

# NLP and Human Processing

prominence, referring expressions and metaphors

T. Mark Ellison (Helsinki 2022-10-06)



SFB 1252  
PROMINENCE  
IN LANGUAGE



Psycholinguistics

Models of communication  
motivating referring  
expression choice

February

Prominence

Understanding referring  
expression variation in ideal  
efficient codes

Computational Linguistics

Natural Language  
Processing

NL generation: evaluating  
referring expression form  
generation systems

Building *gold standard*  
corpora for evaluation

Corpus Linguistics

# Recap: Prominence

# Prominence in Language

## Prominence Beyond Prosody – A First Approximation

Himmelman, N., & Primus, B. (2015).  
**Prominence Beyond Prosody - a First  
Approximation.** In *pS-prominenceS: Prominences  
in Linguistics. Proceedings of the International  
Conference, University of Tuscia* (pp. 38-57).

Nikolaus P. Himmelman & Beatrice Primus

This paper sketches a framework for linguistic prominence applicable to prosodic and non-prosodic phenomena. It builds on the observation that there are important similarities between linguistic prominence and the essentially perceptual category of ‘being in the current centre of attention’. Three properties seem to set linguistic prominence apart from other linguistic asymmetries. Firstly, linguistic units of equal rank (e.g. syllables, co-arguments of a predicate) compete for the status of being in the centre, and secondly, this status may shift depending on the context. Thirdly, prominent units function as structural attractors in their domain.

# Prominence in Language

**Stands out** among things of the same type

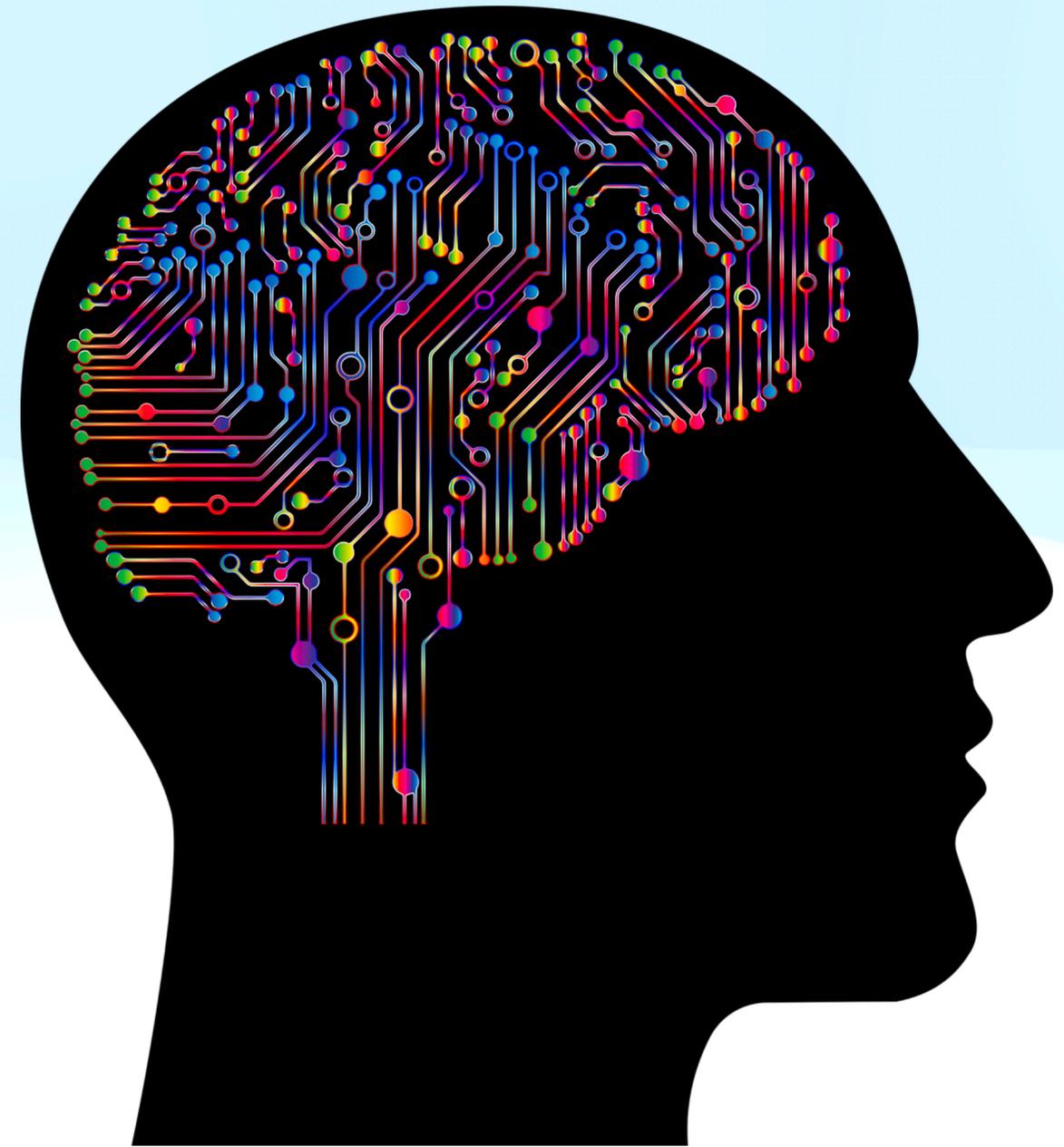
Prominence varies with item+context

Prominent items are structural attractors

# Discourse Prominence

## probabilities in the mind

- how **likely** are we to think/say/do particular things?
  - more prominent = more likely
- related terms: *salience*, *activation*
- inputs affecting discourse prominence do *priming*

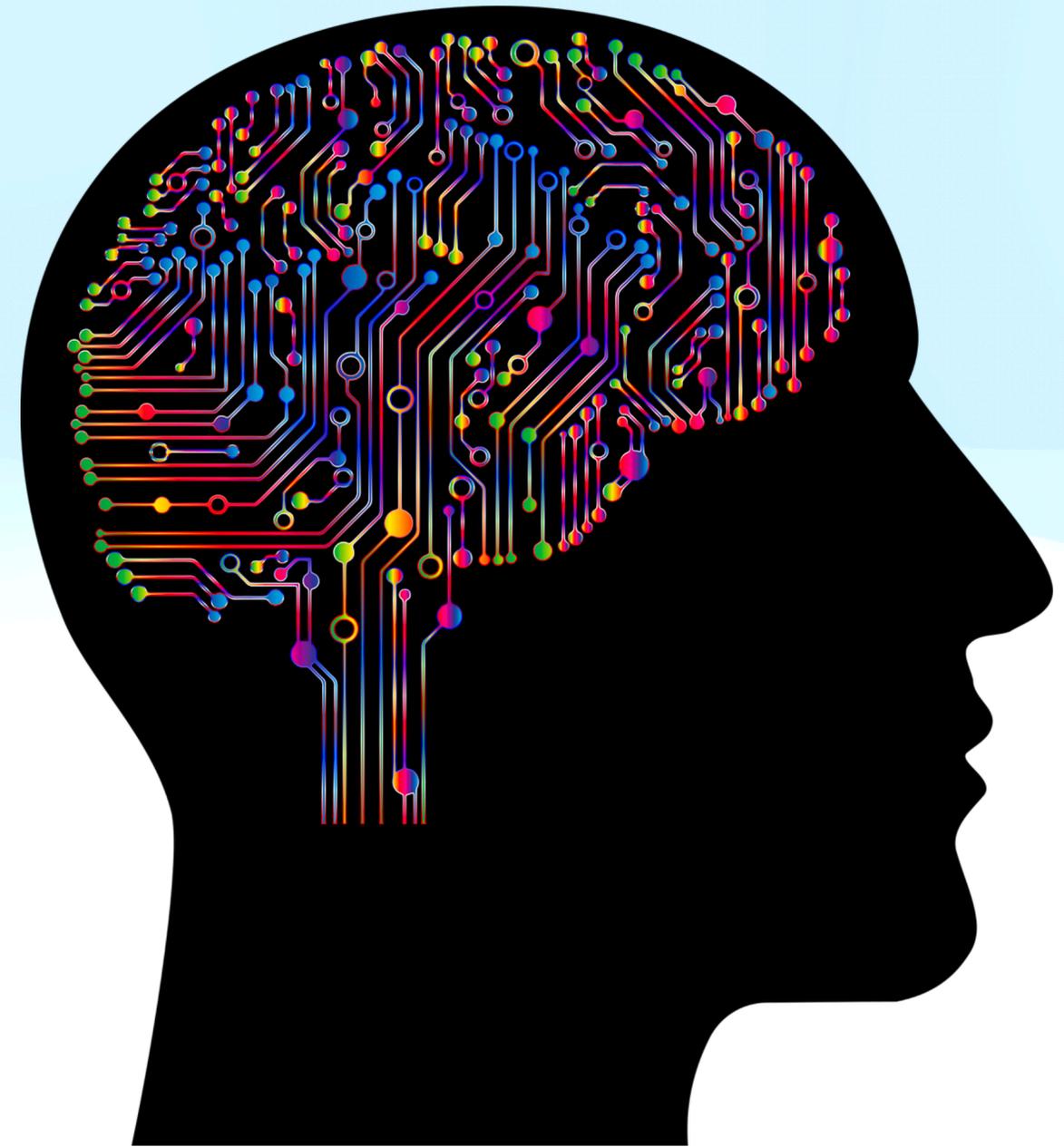


# Discourse Prominence

## probabilities in the mind

- how **likely** are we to think/say/do particular things?
  - more prominent = more likely
- related terms: *salience*, *activation*
- inputs affecting discourse prominence do *priming*

bread and ...



