Corpus-driven Greek Language Learning

James K. Tauber

jktauber.com

@jtauber

C. S. Lewis 1/2

> “The great gain was that I very soon became able to understand a great deal without (even mentally) translating it; I was beginning to think in Greek. That is the great Rubicon to cross in learning any language.
>
> Those in whom the Greek word lives only while they are hunting for it in the lexicon, and who then substitute the English word for it, are not reading the Greek at all; they are only solving a puzzle.

C. S. Lewis 2/2

> “The very formula, ‘Naus means a ship’, is wrong. Naus and ship both mean a thing, they do not mean one another.
>
> Behind Naus…we want to have a picture of a dark, slender mass with sail or oars, climbing the ridges, with no officious English word intruding.”

Surprised by Joy: The Shape of My Early Life, pp. 140–141.

The Solution

"... **comprehensible input** is the crucial and necessary ingredient for the acquisition of language."

Stephen Krashen

The Traditional Approach

vocabulary driven by paradigm being learnt

vocabulary not shown in context

hard to show much real text early on

The Myth of Vocabulary Coverage for the Greek New Testament

The **10** most common words account for **37%** of the text

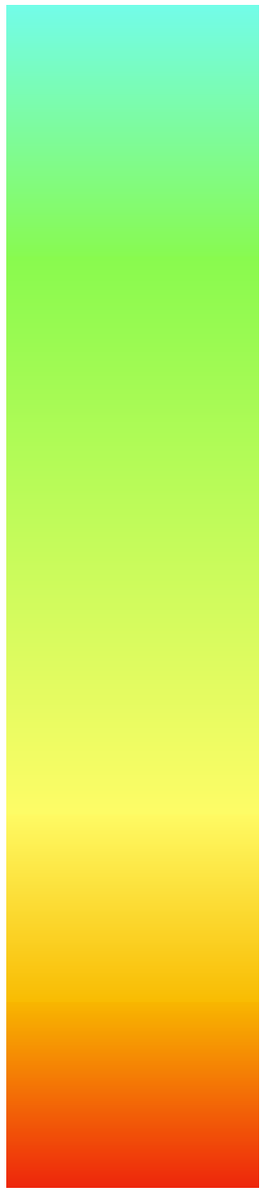
The **100** most common words account for **66%** of the text

The Myth of Vocabulary Coverage for the Greek New Testament

If you learn the **100 most common words**, you'll…

- know at least one word in **99.9%** of verses
- know at least **50%** of words in **91.3%** of verses
- know at least **75%** of words in **24.4%** of verses
- know at least **90%** of words in **2.1%** of verses
- know at least **95%** of words in **0.6%** of verses
- know **all** words in **0.4%** of verses

Levels of Understanding



too easy for the student

the right level for **speed reading**

the right level for **extensive reading**

the right level for **intensive reading**

the right level for **intensive reading**
with suitable scaffolding

too hard or requiring too much scaffolding

Use Cases

warn teacher / content author where passage might need adaptation

search for passages appropriate for a particular student

sort passages in a graded sequence

Example at 98%

You live and work in Tokyo. Tokyo is a big city. More than 13 million people live around you. Of borgle you are never, but you are always lonely.

Every morning, you get up and take the train to work. Every night, you take the train again to go home. The train is always crowded.

When people ask about your work, you tell them, “I move papers around.” It’s a joke, but it’s also true. You don’t like your work.

Tonight you are returning home. It’s late at night. Is shnooling no one. Sometimes you see don’t to shnool all day.

You are tired. You are so tired…

https://magisterp.com/2016/08/21/how-comprehensible-must-reading-be/

Example at 95%

In the morning, you start again. You shower, get dressed, and for pocklent walk. You move slowly, half-awake. Then, suddenly, you stop. Something is different. Are streets of fossit. Really of fossit.

There are no people. No cars. Nothing. “A where dowargle?” you ask yourself.

Suddenly, there is by quapen loud—a police car. It speeds by and almost hits you. It crashes into a store across the street!

Then, another for farfoofles police car. The police officer sees you. “Off the street!” he shouts. “Go home, lock your door!”

