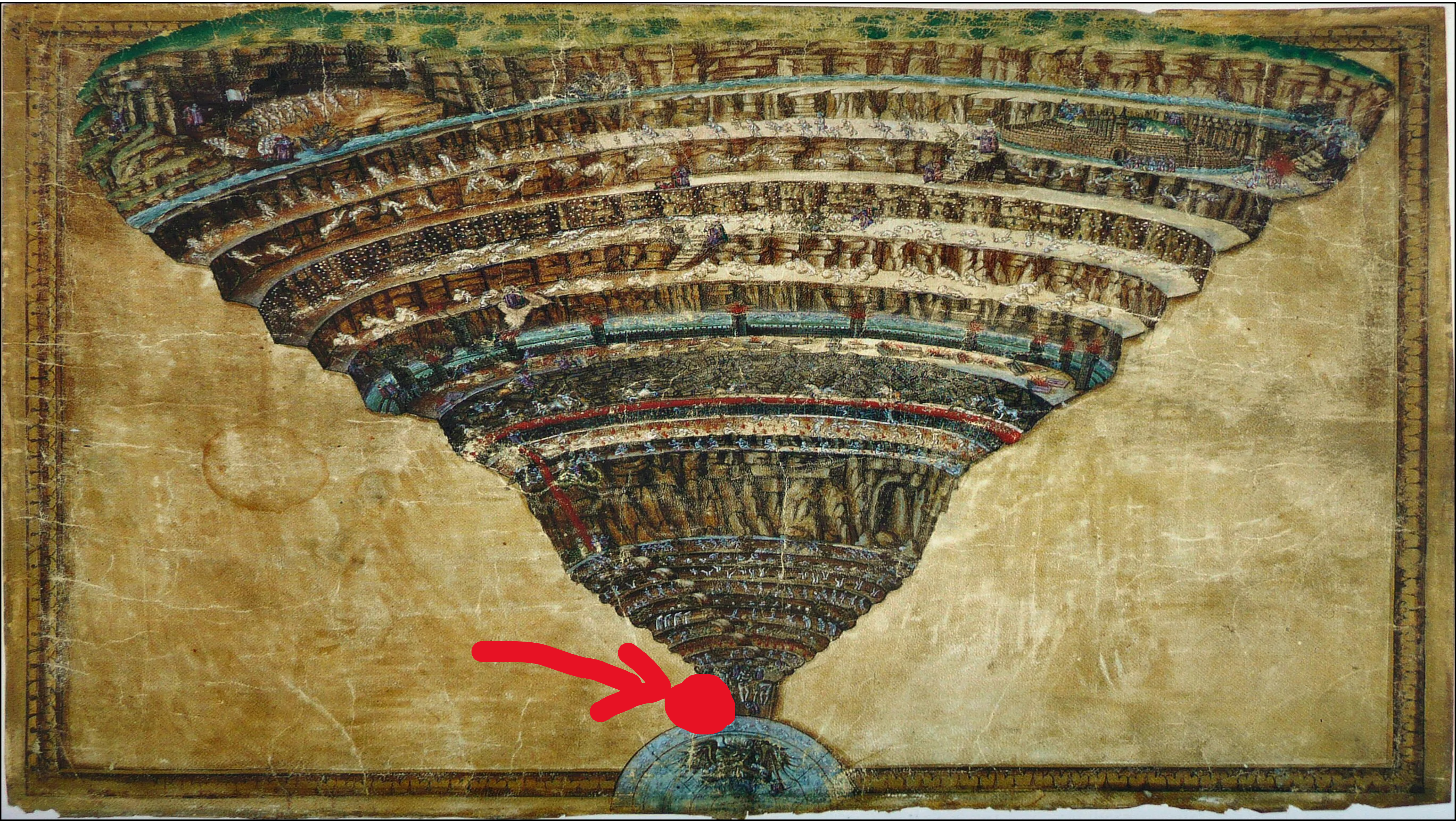
THE RARE, THE NEW, AND THE EXPRESSIVE

Dante's linguistic creativity and a few computational models to take it seriously