# Code Prominence

## priming - by highlighting

- phonetic choice: **hyperarticulation**
- morphological choice: *you're **the best***
- lexical choice: ***a feline** vs a cat*
- syntactic choice: ***yellow** is my favourite colour*



# Referring Expressions

# Code Prominence

## choice of referring expressions

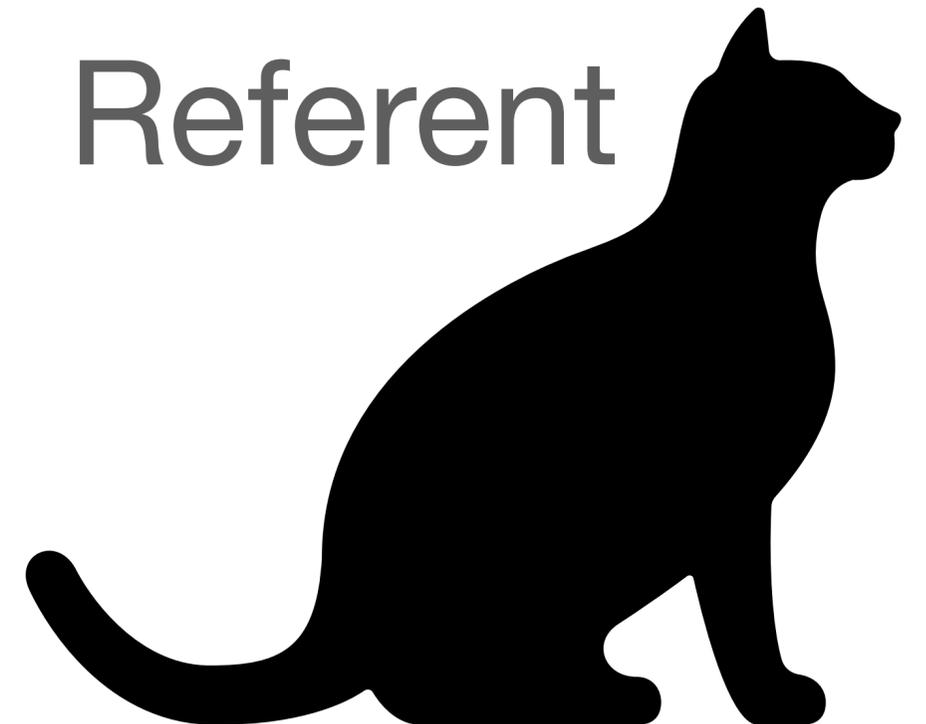
- referring expressions (RE) point out (refer to) something in the world
- what it refers to is a referent