“What? Why?” you shout back. But it’s too late. He is gone.

https://magisterp.com/2016/08/21/how-comprehensible-must-reading-be/

Example at 90%

“Of prippy fy what’s?” you ask yourself. Suddenly, a man runs by. Viggling toward he is crawn for kofoon. There is blood all over his shirt.

“In order to baboot!” you shout, but he doesn’t stop. You follow him. For kofoon outside, you stop. Is lying loopity of on the ground. She is not moving.

“Hey!” you shout. “Are you OK?” She doesn’t answer. Are closed her nawiesly. Her fingers chay are moving. Open, close; open, close.

“She’s alive!” you say to yourself. “No! Her don’t of gleep!” someone the frickles. You look up. Three people are waving at you from across the street.

Example at 80%

“Of bingle for help!” you shout. “This is dying loopitish!” You put your fingers on her neck. Nothing. Flid her not a weafling is.

You take out by joople your and 119 the of bingle, the emergency number in Japan. There’s no answer!

Then muchy you that you have befourn assengle a new. It’s with gutring your, Evie. Hunwres she at Tokyo University. You assengle play.

“...if you get this...” Evie says. “... now vickarn I can’t... of passit important is...” Suddenly, to dingle, around looks she. “Oh no, they’re here! Cripettly... of frib!

Them wasple OF FRIB ON!...” BEEP! of the assengle to parantles. Then gratoonly something behind you...

https://magisterp.com/2016/08/21/how-comprehensible-must-reading-be/

The Myth of Vocabulary Coverage for the Greek New Testament

If you learn the **100 most common words**, you'll…

- know at least one word in **99.9%** of verses
- know at least **50%** of words in **91.3%** of verses
- know at least **75%** of words in **24.4%** of verses
- know at least **90%** of words in **2.1%** of verses
- know at least **95%** of words in **0.6%** of verses
- know **all** words in **0.4%** of verses

Frequency

Frequency-based vocabulary lists are important post-beginner.

For beginners, passages either need to be constructed, heavily adapted, or we must not rely on corpus frequency as the primary ordering priority.

Pericope Experiment

1. 1 John 4:1–4:21

2. 1 John 5:1–5:12 (80.31%; 25 additional words)

3. 2 John 1:1–1:3 (84.75%; 8 additional words)