Giovanni Cassani

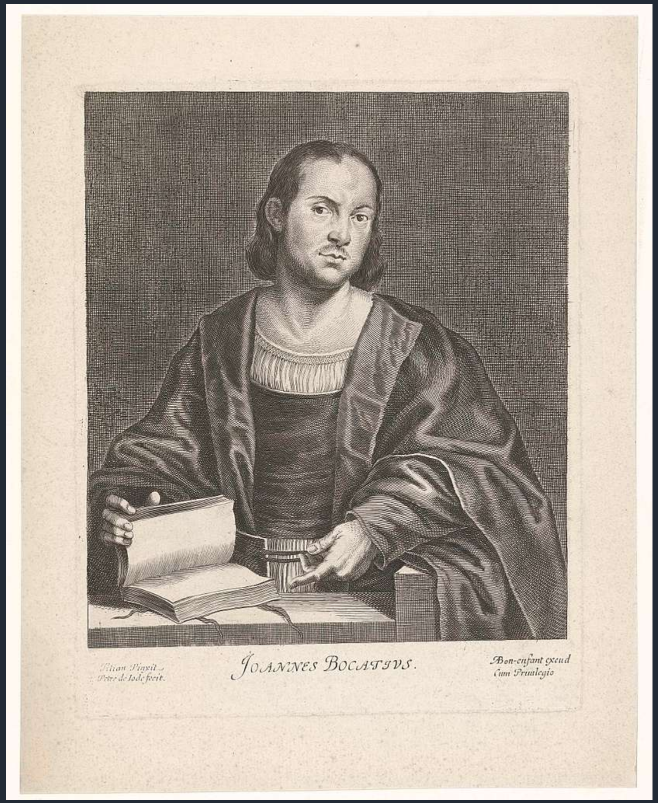
Dante Symposium

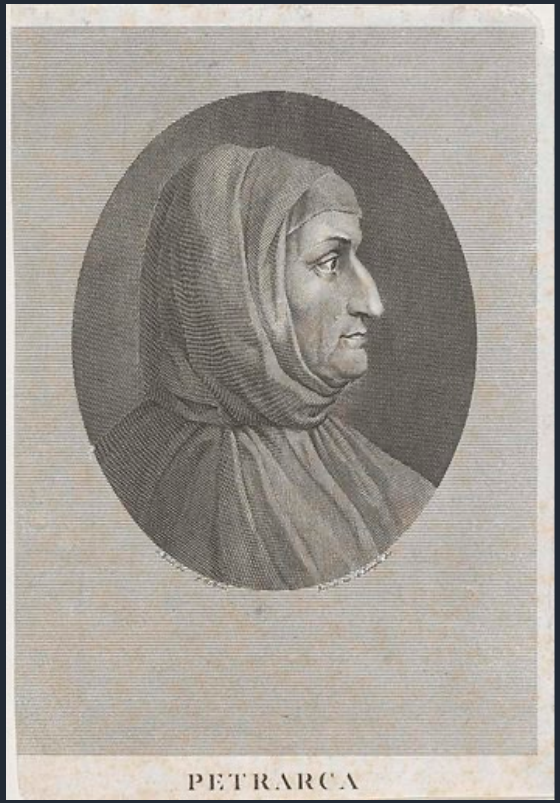
Nov 17th



DANTE'S LINGUISTIC CREATIVITY

Dante and the Italian language





Giovanni Cassani Dante Symposium Nov 17th

Productive morphology

Non altrimenti stupido si turba
lo montanaro, e rimirando ammuta,
quando rozzo e salvatico *s'inurba*.

Purgatorio, XVI, 69

perché non satisface a' miei disii?
Già non attendere' io tua dimanda,
s'io *m'intuassi*, come tu *t'inmii*

Paradiso, IX, 81

Expressive language

Del nostro ponte disse: «O *Malebranche*,
ecco un de li anzian di Santa Zita!

[…]

Tutti gridaron: «Vada *Malacoda*!»;
per ch'un si mosse - e li altri stetter fermi -,
e venne a lui dicendo: «Che li approda?».

[…]

Ma quel demonio che tenea sermone
col duca mio, si volse tutto presto,
e disse: «Posa, posa, *Scarmiglione*!».