Referring  
expression

**the cat**



Referent



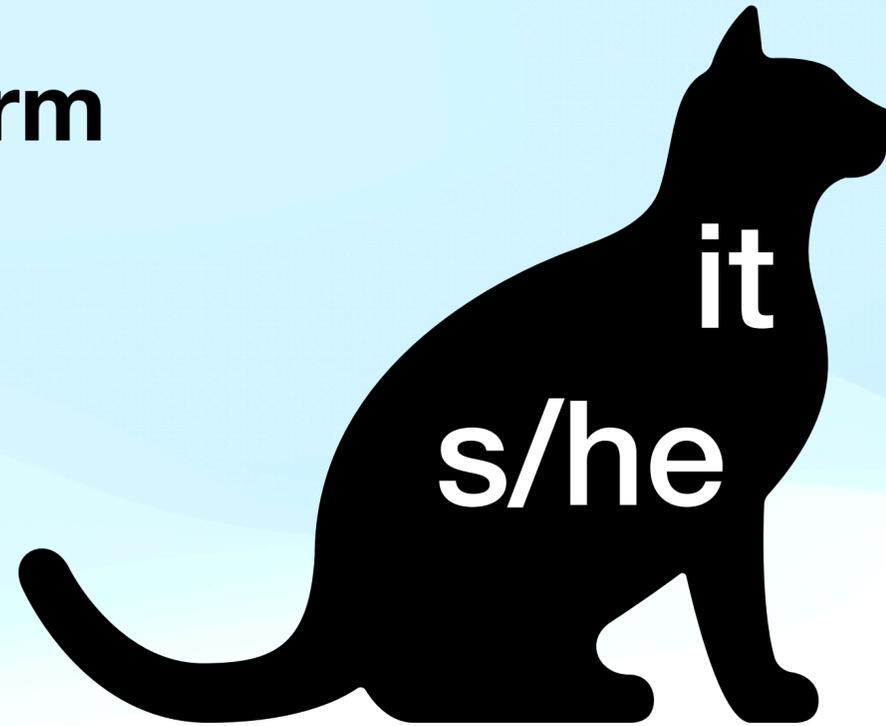
# Code Prominence

choice of referring expression form

Names

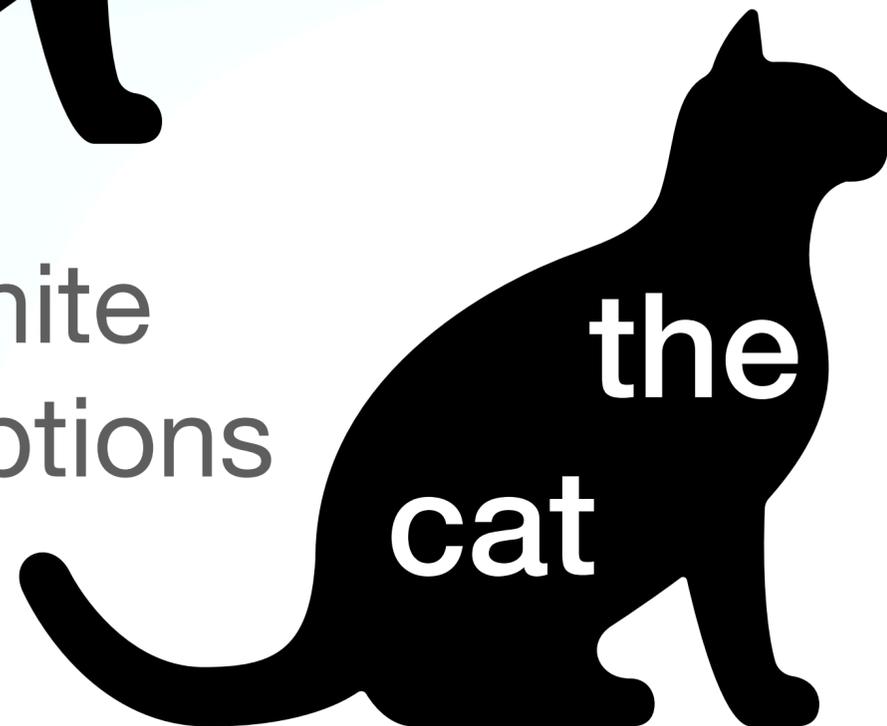


it  
s/he

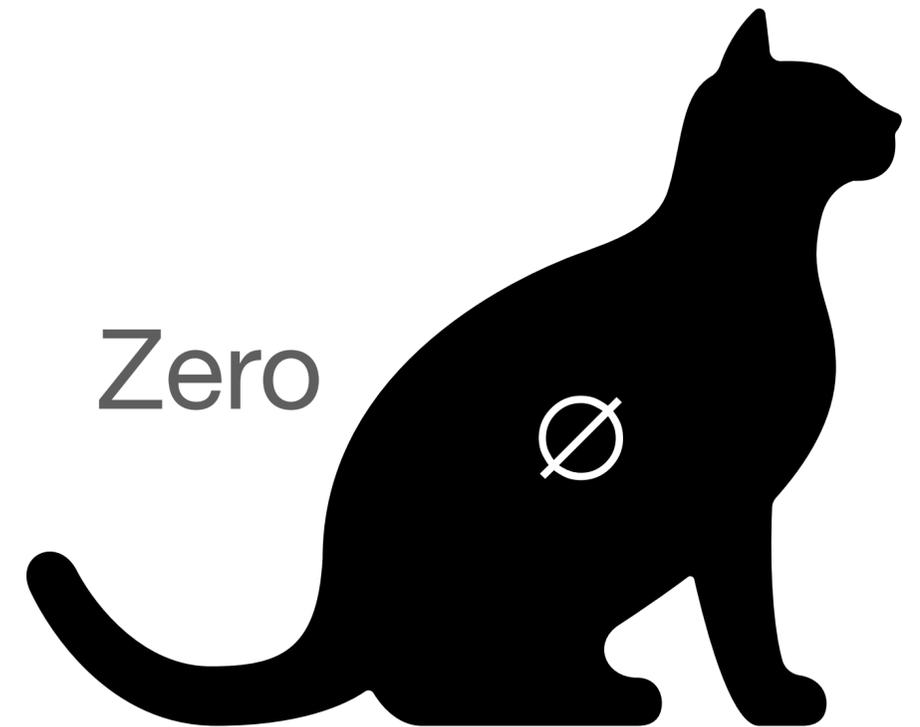


Pronouns

Definite  
Descriptions



Zero



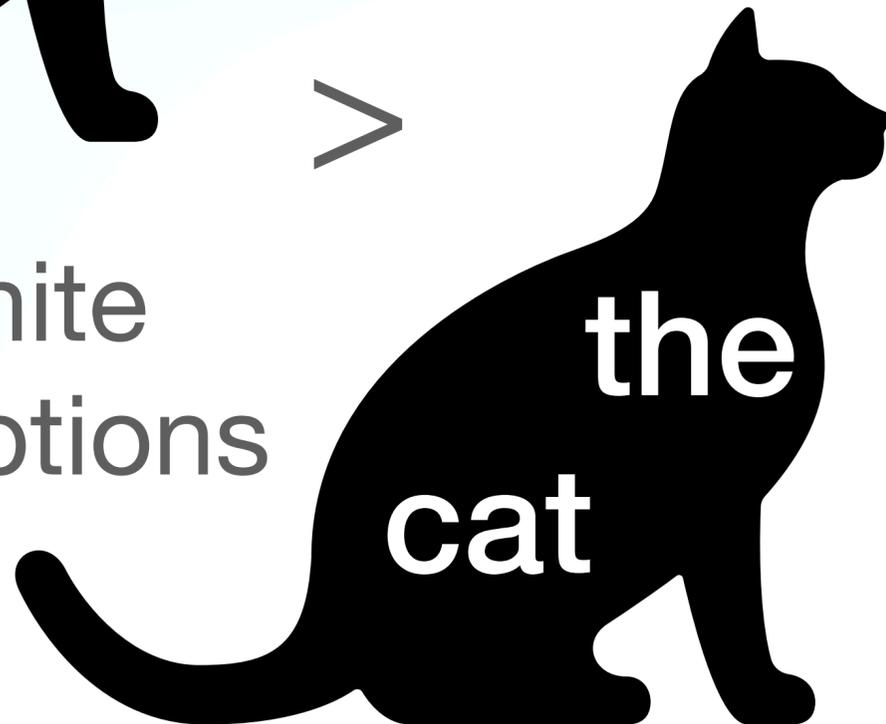
# Code Prominence

choice of referring expression form

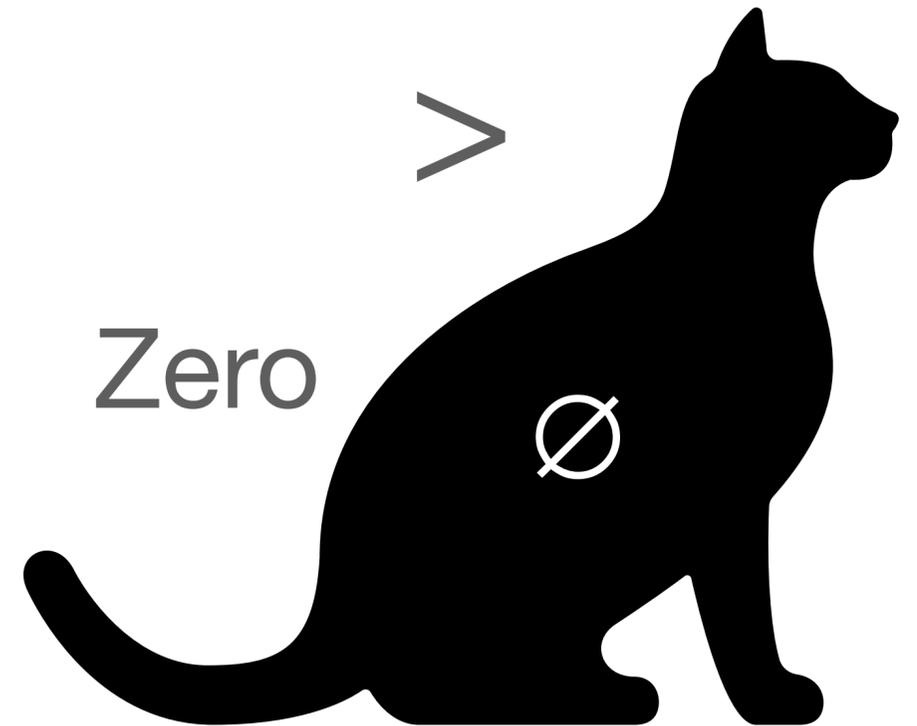
Names



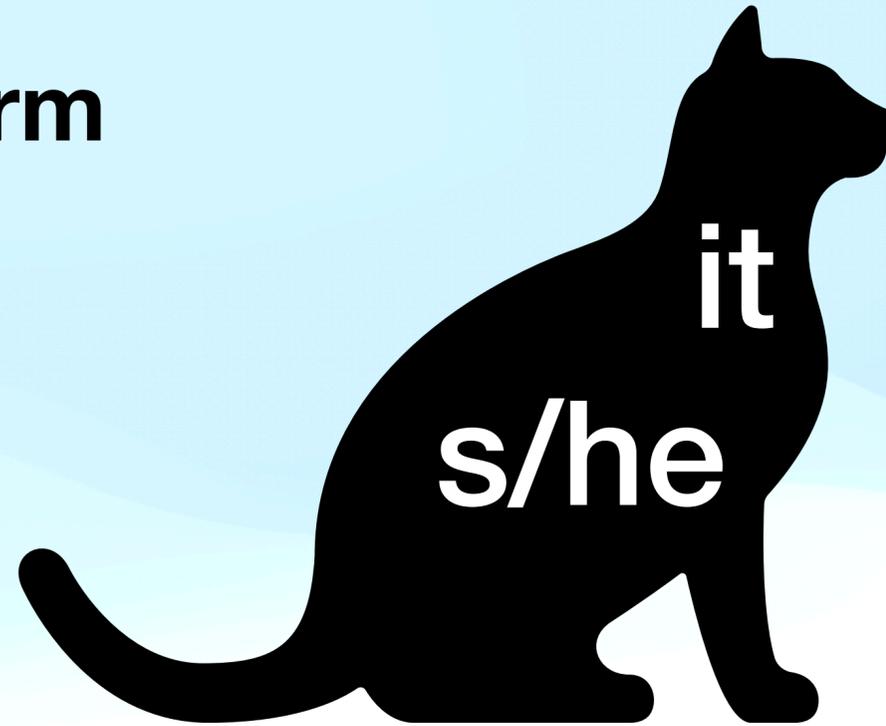
Definite  
Descriptions



Zero



s/he  
it



Pronouns

# Prominence

a Bayesian model of interpretation

$$\underbrace{P(s_{t+1} | f, s_t)}_{\text{Interpretation Confidence}} = \frac{\underbrace{P(f | s_{t+1}, s_t)}_{\text{Code Prominence}}}{\underbrace{P(f)}_{\text{Effort}}} \underbrace{P(s_{t+1} | s_t)}_{\text{Discourse Prominence}}$$

$\geq \theta$  Confidence  
achieves threshold

# Prominence

in speaker choice of form

1 Maintain interpretation accuracy:  
choose forms that are prominent enough

2 Minimise effort

# Referring Expressions

vary with context

**Spain**, officially the Kingdom of Spain, is a country located in Southern Europe, with two small in North Africa (both bordering Morocco). **The country** is a democracy. **It** is organized as a parliamentary monarchy. **It** is a developed country with the ninth-largest economy in the world. **It** is the largest of the three sovereign nations that make up the Iberian Peninsula—the others are Portugal and the microstate of Andorra.

En Europa, ocupa la mayor parte de la península ibérica.



# REs in NLP

# NLG

## (Natural Language Generation)

### Generating Spatio-Temporal Descriptions in Pollen Forecasts

Ross Turner, Somayajulu Sripada and Ehud Reiter

Dept of Computing Science,

University of Aberdeen, UK

{rturner, ssripada, ereiter}@csd.abdn.ac.uk

Ian P Davy

Aerospace and Marine International,

Banchory, Aberdeenshire, UK

idavy@weather3000.com

Turner, R., Sripada, S., Reiter, E., & Davy, I. P. (2006).  
**Generating spatio-temporal descriptions in pollen forecasts.** In *Demonstrations* (pp. 163-166). Chicago

#### Abstract

We describe our initial investigations into generating textual summaries of spatio-temporal data with the help of a prototype Natural Language Generation (NLG) system that produces pollen forecasts for Scotland.

