70.71%	62.52%	S
Spiegler	DEAP MDL-CAT	57.43%	66.59%	61.67%	S
Kohonen	Morfessor S+W+L	58.38%	63.35%	60.76%	S
Kohonen	Morfessor S+W	57.59%	55.21%	56.38%	S
Spiegler	DEAP PROB-NOCAT	56.66%	44.79%	50.03%	S
Kohonen	Morfessor U+W	56.97%	42.98%	49.00%	P
Lignos	Aggressive Compounding	63.07%	39.98%	48.94%	U
Spiegler	DEAP PROB-CAT	51.93%	42.41%	46.69%	S
Spiegler	Promodes-E	43.26%	45.73%	44.46%	P
Spiegler	Promodes	41.59%	47.41%	44.31%	P
Golenia	MAGIP	36.25%	53.73%	43.29%	S
-	Morfessor CatMAP	80.31%	29.51%	43.16%	U
Spiegler	Promodes-H	35.08%	53.45%	42.36%	P
Lignos	Iterative Compounding	75.32%	26.70%	39.43%	U
Lignos	Base Inference	78.98%	24.46%	37.35%	U
-	Morfessor Baseline	90.60%	14.39%	24.	U
-	Letters	5.04%	99.89%	9.6	

FInnish

Morpho Challenge 2010 by Creutz et al.

AUTHOR	METHOD	PRECISION	RECALL	F-MEASURE	TYPE
Kohonen	Morfessor S+W	65.62%	69.28%	67.40%	S
Kohonen	Morfessor S+W+L	67.87%	66.43%	67.14%	S
Lignos	Base Inference	80.77%	53.76%	64.55%	U
Lignos	Iterative Compounding	80.27%	52.76%	63.67%	U
Spiegler	DEAP MDL-NOCAT	51.44%	80.95%	62.91%	S
Spiegler	DEAP PROB-NOCAT	58.52%	63.06%	60.70%	S
Lignos	Aggressive Compounding	71.45%	52.31%	60.40%	U
Kohonen	Morfessor U+W	60.33%	59.55%	59.94%	P
Spiegler	DEAP MDL-CAT	49.83%	75.08%	59.90%	S
Nicolas	MorphAcq	67.83%	53.43%	59.78%	U
Spiegler	DEAP PROB-CAT	55.00%	56.61%	55.79%	S
-	Morfessor Baseline	81.39%	41.70%	55.14%	U

RECISION	RECALL	F-MEASURE	TYPE
6.03%	70.71%	62.52%	S
7.43%	66.59%	61.67%	S
8.38%	63.35%	60.76%	S
7.59%	55.21%	56.38%	S
6.66%	44.79%	50.03%	S
6.97%	42.98%	49.00%	P
3.07%	39.98%	48.94%	U
1.93%	42.41%	46.69%	S
3.26%	45.73%	44.46%	P
1.59%	47.41%	44.31%	P
6.25%	53.73%	43.29%	S
0.31%	29.51%	43.16%	U
5.08%	53.45%	42.36%	P
5.32%	26.70%	39.43%	U
8.98%	24.46%	37.35%	U
0.60%	14.39%	24	
.04%	99.89%	9.6	

FInnish

Morpho Challenge 2010 by Creutz et al.

AUTHOR	METHOD	PRECISION	RECALL	F-MEASURE	TYPE
Kohonen	Morfessor S+W	65.62%	69.28%	67.40%	S
Kohonen	Morfessor S+W+L	67.87%	66.43%	67.14%	S
Lignos	Base Inference	80.77%	53.76%	64.55%	U
Lignos	Iterative Compounding	80.27%	52.76%	63.67%	U
Spiegler	DEAP MDL-NOCAT	51.44%	80.95%	62.91%	S
Spiegler	DEAP PROB-NOCAT	58.52%	63.06%	60.70%	S
Lignos	Aggressive Compounding	71.45%	52.31%	60.40%	U
Kohonen	Morfessor U+W	60.33%	59.55%	59.94%	P
Spiegler	DEAP MDL-CAT	49.83%	75.08%	59.90%	S
Nicolas	MorphAcq	67.83%	53.43%	59.78%	U
Spiegler	DEAP PROB-CAT	55.00%	56.61%	55.79%	S
-	Morfessor Baseline	81.39%	41.70%	55.14%	U