[…]

«Tra'ti avante, *Alichino*, e *Calcabrina*»,
cominciò elli a dire, «e tu, *Cagnazzo*;
e *Barbariccia* guidi la decina.

Libicocco vegn'oltre e *Draghignazzo*,
Ciriatto sannuto e *Graffiacane*
e *Farfarello* e *Rubicante* pazzo.

Inferno, XXI, various verses

Made-up language

«*Pape Satàn, pape Satàn aleppe!*»,
cominciò Pluto con la voce chioccia;
e quel savio gentil, che tutto seppe,

Inferno, VII, 3

«"Ara bell'Ara discesa Cornara”» (Gioacchino Belli, translation in Milanese dialect)

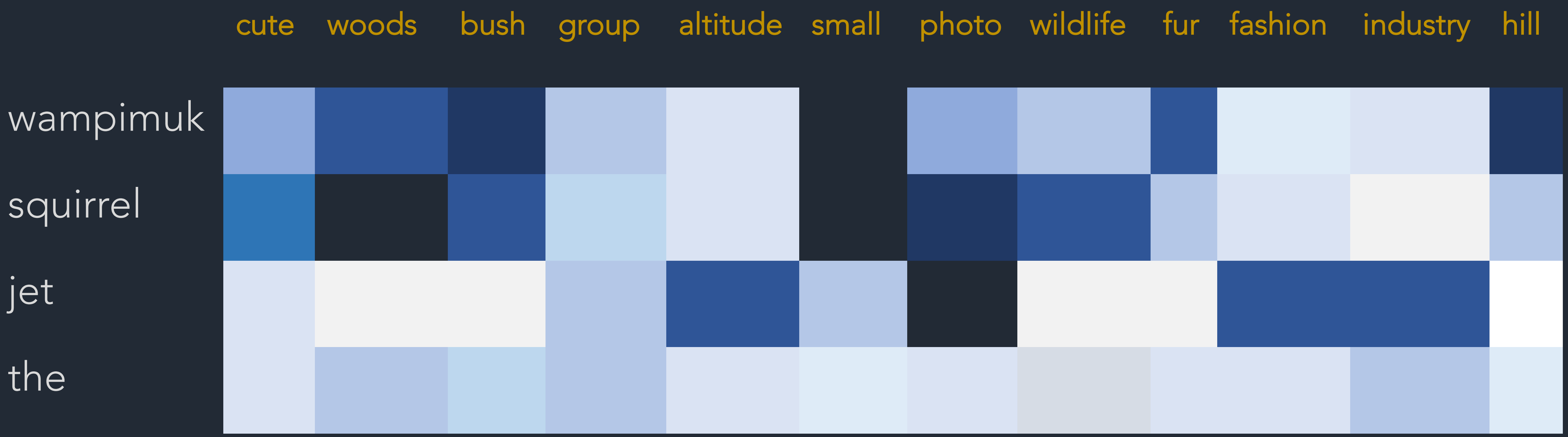
«Pala 'nzità, pala 'nzitata, allippa!» (Nino Martoglio,. Translation in Sicilian dialect)