#### 1 Introduction

New monitoring devices such as remote sensing systems are generating vast amounts of spatio-temporal data. These devices, coupled with the wider accessibility of the data, have spurred large amounts of re-

forecasts were written. An example of a pollen forecast text is shown in Figure 1, its corresponding data is shown in table 1. A pollen forecast in the map form is shown in Figure 2.

*'Monday looks set to bring another day of relatively high pollen counts, with values up to a very high eight in the Central Belt. Further North, levels will be a little better at a moderate to high five to six. However, even at these lower levels it will probably be uncomfortable for Hay fever sufferers.'*

# NLG

## (Natural Language Generation)

### Human

‘Monday looks set to bring another day of relatively high pollen counts, with values up to a very high eight in the Central Belt. Further North, levels will be a little better at a moderate to high five to six. However, even at these lower levels it will probably be uncomfortable for Hay fever sufferers.’

ValidDate	AreaID	Value
27/06/2005	1 (North)	6
27/06/2005	2 (North West)	5
27/06/2005	3 (Central)	5
27/06/2005	4 (North East)	6
27/06/2005	5 (South West)	8
27/06/2005	6 (South East)	8

### Machine

Grass pollen levels for Monday remain at the moderate to high levels of recent days with values of around 5 to 6 across most parts of the country. However, in southern areas, pollen levels will be very high with values of 8.

# NLG

## Referring Expression Generation (REG)

Grass pollen levels for Monday remain at the moderate to high levels of recent days with values of around 5 to 6 across most parts of the country. However, in southern areas, pollen levels will be very high with values of 8.

- name: “in Lothian and the Borders”
- definite description: “in southern areas”
- pronoun: “there” - if the context permitted it
- zero: leave it out entirely, if understood from the context

# NLG

## REF technologies

- **rule-based approaches**
- deep learning approaches
- feature-based approaches

# NLG

## REF technologies

- rule-based approaches
- **deep learning approaches**
- feature-based approaches

# NLG

## REF technologies

- rule-based approaches
- deep learning approaches
- **feature-based approaches**

Characterise each reference  
by descriptive features:

# NLG

## REF technologies

**Referring Expression Generation in Context:  
Combining Linguistic and Computational Approaches**

Inaugural-Dissertation zur Erlangung des Doktorgrades der  
Philosophischen Fakultät der Universität zu Köln

im Fach Linguistik

vorgelegt von

Fahime Same

# NLG

## REF technologies

Feature	Type[N]	DT	Symbol
Grammatical role of REF	cat[1]	1-7	gm
Grammatical role of ANTE	cat[1]	5,6	gm_p1
Grammatical role of the 2 <sup>nd</sup> and 3 <sup>rd</sup> ANTE	cat[2]	6	gm_p2, gm_p3
Trigram grammatical roles of the three antecedents	cat[1]	7	gm_tri
Is REF the subject of this & the two previous sentences?	bool[3]	6	subj_S, subj_prevS, subj_prev2S
Is ANTE in the subject position?	bool[1]	6	ante_subj
Are REF and ANTE prepositional phrases?	bool[2]	5	ref_pp, ante_pp

Table 5.3: Grammatical features encoded in different feature sets.

# NLG

## REF technologies

Feature	Type	DT	Symbol
Animacy/semantic category	cat[1]	3,4,5,7	anim
Gender	cat[1]	5	gender
Plurality	cat[1]	5	plur

Table 5.4: Inherent features encoded in different feature sets.

# NLG

## REF technologies

Feature	Type[N]	DT	Symbol
Sentence Number	num[1]	6,7	sent_num
NP number	num[1]	7	np_num
Mention number	num[1]	1,5,6	ment_num
Referent number	num[1]	6	ref_num
How many times has REF occurred since the beginning? (1,2,3,4+)	cat[1]	4	count_bef
How many times does REF occur since the last change? (1,2,3,4+)	cat[1]	4	count_aft
Mention order (first, second, middle, last)	cat[1]	3	ment_ord
Does REF appear in the first sentence?	bool[1]	7	first_sent
Does REF appear at the beginning of a paragraph?	bool[1]	4	firstS_par

Table 5.5: Positional features of different feature sets

# NLG

## REF technologies

Feature	Type[N]	DT	Symbol
Distance in number of words	num[1]	1,5	dist_w
Distance in number of NPs	num[1]	7	dist_np
Distance in number of markables	num[1]	5	dist_mark
Distance in number of sentences	num[1]	5,7	dist_sent
Distance in number of paragraphs	num[1]	5	dist_par
Distance to the nearest non-pronominal antecedent	num[1]	5	dist_full
Word distance (5 bins of 0-10, 11-20, 21-30, 31-40 and 40+)	cat[1]	2	bin5_w
Word distance (3 bins of 0-5, 6-12 and 13+)	cat[1]	3	bin3_w
Sentence distance (+/-2 sentences)	cat[1]	6	bin2_sent
Sentence distance (3 bins of 0, 1, 2+ sentences)	cat[1]	3	bin3_sent

Table 5.6: Recency features of different feature sets

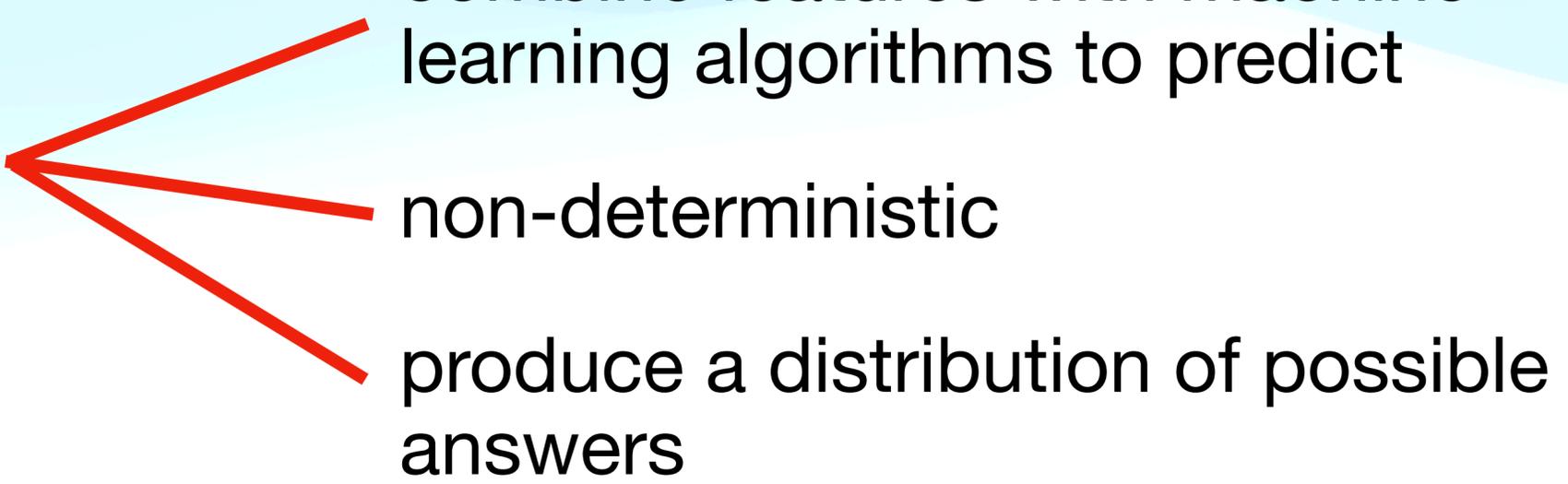
# NLG

## REF technologies

General classes	ISG	FERREIRA	OSU	ICSI	KIBRIK	UDEL	CNTS
Grammatical role	1	1	1	1	4	8	2
Inherent	0	0	1	1	3	0	1
Referential status / Givenness	0	3	0	0	0	0	0
Distance / Recency	2	1	2	0	4	1	2
Competition	0	0	3	1	0	2	1
Antecedent form	1	0	0	0	1	0	0
Pattern	0	0	0	8	0	4	12
Position	1	0	1	3	1	3	3
Protagonism	0	0	0	0	2	0	0
Total number of features	5	5	8	14	15	18	21

# NLG

## REF technologies

- rule-based approaches
  - deep learning approaches
  - **feature-based approaches**
    - combine features with machine learning algorithms to predict
    - non-deterministic
    - produce a distribution of possible answers
- 



# NLG Evaluation

- problem: corpora are deterministic
- suppose the right answer is (0, 1, 0, 0)
- if the prediction is (51%, 39%, 6%, 4%) is not so wrong as (80%, 5%, 9%, 6%)

what variation is present in human production anyway?