RECISION	RECALL	F-MEASURE	TYPE
6.03%	70.71%	62.52%	S
7.43%	66.59%	61.67%	S
8.38%	63.35%	60.76%	S
7.59%	55.21%	56.38%	S
6.66%	44.79%	50.03%	S
6.97%	42.98%	49.00%	P
3.07%	39.98%	48.94%	U
1.93%	42.41%	46.69%	S
3.26%	45.73%	44.46%	P
1.59%	47.41%	44.31%	P
6.25%	53.73%	43.29%	S
0.31%	29.51%	43.16%	U
5.08%	53.45%	42.36%	P
5.32%	26.70%	39.43%	U
8.98%	24.46%	37.35%	U
0.60%	14.39%	24	
.04%	99.89%	9.6	

FInnish

Morpho Challenge 2010 by Creutz et al.

AUTHOR	METHOD	PRECISION	RECALL	F-MEASURE	TYPE
Kohonen	Morfessor S+W	65.62%	69.28%	67.40%	S
Kohonen	Morfessor S+W+L	67.87%	66.43%	67.14%	S
Lignos	Base Inference	80.77%	53.76%	64.55%	U
Lignos	Iterative Compounding	80.27%	52.76%	63.67%	U
Spiegler	DEAP MDL-NOCAT	51.44%	80.95%	62.91%	S
Spiegler	DEAP PROB-NOCAT	58.52%	63.06%	60.70%	S
Lignos	Aggressive Compounding	71.45%	52.31%	60.40%	U
Kohonen	Morfessor U+W	60.33%	59.55%	59.94%	P
Spiegler	DEAP MDL-CAT	49.83%	75.08%	59.90%	S
Nicolas	MorphAcq	67.83%	53.43%	59.78%	U
Spiegler	DEAP PROB-CAT	55.00%	56.61%	55.79%	S
-	Morfessor Baseline	81.39%	41.70%	55.14%	U

RECISION	RECALL	F-MEASURE	TYPE
6.03%	70.71%	62.52%	S
7.43%	66.59%	61.67%	S
8.38%	63.35%	60.76%	S
7.59%	55.21%	56.38%	S
6.66%	44.79%	50.03%	S
6.97%	42.98%	49.00%	P
3.07%	39.98%	48.94%	U
1.93%	42.41%	46.69%	S
3.26%	45.73%	44.46%	P
1.59%	47.41%	44.31%	P
6.25%	53.73%	43.29%	S
0.31%	29.51%	43.16%	U
5.08%	53.45%	42.36%	P
5.32%	26.70%	39.43%	U
8.98%	24.46%	37.35%	U
0.60%	14.39%	24	
.04%	99.89%	9.6	

FInnish

Unsettling interim observations

1. Still operating within an 2^N-time exhaustive search framework. No tractability guarantee

2. Still considering only prefixes and suffixes

3. No regard for well-formedness of the analyses, e.g. suffixes can also appear as prefixes.

seems intractable indeed: we should discuss greedy and hybrid approaches instead

current solutions still struggle even with strictly suffixing and prefixing morphologies

use hard-wired markedness penalties **or** attach a grammar (≠ lexicon) to the compressed (MDL-ized) corpus

will depart from this agenda
to explore non-MDL approaches

morpho-phonological analogies

	foods	unfed	booted	bets
foods				
unfed				
booted				
bets				

morpho-phonological analogies

	foods	unfed	booted	bets
foods		foods : unfed	foods : booted	bets : foods
unfed			unfed : booted	bets : unfed
booted				booted : bets
bets				

morpho-phonological analogies

	foods	unfed	booted	bets
foods		f d : f d	ood : oo d	s : s
unfed			ed : ed	e : e
booted				b t : b t
bets				