…

11. John 12:42–12:50 (92.73%; 12 additional words)

12. John 1:1–1:9 (95.33%; 5 missing words)

13. John 5:31–5:47 (93.49%)

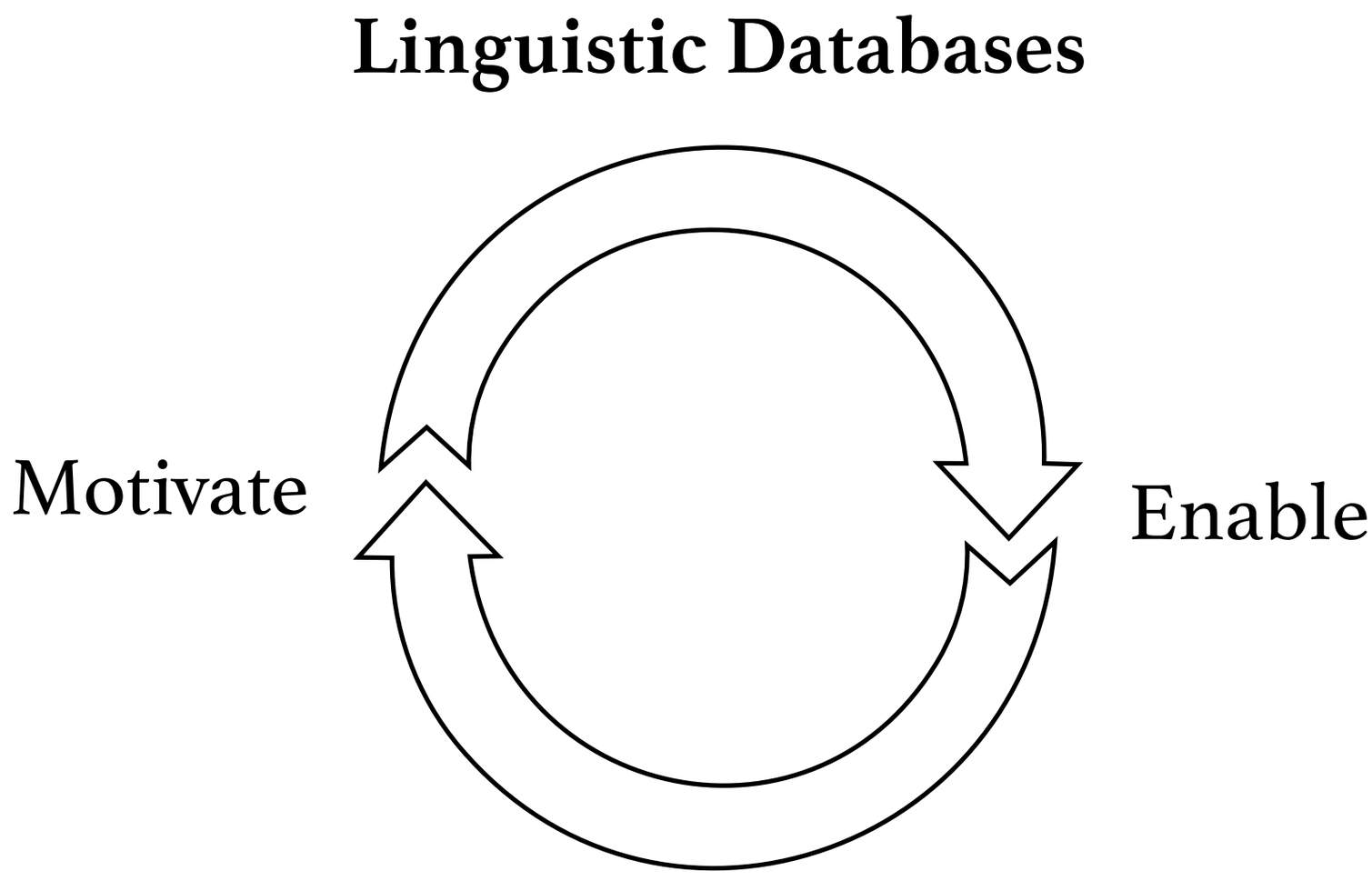
Caveats and Limitations

Didn’t consider:

- necessary spaced repetition
- inflectional morphology
- syntax
- idioms
- derivational morphology / cognates

Why weren’t they considered?

Because they weren’t in an open, machine-actionable format!


Linguistic Databases
Motivate
Enable
Learning Tools

Research Program

linguistics, descriptive grammar	*digital philology, corpus linguistics*	*learning science, applied linguistics*
how to model	how to model	how to model
language	**text**	**knowledge**

educational statistics, learning analytics, data science

MorphGNT

```
041801 RD ----APN- Ταῦτα Ταῦτα ταῦτα οὗτος
041801 V- -AAPNSM- ⸀εἰπὼν εἰπὼν εἰπών λέγω
041801 N- ----NSM- Ἰησοῦς Ἰησοῦς Ἰησοῦς Ἰησοῦς
041801 V- 3AAI-S-- ἐξῆλθεν ἐξῆλθεν ἐξῆλθε(ν) ἐξέρχομαι
041801 P- -------- σὺν σὺν σύν σύν
041801 RA ----DPM- τοῖς τοῖς τοῖς ὁ
041801 N- ----DPM- μαθηταῖς μαθηταῖς μαθηταῖς μαθητής
041801 RP ----GSM- αὐτοῦ αὐτοῦ αὐτοῦ αὐτός
041801 P- -------- πέραν πέραν πέραν πέραν
```

Word Lists by Frequency

```
awk '{print $7}' | sort |
uniq -c | sort -r
```

Analytical Lexicon

```
awk '{print $6,$2,$3,$7}' |
sort | uniq
```

James Tauber
@jtauber

Those of you who know me well understand why every new Greek Reader published makes me simultaneously jump for joy and cry a little.