FOREWORD: EMBEDDINGS

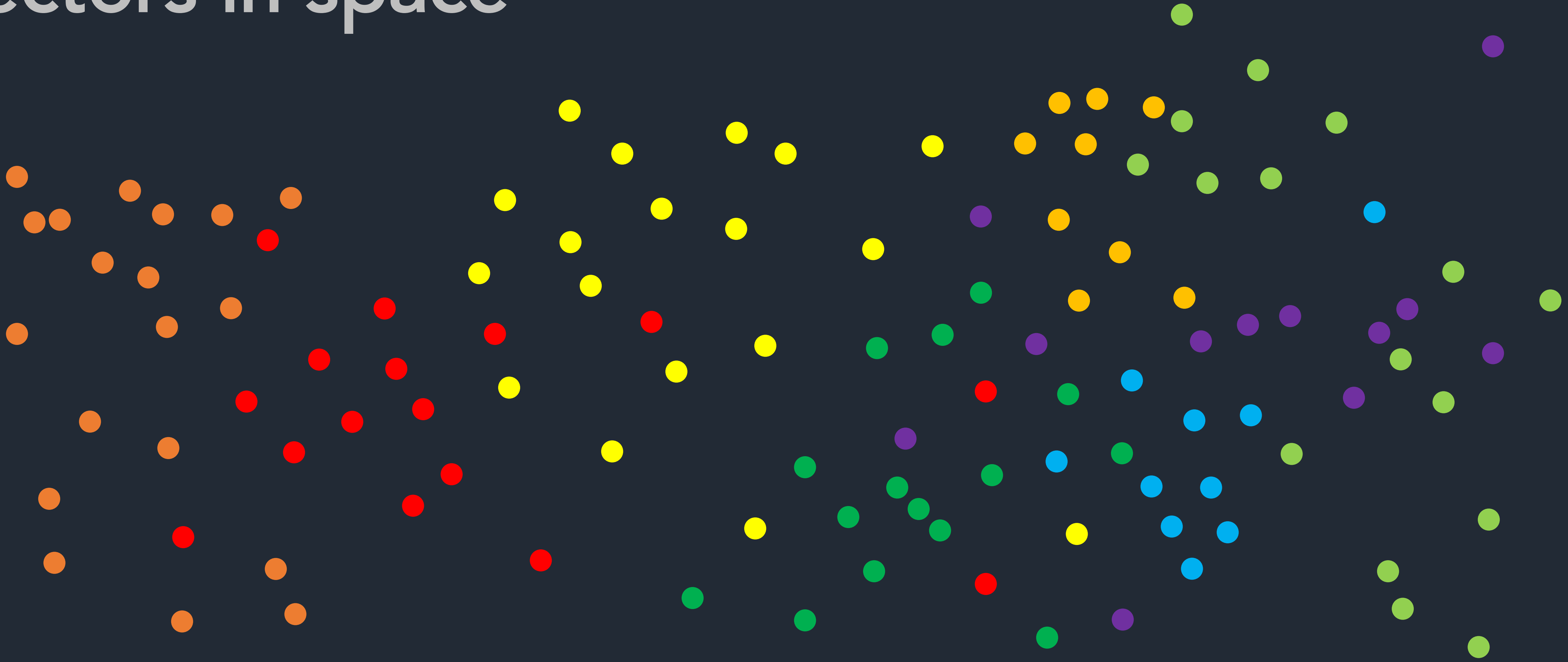
By the company it keeps

The cute *wampimuk* hid behind the bush.
Yesterday, I went to hike in the woods and saw a rare *wampimuk*.
Wampimuks live in small groups at an altitude of approximately 1000m.
A photo of a *wampimuk* won the world wildlife photography prize.
Wampimuks’ fur is very coveted in the fashion industry.
My dog chased a *wampimuk* down the hill and got lost.

From context to vectors



Vectors in space



What if a word doesn’t keep a company?
What if we see it for the first time?
What if it has been made-up?

Every word has been an innovation at some point.
Can we study it with NLP while it’s fresh and understand more about it?

COMPUTATIONAL APPROACHES TO MODEL INNOVATIONS

Combinatorics

Even if the contexts in which *s'inurba, m'intuassi*, or *t'inmii* are not very informative, these new words re-combine known morphemes, such as *in-, urb, tu, mi*.