morpho-phonological analogies

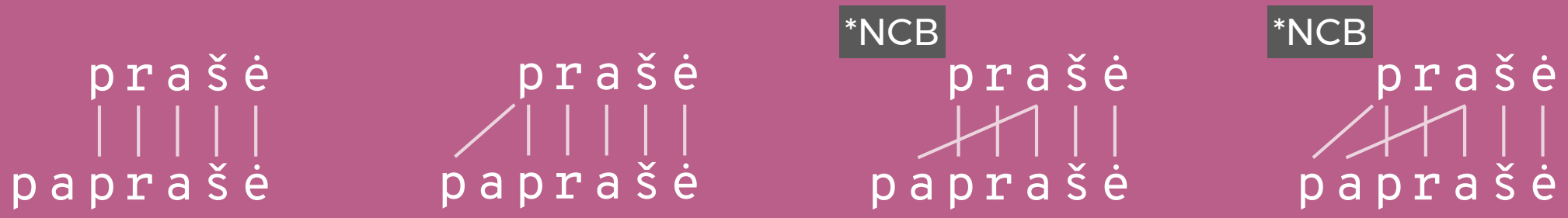
	foods	unfed	booted	bets
foods		foods : unfed	foods : booted	bets : foods
unfed			unfed : booted	bets : unfed
booted				booted : bets
bets				

morpho-phonological analogies

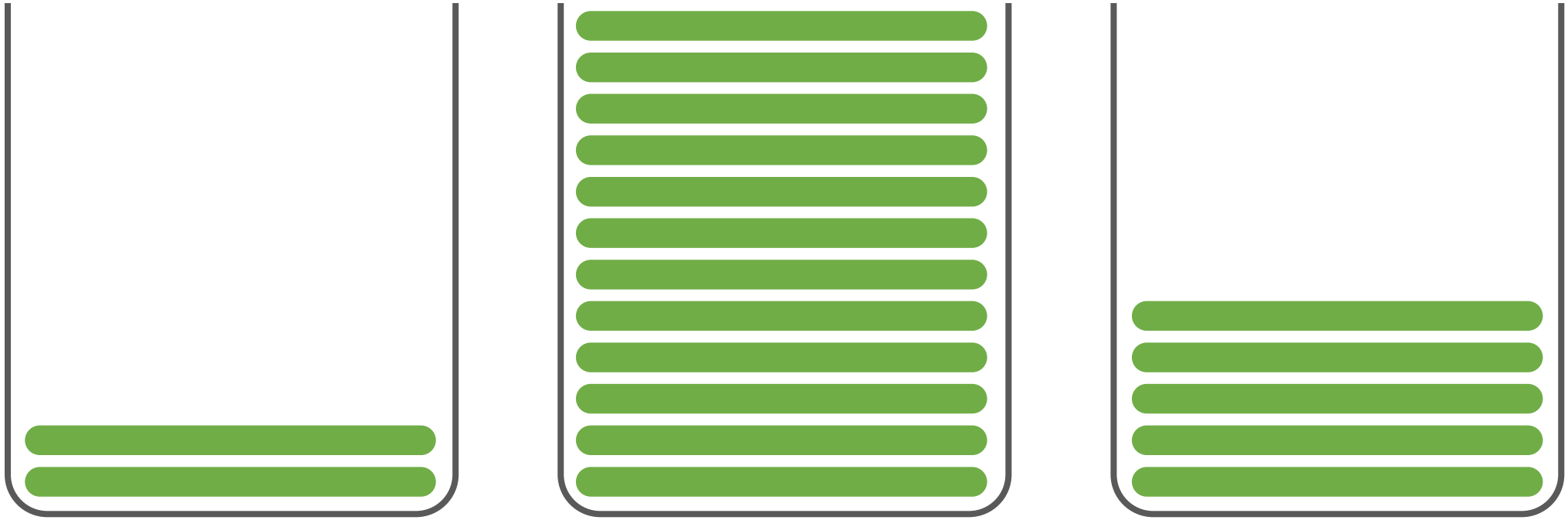
	foods	unfed	booted	bets
foods		foods : unfed loops : *unlep	foods : booted rails : *baitel	bets : foods beta : *fooda
unfed			unfed : booted unfix : *bootix	bets : unfed bits : *unfid
booted				booted : bets mooded : meds
bets				

because we anticipate **reduplication**,

some analogies will **not** have an unique solution to be singled out on grounds of general pattern detection alone