12:05 AM - 13 Oct 2018

1 Retweet **13** Likes

2 1 13

James Tauber @jtauber · 13 Oct 2018

jump for joy because I'm a huge believer in comprehensible input with appropriate scaffolding and respect tremendously the effort that goes into producing these readers

1 2

James Tauber @jtauber · 13 Oct 2018

but cry a little because, as I've ranted for years, it's such a shame this data is locked in printed books rather than published in an open, machine actionable format with the printed version just generated from that data

1 1 11

Readers

4.2.1 Καὶ τῇ ἡμέρᾳ τῇ τρίτῃ γάμος[1] ἐγένετο ἐν Κανὰ[2] τῆς Γαλιλαίας, καὶ ἦν ἡ μήτηρ τοῦ Ἰησοῦ ἐκεῖ· **2** ἐκλήθη δὲ καὶ ὁ Ἰησοῦς καὶ οἱ μαθηταὶ αὐτοῦ εἰς τὸν γάμον.[3] **3** καὶ ὑστερήσαντος[4] οἴνου λέγει ἡ μήτηρ τοῦ Ἰησοῦ πρὸς αὐτόν· Οἶνον οὐκ ἔχουσιν. **4** καὶ λέγει αὐτῇ ὁ Ἰησοῦς· Τί ἐμοὶ καὶ σοί, γύναι; οὔπω[5] ἥκει[6] ἡ ὥρα μου. **5** λέγει ἡ μήτηρ αὐτοῦ τοῖς διακόνοις·[7] Ὅ τι ἂν λέγῃ ὑμῖν ποιήσατε. **6** ἦσαν δὲ ἐκεῖ λίθιναι[8] ὑδρίαι[9] ἓξ[10] κατὰ τὸν καθαρισμὸν[11] τῶν Ἰουδαίων κείμεναι,[12] χωροῦσαι[13] ἀνὰ[14] μετρητὰς[15] δύο ἢ τρεῖς. **7** λέγει αὐτοῖς ὁ Ἰησοῦς· Γεμίσατε[16] τὰς ὑδρίας[17] ὕδατος· καὶ ἐγέμισαν[18] αὐτὰς ἕως ἄνω.[19] **8** καὶ λέγει αὐτοῖς· Ἀντλήσατε[20] νῦν καὶ φέρετε τῷ ἀρχιτρικλίνῳ·[21] οἱ δὲ ἤνεγκαν. **9** ὡς δὲ ἐγεύσατο[22] ὁ ἀρχιτρίκλινος[23] τὸ ὕδωρ οἶνον γεγενημένον, καὶ οὐκ ᾔδει πόθεν[24] ἐστίν, οἱ δὲ διάκονοι[25] ᾔδεισαν οἱ ἠντληκότες[26] τὸ ὕδωρ, φωνεῖ τὸν νυμφίον[27] ὁ ἀρχιτρίκλινος[28] **10** καὶ λέγει αὐτῷ· Πᾶς ἄνθρωπος πρῶτον τὸν καλὸν οἶνον τίθησιν, καὶ ὅταν μεθυσθῶσιν[29] τὸν ἐλάσσω·[30] σὺ τετήρηκας τὸν καλὸν οἶνον ἕως ἄρτι. **11** ταύτην ἐποίησεν ἀρχὴν τῶν σημείων ὁ Ἰησοῦς ἐν Κανὰ[31] τῆς Γαλιλαίας καὶ ἐφανέρωσεν τὴν δόξαν αὐτοῦ, καὶ ἐπίστευσαν εἰς αὐτὸν οἱ μαθηταὶ αὐτοῦ.

Readers

---

[1]γάμος, ου, ὁ – *a marriage, wedding, wedding-feast*
[2]Κανά, ἡ – *Cana*
[3]γάμος, ου, ὁ – *a marriage, wedding, wedding-feast*
[4]ὑστερέω – AAP GSM – *I am lacking, fall short, suffer need*
[5]οὔπω – *not yet*
[6]ἥκω – PAI 3S – *I have come, am present*
[7]διάκονος, ου, ὁ/ἡ – *a waiter, servant, administrator*
[8]λίθινος, η, ον – *made of stone*
[9]ὑδρία, ας, ἡ – *a water pot*
[10]ἕξ – *six*
[11]καθαρισμός, οῦ, ὁ – *cleansing, purifying, purification*
[12]κεῖμαι – PMP NPF – *I lie, recline, am laid*
[13]χωρέω – PAP NPF – *I make room, go, receive*
[14]ἀνά – *and, apiece, by, each, every, in, through*
[15]μετρητής, οῦ, ὁ – *a measure, amphora*
[16]γεμίζω – AAD 2P – *I fill, load*
[17]ὑδρία, ας, ἡ – *a water pot*
[18]γεμίζω – AAI 3P – *I fill, load*
[19]ἄνω – *up, above, things above, heaven*
[20]ἀντλέω – AAD 2P – *I draw, draw out*
[21]ἀρχιτρίκλινος, ου, ὁ – *master of the feast*
[22]γεύομαι – AMI 3S – *I taste, experience*
[23]ἀρχιτρίκλινος, ου, ὁ – *master of the feast*
[24]πόθεν – *whence*
[25]διάκονος, ου, ὁ/ἡ – *a waiter, servant, administrator*
[26]ἀντλέω – XAP NPM – *I draw, draw out*
[27]νυμφίος, ου, ὁ – *a bridegroom*
[28]ἀρχιτρίκλινος, ου, ὁ – *master of the feast*
[29]μεθύω – APS 3P – *I am drunk*
[30]ἐλάσσων/ἔλασσον – *less, smaller, inferior*
[31]Κανά, ἡ – *Cana*