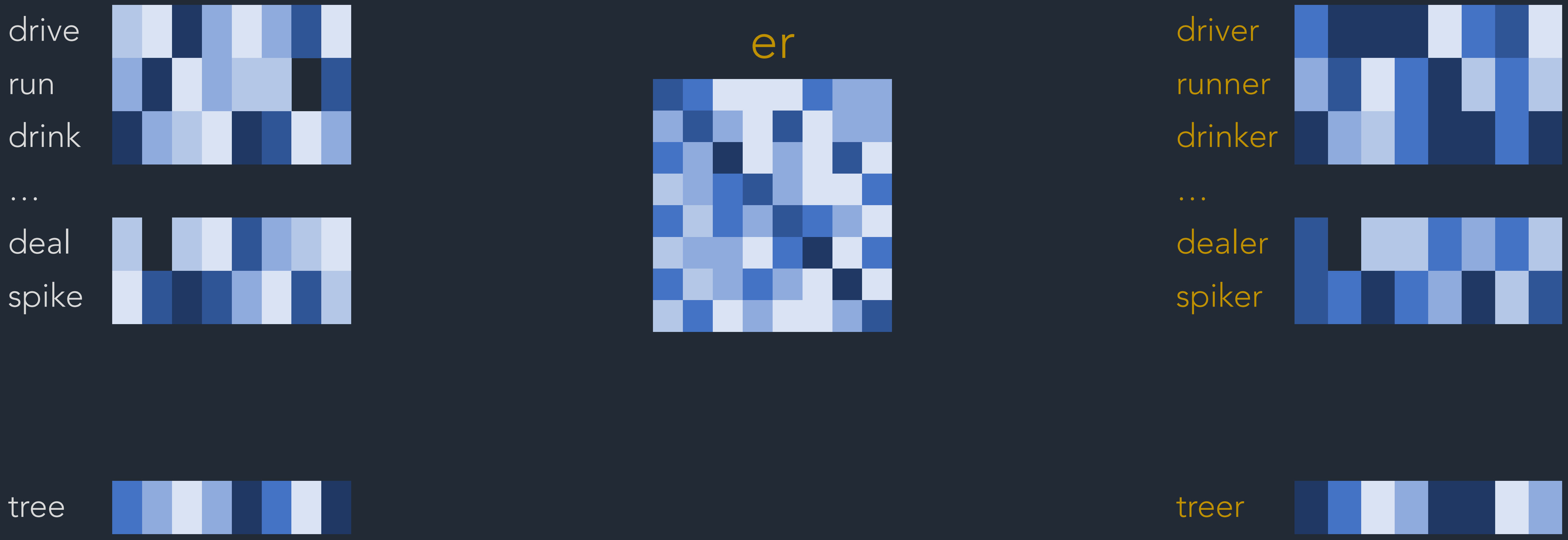
How do we derive the whole word meaning from the meaning of the parts?

FRACSS

Functional Representations of Affixes in Compositional Semantic Space

We can represent affixes as functions that we apply to stems to transform them. Affixes are represented as matrices, stems as vectors: we can multiply the two and obtain a new vector in the same semantic space.

How does it work?