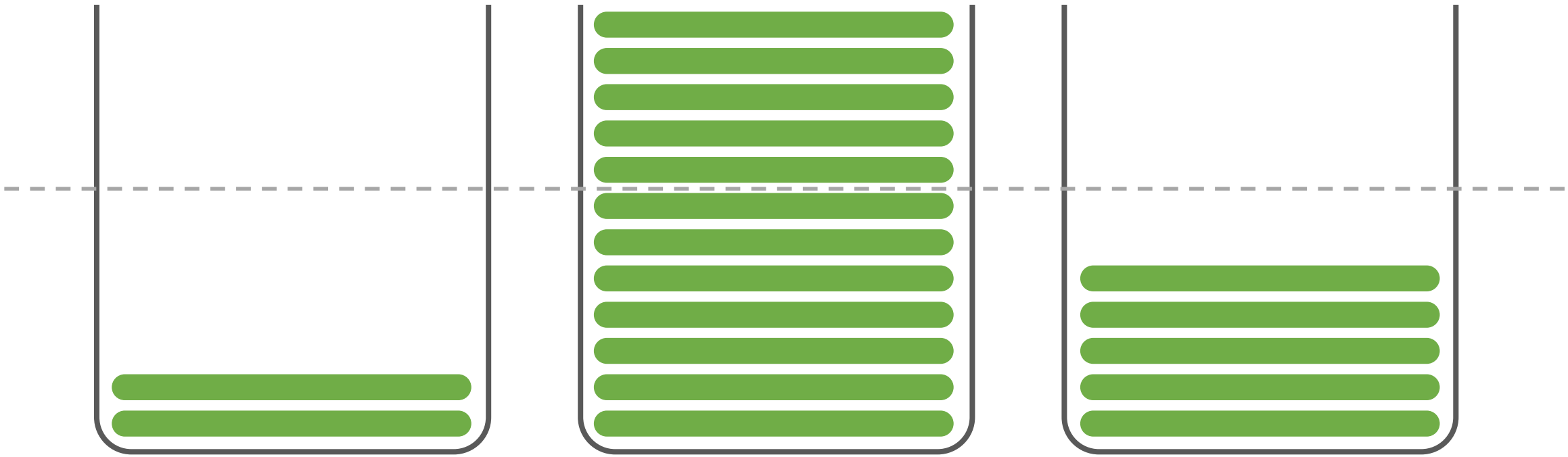


(slėpė → **pa**slėpė) (slėpė → *__sa__slėpė) (slėpė → *__pė__slėpė) (slėpė → *__sė__slėpė)



$\mathbf{C_1}$oo$\mathbf{C_2}$s → un$\mathbf{C_1}$e$\mathbf{C_2}$ **X** → **X**s **X** → **X**e