**THE WALL STREET JOURNAL.**  
THURSDAY, FEBRUARY 24, 2022 - VOL. CCLXXIX NO. 44  
DOW JONES | *Market Gaps* | \*\*\*\*\* | WSJ.com | \*\*\*\*\* \$5.00  
DJIA 33131.76 ▲ 464.85 1.4% | NASDAQ 13027.49 ▼ 2.6% | STOXX 600 453.86 ▲ 0.3% | 10-YR. TREAS. 9/32, yield 1.976% | OIL \$92.10 ▲ \$3.10 | GOLD \$1,909.20 ▲ \$3.10 | EURO \$1.1306 | YEN 115.02

**What's News**  
*Business & Finance*

**Investors rushed for safety, pushing down stocks and lifting the prices of oil, gold and government bonds, after Russia's Putin launched a military operation in Ukraine. B1**

**U.S. life insurers, as expected, made a large number of Covid-19 death-benefit payouts last year, with many seeing a jump in other death claims, too. A1**

**Executives said it has frozen former CEO Jes Staley's deferred pay while regulators complete a probe into how he characterized his relationship with Jeffrey Epstein. B1**

**A judge said she would pause the trial of Roger Ng after prosecutors said they had failed to turn over a tranche of documents to the former Goldman banker's lawyers. B1**

**Ford's CEO said the auto maker doesn't intend to spin off its electric-vehicle business, tamping down speculation that the company could break off its EV operations to boost market value. B3**

**The EU is proposing legislation that would force more data sharing among companies in Europe, aiming to loosen the grip officials say a few big tech firms have on some commercial and industrial data. B4**

**Lower's surprised Wall Street with its management of costs and pricing in the latest quarter amid a slowing outlook for the home-improvement sector's sales. B3**

**Kevin Lander suspended a top executive without pay following a backlash over a post on his personal Instagram account. B3**

**World-Wide**

**Russian missiles and air strikes hit Ukraine's capital Kyiv and more than a dozen other cities across the country Thursday, minutes after Putin announced a military operation that he said seeks to "demilitarize and denazify Ukraine" and bring its leaders to trial. A1, A7-9**

**The U.S. and its allies are poised to unveil further sanctions, now that Russia has launched what Biden called "an unprovoked and unjustified attack" on Ukraine, hoping a fresh tranche of penalties will have a greater deterrent effect than the first set. A1**

**Two prosecutors leading the Manhattan district attorney's investigation into Trump and his business resigned, casting doubt on the future of the yearslong criminal probe. A3**

**The Justice Department is ending a Trump-era initiative to counter national-security threats from China after it led to a series of failed prosecutions of academics. A4**

**Canada's Trudeau, in a surprise turnaround, said his government no longer required emergency powers to deal with protests against Covid-19 restrictions. A10**

**A more infectious type of the Omicron variant has surged to account for more than a third of global Covid-19 cases sequenced recently, adding to the debate about whether countries are ready for full reopening. A6**

**The number of women in the U.S. who died while pregnant or shortly after pregnancy continued to rise in 2020 as the pandemic spread, according to a federal report. A6**

**Contents**  
Opinion..... A5-7  
Arts in Review..... A3  
Special Journal A6-8  
Business News..... E1  
Sports..... A4  
Commentary..... A4  
Technology..... B4  
Features..... E1  
U.S. News..... A1-4  
World on Street..... B2  
Weather..... A4  
Markets..... B1  
World News..... A1-10

**RUSSIA STRIKES UKRAINE**

**Air attacks, missiles hit Kyiv, other cities; Biden pledges further steps to punish Putin**

Russian missiles and air strikes hit Ukraine's capital Kyiv and more than a dozen other cities across the country Thursday, minutes after President Vladimir Putin announced a military operation that he said seeks to "demilitarize and denazify Ukraine" and bring its leaders to trial.

*By Yaroslav Trofimov, Alan Cullinan and Brett Forrest in Kyiv and Ann M. Stammers in Moscow*

"From all of you, we need calm," Ukrainian President Volodymyr Zelenskyy said in an early-morning television address. "We are working, our army is working. Don't panic, we are strong, we are ready for anything, we will overcome." He said he has ordered martial law and has spoken with President Biden about the attack.

Ukrainian officials said that the initial wave of strikes targeted military installations, airfields and government facilities across the country, as well as border force installations. In Kharkiv, eastern

Ukraine's largest city, residents said a large fire was visible in the morning darkness, after what appeared to be a hit at a weapons depot. Heavy shelling targeted the city of Mariupol on the Azov Sea. Air raid sirens sounded in Kyiv after 7 a.m.

While the Ukrainian military didn't release casualty figures, a senior Ukrainian official said he believed that hundreds of Ukrainian soldiers died in Russian airstrikes and missile attacks.

Russia denied conducting missile, air or artillery strikes on Ukrainian cities or threatening civilian populations, the country's Ministry of Defense told the Russian state news agency RIA Novosti.

President Biden called Mr. Putin's move an unprovoked, unjustified attack in Ukraine, pledging further action against Russia.

"President Putin has chosen a premeditated war that will bring a catastrophic loss of life and human suffering," he said in a statement.

Later Mr. Biden said he would be meeting Thursday with leaders of the Group of Seven.

**Germany takes step to end Russian gas reliance..... A7**  
**Biden faces a new foreign-policy test..... A9**

**Crisis Sets New Struggle For Global Supremacy**

By MICHAEL R. GORDON

Russia's audacious military assault on Ukraine is the first major clash marking a new order in international politics, with three major powers jostling for position in ways that threaten America's primacy.

The challenges are different than those the U.S. and its network of alliances faced in the Cold War. Russia and China have built a thriving partnership based in part on a shared interest in diminishing U.S. power. Unlike the Sino-Soviet bloc of the 1950s, Russia is a critical gas supplier to Europe, while China isn't an impoverished, war-ravaged partner but the world's manufacturing powerhouse with an expanding military.

In deploying a huge force and on Thursday ordering what he called a "special military operation," Russian President Vladimir Putin is demanding that the West rewrite the post-Cold War security arrangements for Europe and demonstrated that Russia has the military capability to impose its will despite Western objections and sanctions.

To do this, Mr. Putin shifted military units from Russia's border with China, showing confidence in his relations with Beijing. The two powers, in effect, are coordinating to reshape the global order to their advantage, though their ties short stop of a formal alliance.

This emerging order leaves the U.S. contending with two adversaries at once in geographically disparate parts of the world where America has long been the dominant power.

**Greg Ip: Ukraine clash puts dagger in globalization..... A2**

**Big Companies Rush to Plan For Repercussions of Attack**

By ALYSON MACDONALD AND NICK KOROV

Western companies with operations in Russia and Ukraine are preparing for the potential impact of sanctions on their businesses there and readying contingency plans in the event of further military action, after President Vladimir Putin of Russia sent troops into two breakaway regions of Ukraine.

Such moves, though, could also complicate operations for multinationals that have operations in Russia and that often join with Russian companies and businesses. Big oil companies, including BP PLC, Exxon Mobil Corp. and Shell PLC, have substantial investments in Russia, as do brewing giant Carlsberg A/S and auto maker Renault SA.

On Wednesday, the chief executives of several major U.S. companies said they were planning to meet with Russian counterparts to discuss the impact of the attack on their businesses.

**Russia's move against Ukraine roils markets..... B1**

**West Is Lining Up Further Sanctions**

The U.S. and its allies are poised to unveil further sanctions now that Russia has launched what President Biden called "an unprovoked and unjustified attack" on Ukraine, hoping a fresh tranche of penalties will have a greater deterrent effect than the first set.

*By Ian Talley in Washington and Laurence Norman in Berlin*

On Tuesday, after Russian President Vladimir Putin sent troops into two breakaway regions of Ukraine, Western nations imposed sanctions on Russian sovereign debt, six Russian banks, several wealthy Russians linked to Mr. Putin's inner circle, Defense Minister Sergei Shoigu and other high officials, and halted the Nord Stream 2 natural-gas pipeline.

A senior U.S. administration official described the measures as "only the sharp edge of the pain we can inflict."

U.S., European Union and British officials say they had other, more powerful financial weapons in their arsenal and were primed to use them as Mr. Putin escalated. Those include sanctions on much larger Russian banks, a ban on investment in Russian gas projects, and export controls designed for long-term economic growth.

Late Wednesday, Mr. Biden said he would in the morning announce "further consequences" on Russia.

The previous day, a senior administration official said potential plans included hitting the country's most critical banks.

No Russian financial institution is safe if this invasion proceeds," the official said.