drive
run
drink
…
deal
spike

tree

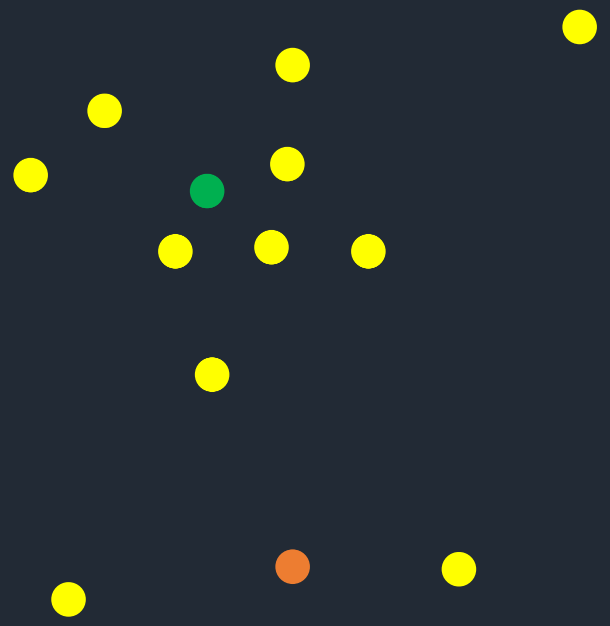
er

driver
runner
drinker
…
dealer
spiker

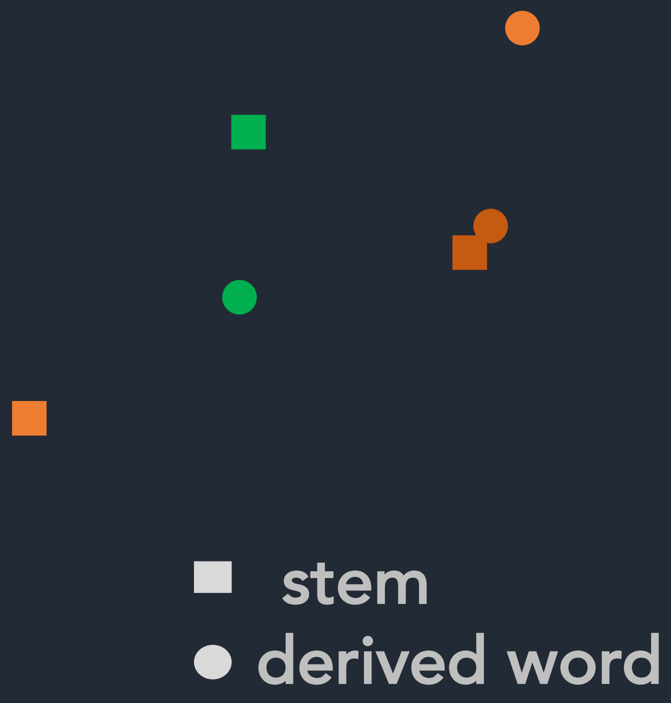
treer

Making sense of new forms

neighborhood density



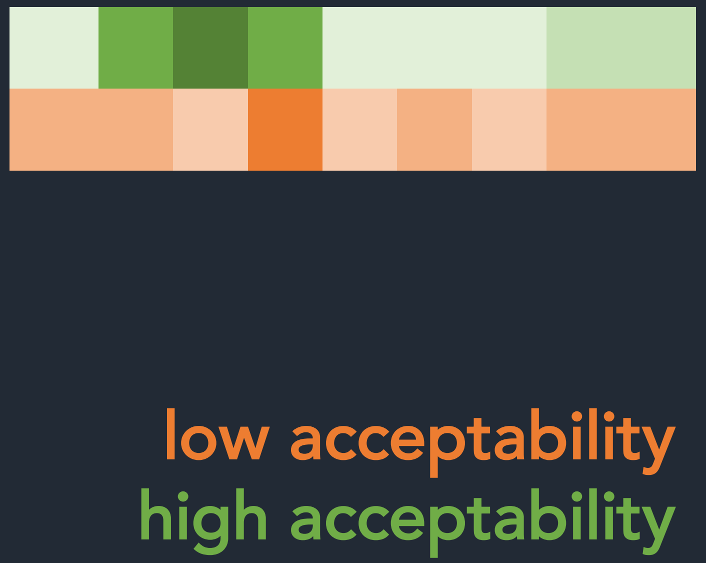
stem proximity



■ stem

● derived word

vector entropy



low acceptability

high acceptability

CAOSS

Compounding as Abstract Operation in Semantic Space

Very similar to FRACSS but looks at compounds, estimating functions H and M that characterize general properties of heads and modifiers as learned from a set of compounds, so