we expect **frequency** to sift the grain from the chaff



C_1oo**C_2**s → un**C_1**e**C_2**

X → **X**s

X → **X**e

morpho-phonological analogies: **issues**

rēgis, *rēgī*, *rēgem* etc. are in the corpus, but **rēg* is not

for each pair of words, also include their LCS in the lexicon

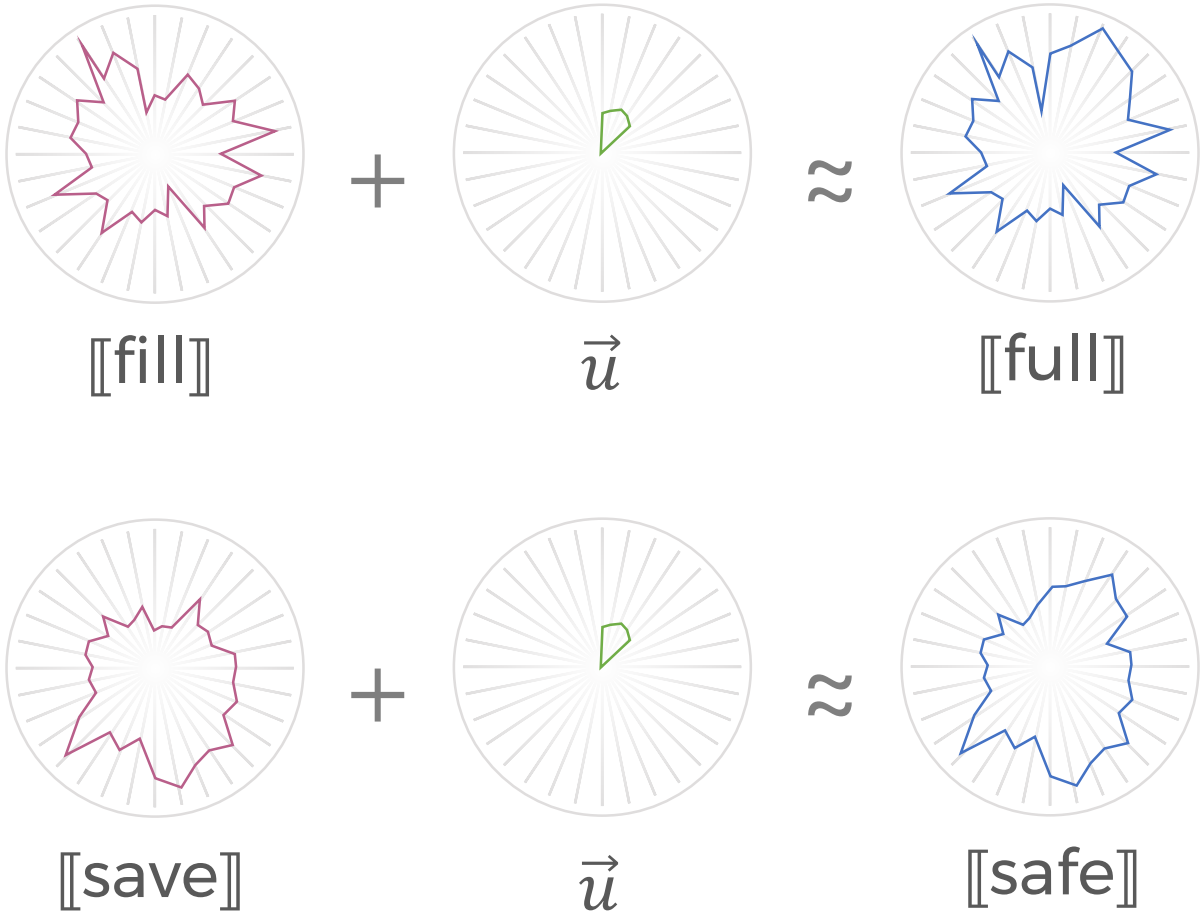
subtyping relations on morphemes:
*crate**+s*** and *box**+es*** will be counted separately

reapply top *N* morphemes on all words, the result might be just one consistent phonological rule away

alignments are done strictly on 3-phoneme stems (and 4, 5, ...)

slot splitting and slot merging: complicated

contesting the “form → meaning” unidirectionality: **Word2Vec**



the radial graphs represent numeric feature sequences

the feature sequences are obtained by reducing the dimensionality of **word neighborhood** data

additive compositionality emerges

contesting the “form → meaning” unidirectionality: **Word2Vec**

search for correlations between formal and (approximate) semantic additions

bibliography

Soricut, R. & Och, J. F. (2015) Unsupervised Morphology Induction using Word Embeddings

Mikolov, T. et al. (2013) Distributed Representations of Words and Phrases and their Compositionality

Conclusions

MDL etc.	Self-evident. Intractable in pure form. Boosting accuracy introduces complexity. Generalization to non-concatenative (NC) morphologies appears difficult. Research **ample**.
Analogies	Seemingly tractable in pure form. Naturally generalizes to NC morphologies. Research **scarce**. Promising so far.
Word2Vec	Research **in its infancy**. Extraordinary theoretical appeal, but any conclusions inevitably premature.
Connectionism	**Not covered here.** Sizable body of research. Rapidly gaining traction in industry. Highly distinct variety of artificial learners.

THANK YOU

THANK YOU