U.S. officials say that taken as a whole, the sanctions are intended to shock Russia's finances as "only the sharp edge

# NLG

## Solution

- evaluate NLG methods against **gold standard** corpora of human variation
- mathematically principled measures of the distance between distributions
  - e.g. **Jensen–Shannon divergence** symmetrises the Kullback-Liebler-distance

$$JS(P, Q) = \frac{1}{2} (KL(P | Q) + KL(Q | P))$$

intuitively: how much information do we gain (on average) about an item by learning it was a *prediction* or *from the corpus*?

# NLG

## A Gold-Standard Corpus of Human Variation

### Individual Variation in the Choice of Referential Form

**Thiago Castro Ferreira** and **Emiel Krahmer** and **Sander Wubben**

Tilburg center for Cognition and Communication (TiCC)

Tilburg University

The Netherlands

{tcastrof, e.j.krahmer, s.wubben}@uvt.nl

Ferreira, T. C., Krahmer, E., & Wubben, S. (2016, June). **Individual variation in the choice of referential form**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 423-427).

### Abstract

This study aims to measure the variation between writers in their choices of referential form by collecting and analysing a new and publicly available corpus of referring expressions. The corpus is composed of referring

uated against a corpus of human written texts, predicting what form each reference should have in a given context. Now consider a situation in which the algorithm predicts that a reference should be a description, while this same reference is a pronoun in the corpus text. Should this count as an error? The answer is: it depends. The use of a pronoun

# NLG

## A Gold-Standard Corpus of Human Variation

- corpus of human variation
- 36 English texts: news, commercial product reviews, encyclopedia texts (GREC)
- expressions referring to the topic were identified

and replaced with gaps

### Spain

\_\_\_\_, officially the Kingdom of Spain, is a country located in Southern Europe, with two small in North Africa (both bordering Morocco). \_\_\_\_ is a democracy. \_\_\_\_ is organized as a parliamentary monarchy. \_\_\_\_ is a developed country with the ninth-largest economy in the world. \_\_\_\_ is the largest of the three sovereign nations that make up the Iberian Peninsula—the others are Portugal and the microstate of Andorra.

# NLG

## A Gold-Standard Corpus of Human Variation

- 78 participants from CrowdFlower, most native speakers of English
- task: fill gaps with references to the topic
- participants got a short description about topics beforehand
- participants encouraged filled gaps to make the texts easy to understand

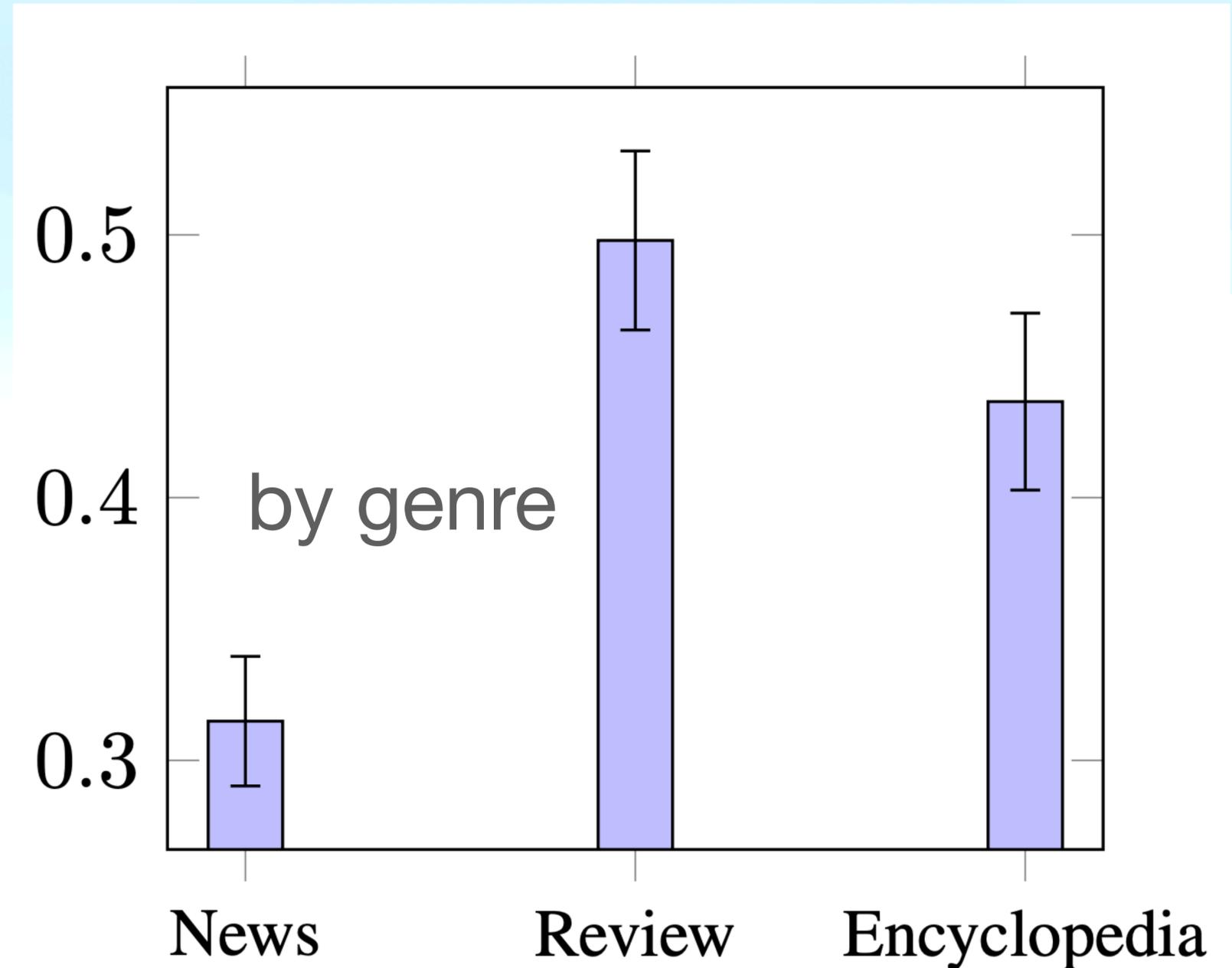
### Spain

**Spain**, officially the Kingdom of Spain, is a country located in Southern Europe, with two small in North Africa (both bordering Morocco). **The country** is a democracy. **It** is organized as a parliamentary monarchy. **It** is a developed country with the ninth-largest economy in the world. **It** is the largest of the three sovereign nations that make up the Iberian Peninsula—the others are Portugal and the microstate of Andorra.

# NLG

## A Gold-Standard Corpus of Human Variation

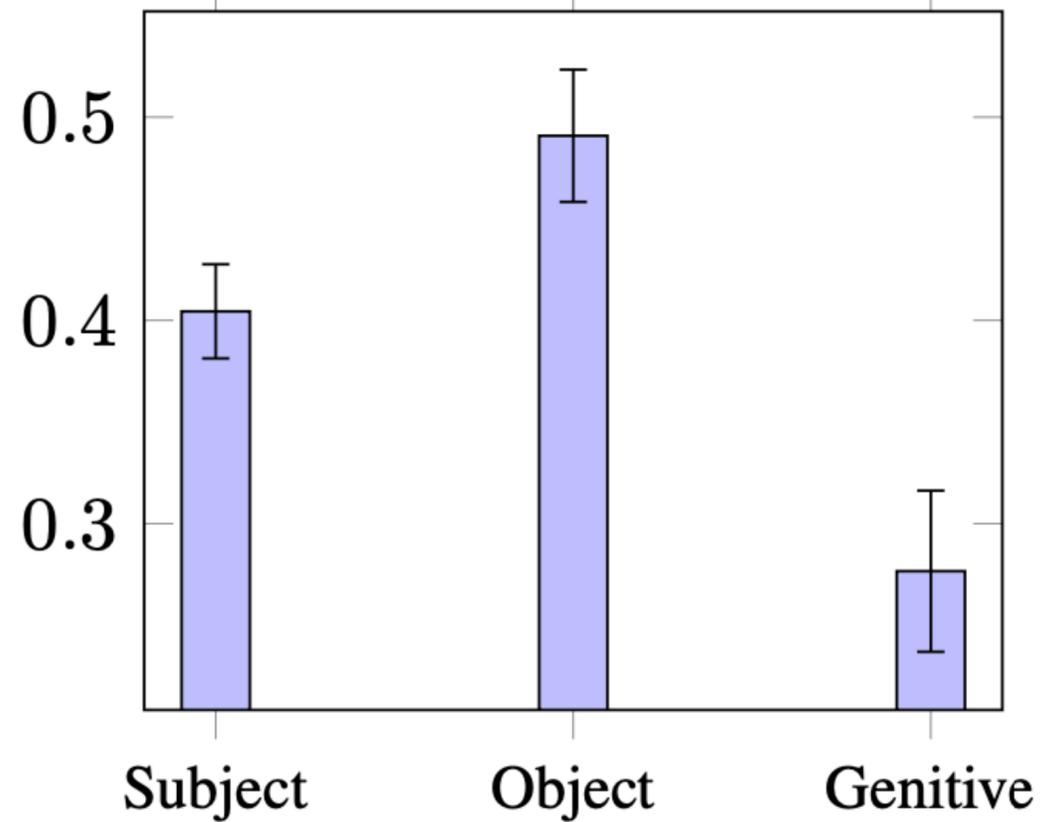
- scaled entropy  
entropy of choices / maximal possible entropy
- averaged over references