new compound = w1*M + w2*H

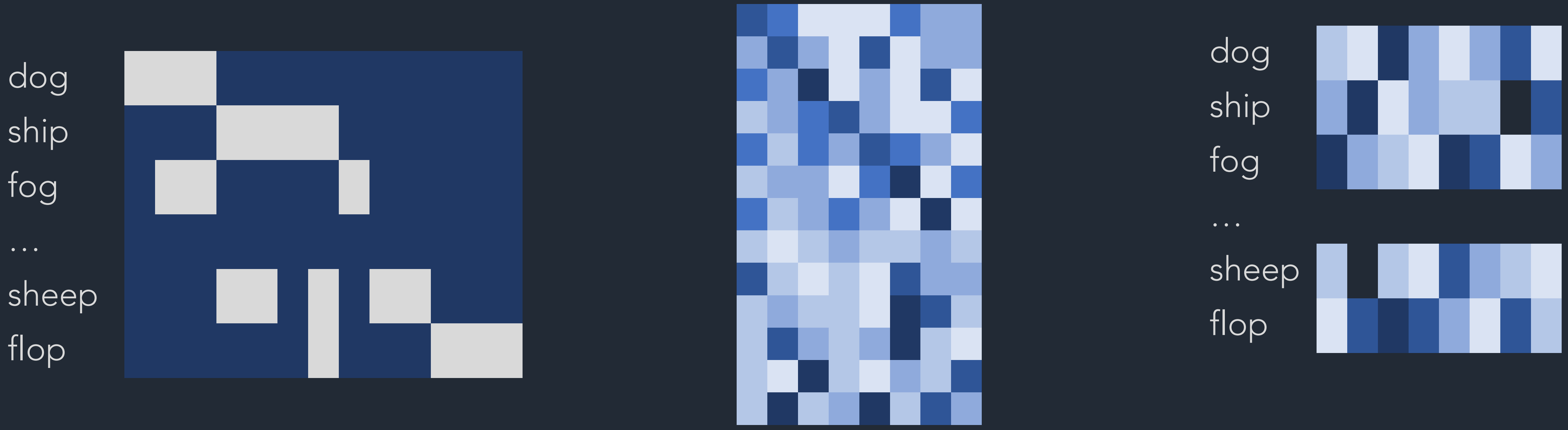
Interim summary

We can get **word** vectors for words we've never seen exploiting derivational morphology and compounding.

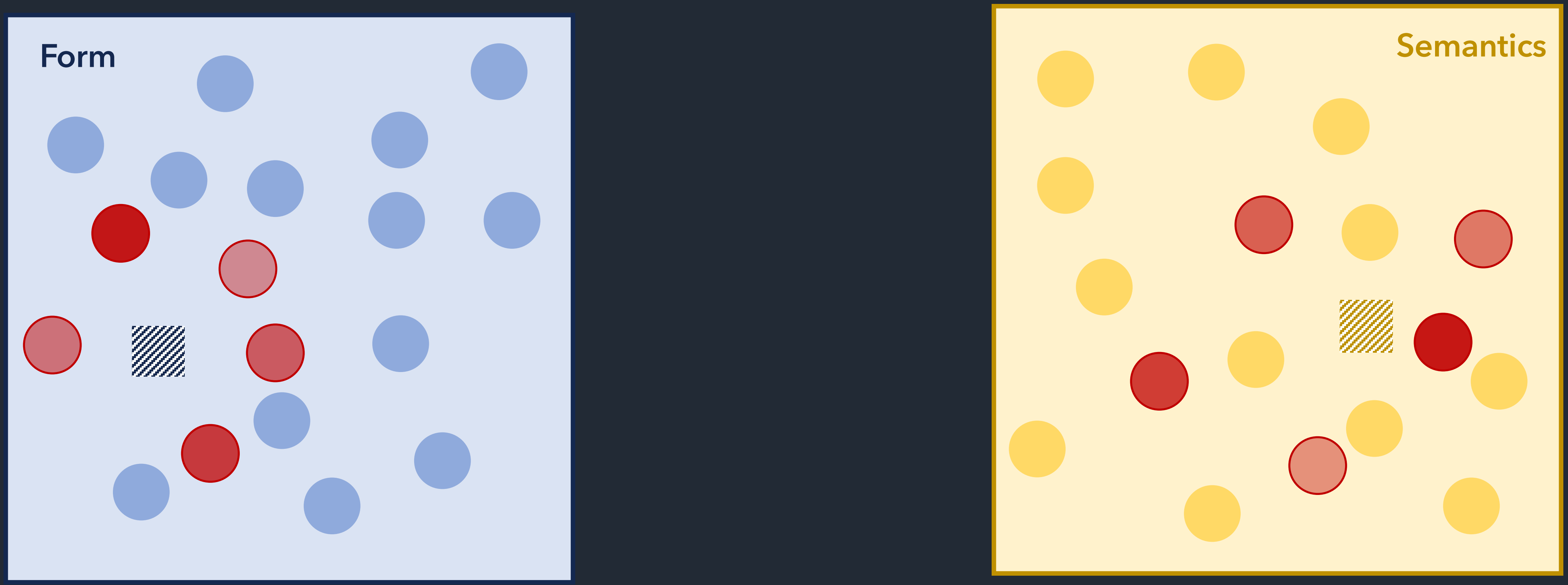
We can probe how acceptable these novel derived words or compounds would be to language users by checking relations with other words.