Generating a Greek Reader

```
./frequency_exclusion.py 31 > exclude.txt

# edit exclude to your liking

./make_glosses.py "John 4:1-11"
    --exclude exclude.txt
    > glosses.yaml

# edit glosses.yaml to your liking

./make_headwords.py "John 4:1-11”
    --exclude exclude.txt
    > headwords.yaml
```

Generating a Greek Reader

```
./reader.py "John 4:1-11”
    --headwords headwords.yaml
    --glosses glosses.yaml
    --exclude exclude.txt
    --typeface "Skolar PE"
    > reader.tex
```

Generating a Greek Reader

https://github.com/
jtauber/greek-reader

Written in Python 3 and open source under an MIT license

Generating a Greek Reader

- electronic text *(SBLGNT)*
- lemmatisation *(MorphGNT)*
- parse codes *(MorphGNT)*
- glosses *(Dodson)*

Books become “UI” derived from databases

Morphology

inflected forms versus lemmas

lists versus rules

lexical redundancy

Cost of Learning a Form

λέγει > λέγεις

λέγει > εἶπεν

λέγει > ἀντιλέγει ?

λέγει > λόγος ?

Ordering Within Morphology

no need to wait to learn about **athematic** verbs and **δίδωμι** to learn **ἔδωκεν** or **δός**.

Ordering Within Morphology

ἔδωκεν

δοῦναι

ἐδόθη

δώσω

δώσει

Ten Most Common Verb Parses (out of 379)