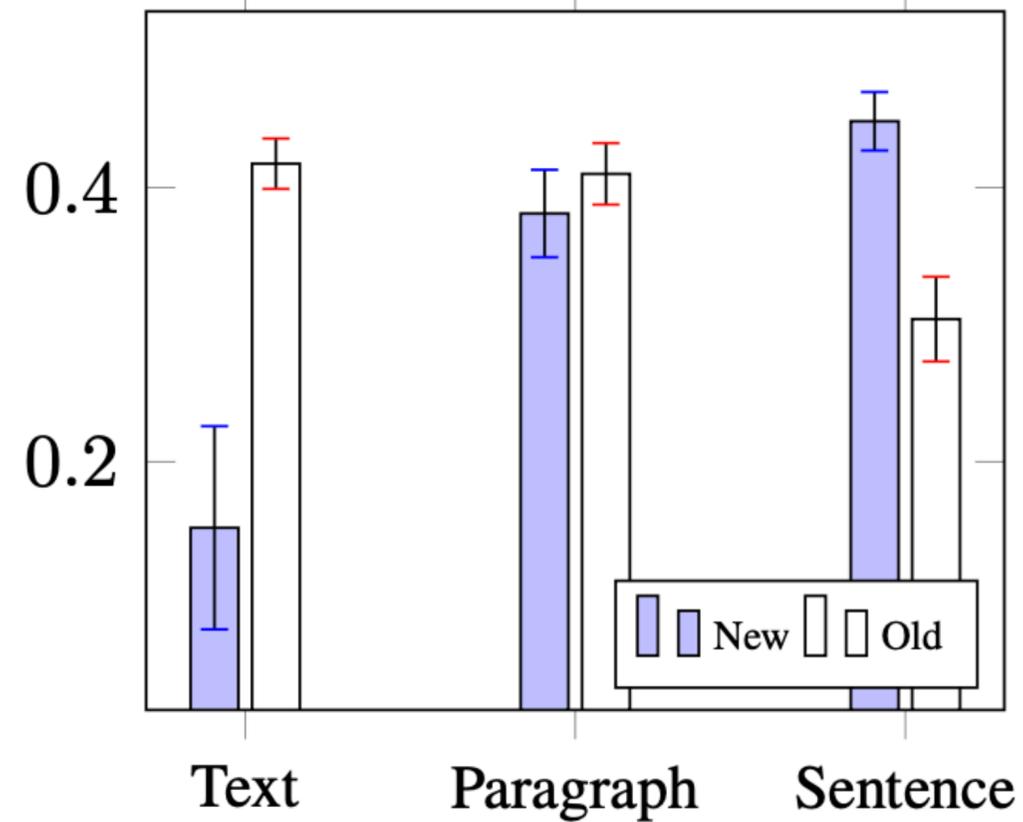
# NLG

## A Gold-Standard Corpus of Human Variation

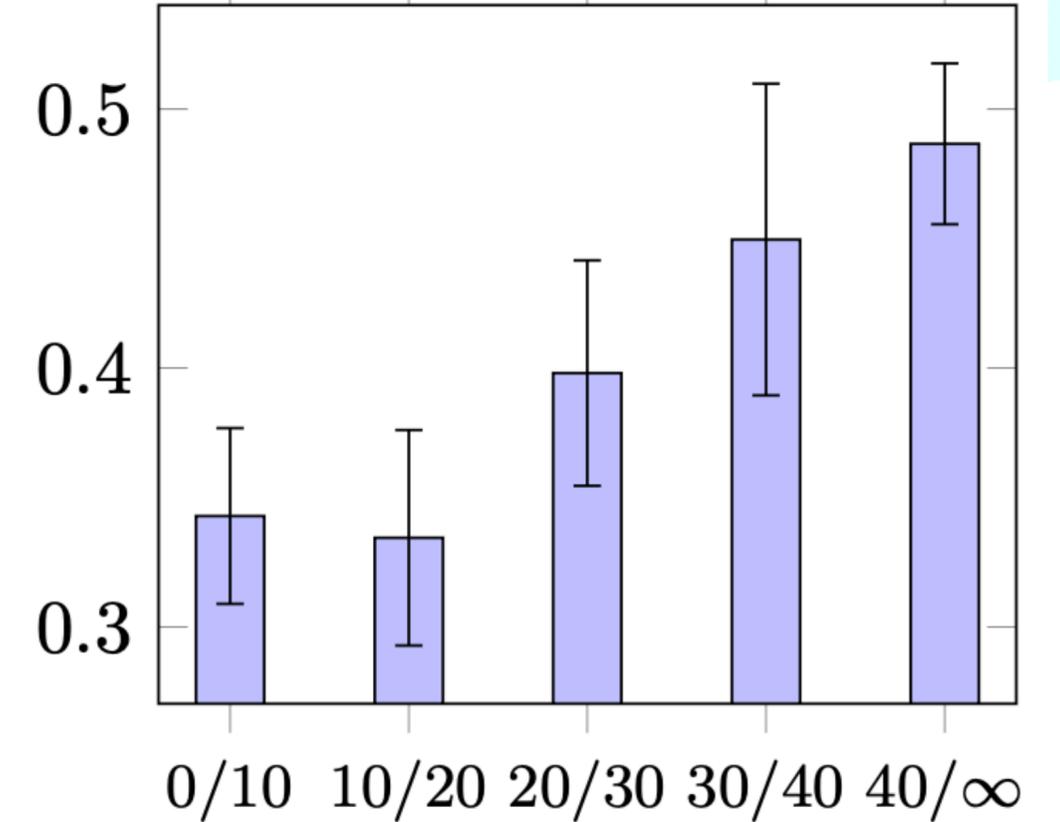
by syntactic role



by new in ...



by antecedent distance



# Beyond Parallel Corpora

# Beyond Parallel Corpora

## Ellison & Same

### Constructing Distributions of Variation in Referring Expression Type from Corpora for Model Evaluation

Ellison, T. M., & Same, F. **Constructing Distributions of Variation in Referring Expression Type from Corpora for Model Evaluation.** In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 2989–2997.

**T. Mark Ellison, Fahime Same**  
University of Cologne  
Cologne, Germany  
{t.m.ellison, f.same}@uni-koeln.de



#### Abstract

The generation of referring expressions (REs) is a non-deterministic task. However, the algorithms for the generation of REs are standardly evaluated against corpora of written texts which include only one RE per each reference. Our goal in this work is firstly to reproduce one of the few studies taking the distributional nature of the RE generation into account. We add to this work, by introducing a method for exploring variation in human RE choice on the basis of longitudinal corpora - substantial corpora with a single human judgement (in the process of composition) per RE. We focus on the prediction of RE types, *proper name*, *description* and *pronoun*. We compare evaluations made against distributions over these types with evaluations made against parallel human judgements. Our results show agreement in the evaluation of learning algorithms against distributions constructed from parallel human evaluations and from longitudinal data.

**Keywords:** Referring Expression, Prediction, Evaluation, Variation

#### 1. Introduction

task and not a deterministic one. Although a certain



# Beyond Parallel Corpora

## Problems with them

Parallel corpora construction processes are **unnatural**

so the distributions produced are likely to be very noisy

**Filling in slots in someone else's text**

VS

**composing your own text**

### Spain

\_\_\_\_, officially the Kingdom of Spain, is a country located in Southern Europe, with two small in North Africa (both bordering Morocco). \_\_\_\_ is a democracy. \_\_\_\_ is organized as a parliamentary monarchy. \_\_\_\_ is a developed country with the ninth-largest economy in the world. \_\_\_\_ is the largest of the three sovereign nations that make up the Iberian Peninsula—the others are Portugal and the microstate of Andorra.