But we need annotations, lots of them, to train functions. Can we do without? Also, what if the new word doesn't use known morphemes?

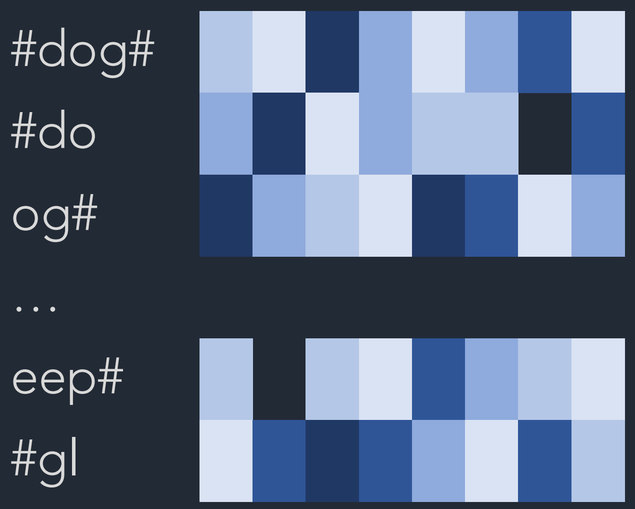
Linear Discriminative Learning

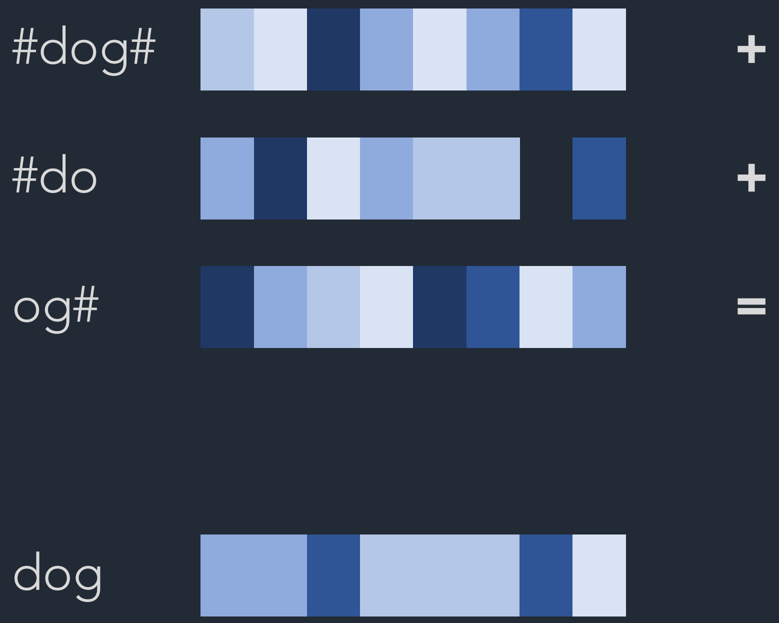


Form-To-Semantics Consistency



FastText





A multi-purpose toolkit

These methods do without explicit morphemes, although they often retrieve them: as long as a string has a form, it can be represented semantically exploiting correspondences in the language, at the interface between word forms and their meanings.

Bringing NLP closer to Dante’s creativity

We can use FRACSS to look at whether semantics has anything to do with why *inurbarsi* made it until today but *intuarsi* didn’t.

We can use CAOSS, FastText, LDL, and OSC to see what kind of connotations do the devils’ names evoke and better understand expressive language use.

What do we need though?

Quite some primary linguistic data to learn the representations of frequent words, as we still need them. So, to study Dante’s neologisms we’d need representative corpora of the language his contemporaries experienced, which is hard to come by in sufficient quantities.

Why this talk then?

The same linguistic creativity we find in the *Commedia* is always around, some of the innovations that appear today might make up the language 700 years from now.

Having tools to study these processes on a large scale while they happen can help us understand a lot about how people reinvent their languages.

What about made-up language?

No NLP tools can get you an answer to what Dante meant to say with that infamous verse, not even his contemporaries knew.

We can still project that verse to some semantic space and see where it lands. Probably an empty portion, far from everything known. Leaving the reader lost. Much like in hell.

THANKS!