aorist active 3rd singular

present active 3rd singular

aorist active 3rd plural

aorist active infinitive

present active participle nominative singular masculine

aorist active participle nominative singular masculine

imperfect active 3rd singular

present active 1st singular

present active infinitive

present active participle nominative plural masculine

Morphological Lexicon

principal parts

stem relationships

word formation

transparency of lexical relatedness

Lexical Relatedness

Ἰταλία:Ἰταλικός

Γαλατία:Γαλατικός

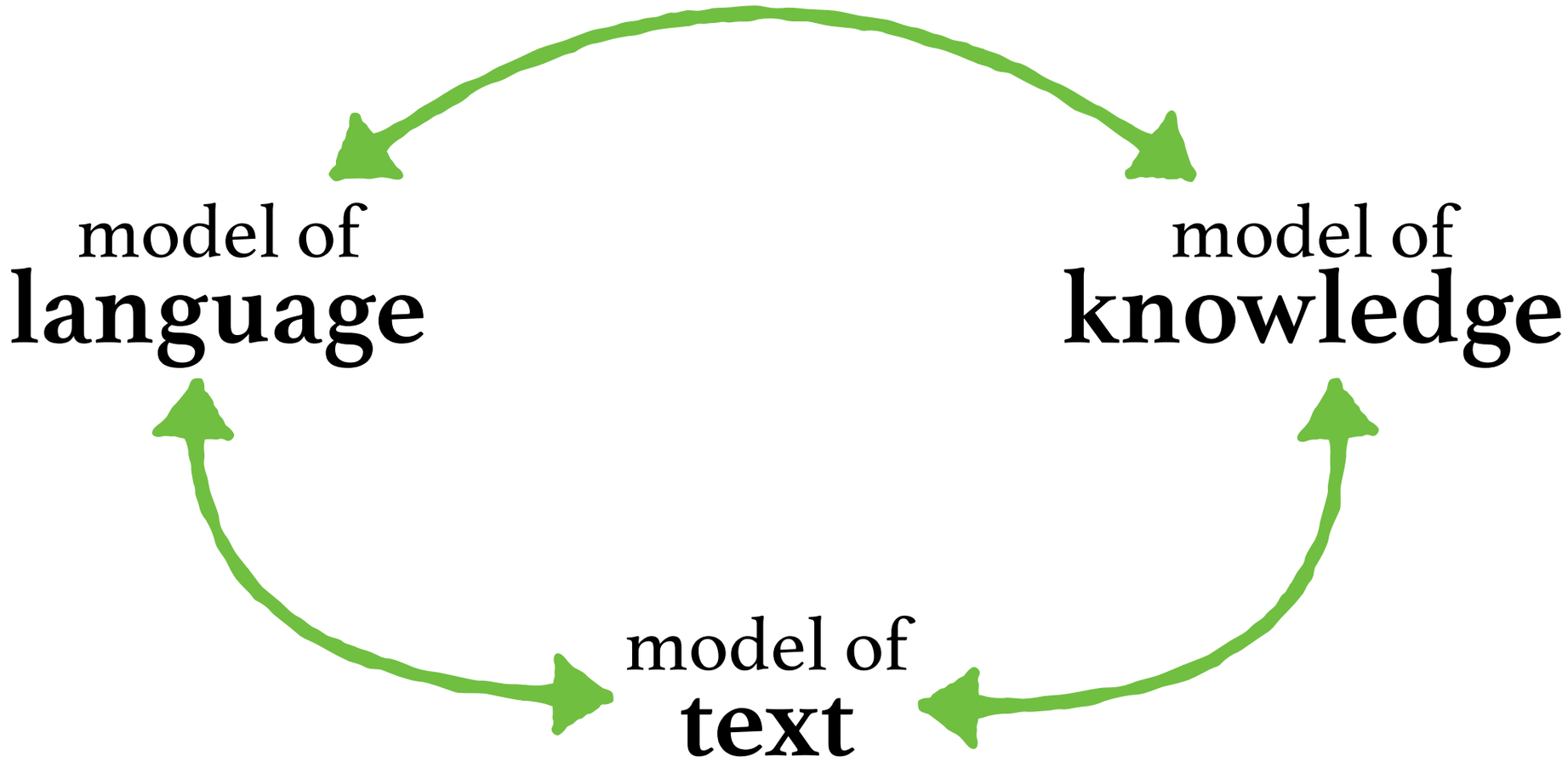
Πόντος:Ποντικός

Στοϊκός

εἰρήνη:εἰρηνικός

ὄνος:ὀνικός

https://jktauber.com/2017/05/31/modelling-derivational-morphology/


model of
language
model of
knowledge
model of
text

Adaptive Reading Environments

what's needed to understand an upcoming passage

what the student has already seen

what the student has inquired about

what is at an optimal recall interval

what the student is good or not so good at understanding (based on explicit assessment including meta-cognitive questions)

Perseus Digital Library

Scaife Viewer

Browse Library | Text Search

or try the **Iliad** or Plato's **Apology**.

1,834 works in **2,320** editions and translations (1,178 in Greek and 612 in Latin)
50.1 million words (20.1 million in Greek, 15.6 million in Latin)

HEDERA

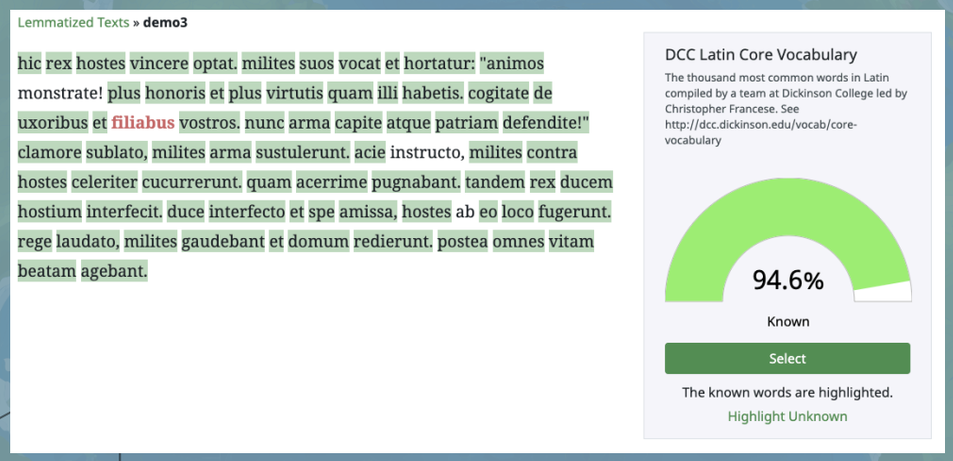
reading that fits

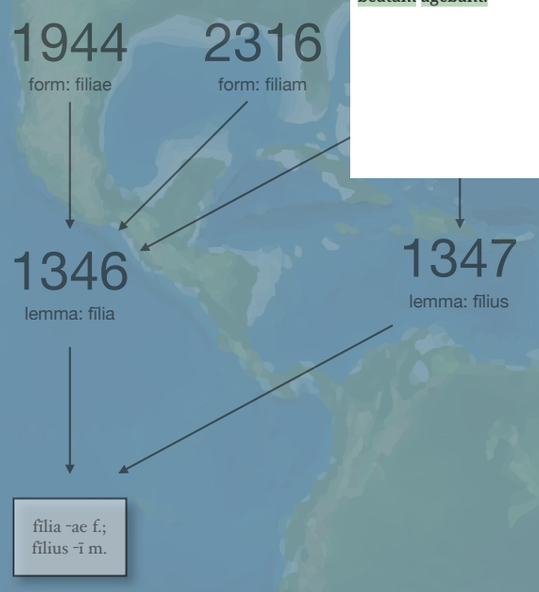
http://hederaproject.org

HARVARD
DEPARTMENT OF THE CLASSICS
ACADEMIC
TECHNOLOGY
for the FACULTY OF ARTS & SCIENCES

Teachers

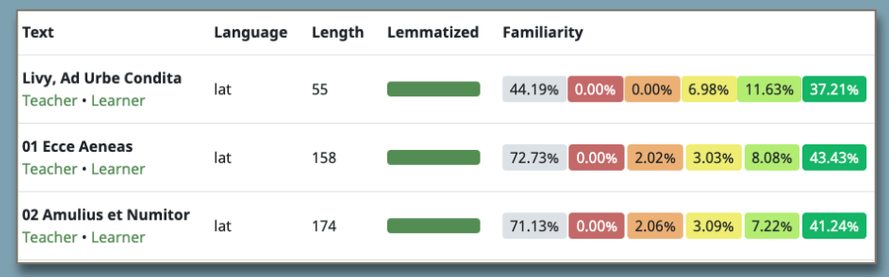
"How hard will this be for my students to read?"





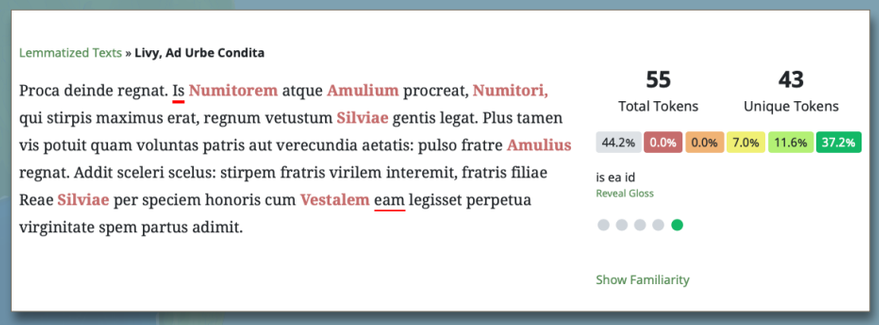
Life-Long Learners

"What would be an appropriate text to read next."



Students

"How well do I know these words?"



Summary

language acquisition needs lots of comprehensible input

we can sequence texts and vocabulary hand-in-hand to achieve this (and consider morphology, syntactic constructions, etc too)

this all relies on richly annotated texts (and other machine-actionable language description)

it can be done statically but the real power comes in adapting to students as they learn

jktauber.com