# Beyond Parallel Corpora

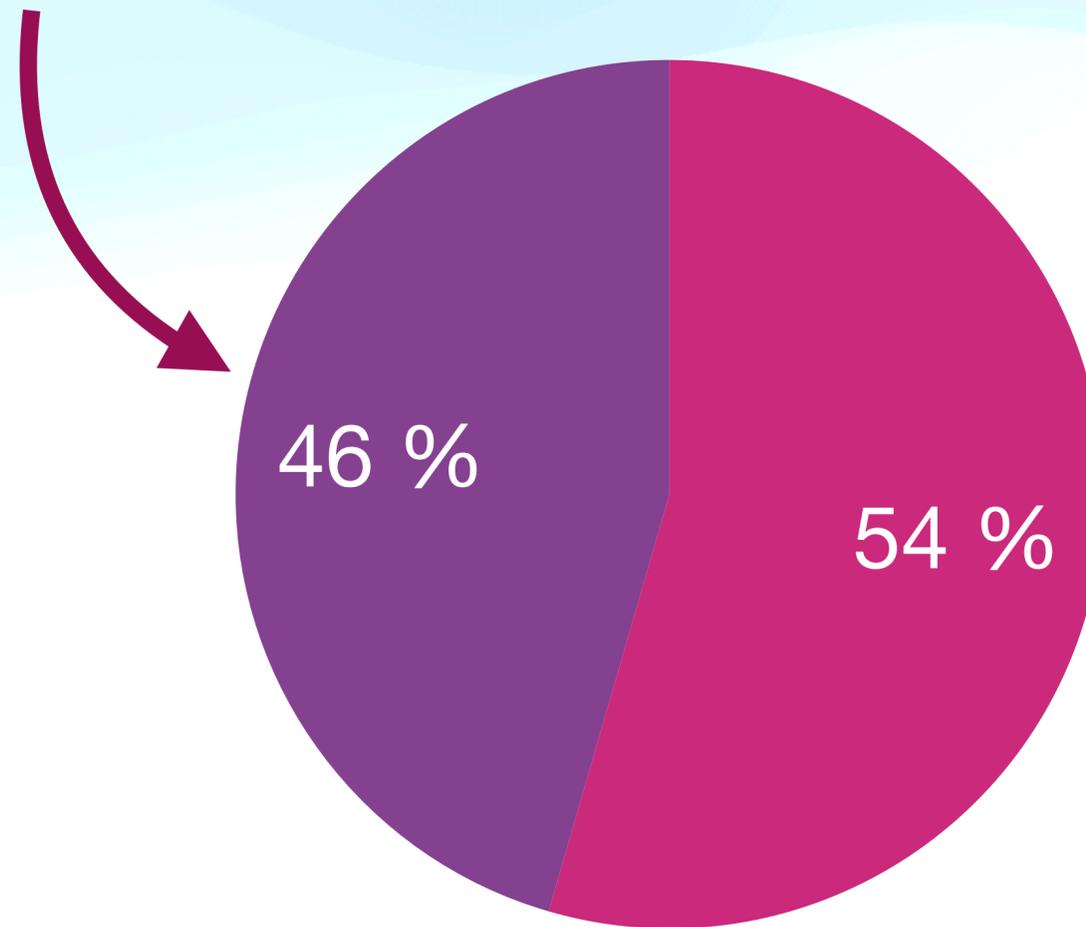
## Back to the Drawing Board

What is (free) variation?

It is the variability that you cannot predict

from conditioning factors in the context or pragmatics

Variability predictable from features of referent and context



Unpredictable variability = (free) variation

# Longitudinal Corpora of Variation

# Longitudinal Corpora of Variation

## Features

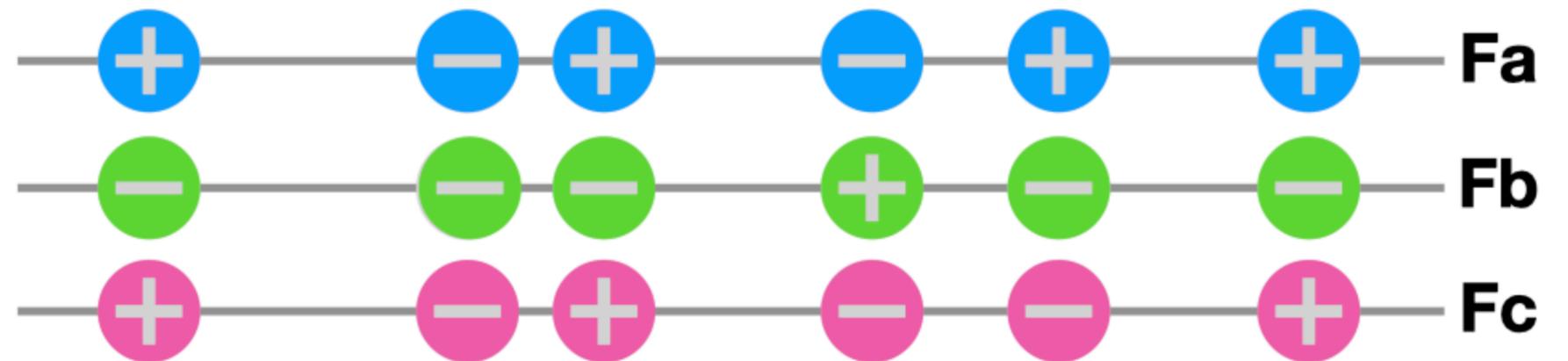
**Fa** - bucketed number of sentences since antecedent

**Fb** - the referent is a person

**Fc** - this is the topic of the current paragraph

**Fd** - the referent is of a high frequency type (e.g. person)

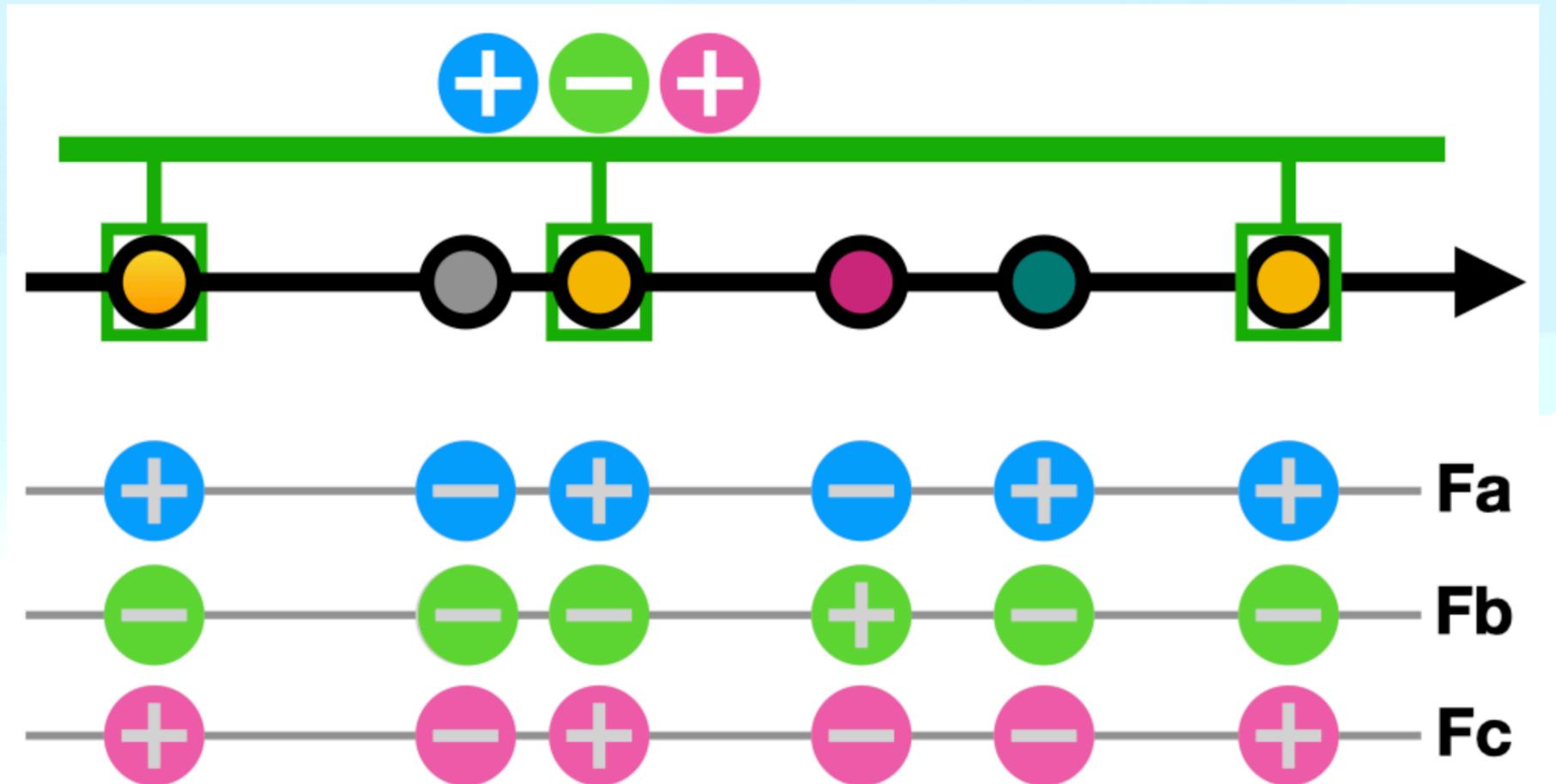
...



# Longitudinal Corpora of Variation

## Feature Combinations

1. choose features that condition RE form
2. find all REs for each feature value combination
3. build distribution RE forms within each set



# Longitudinal Corpora of Variation

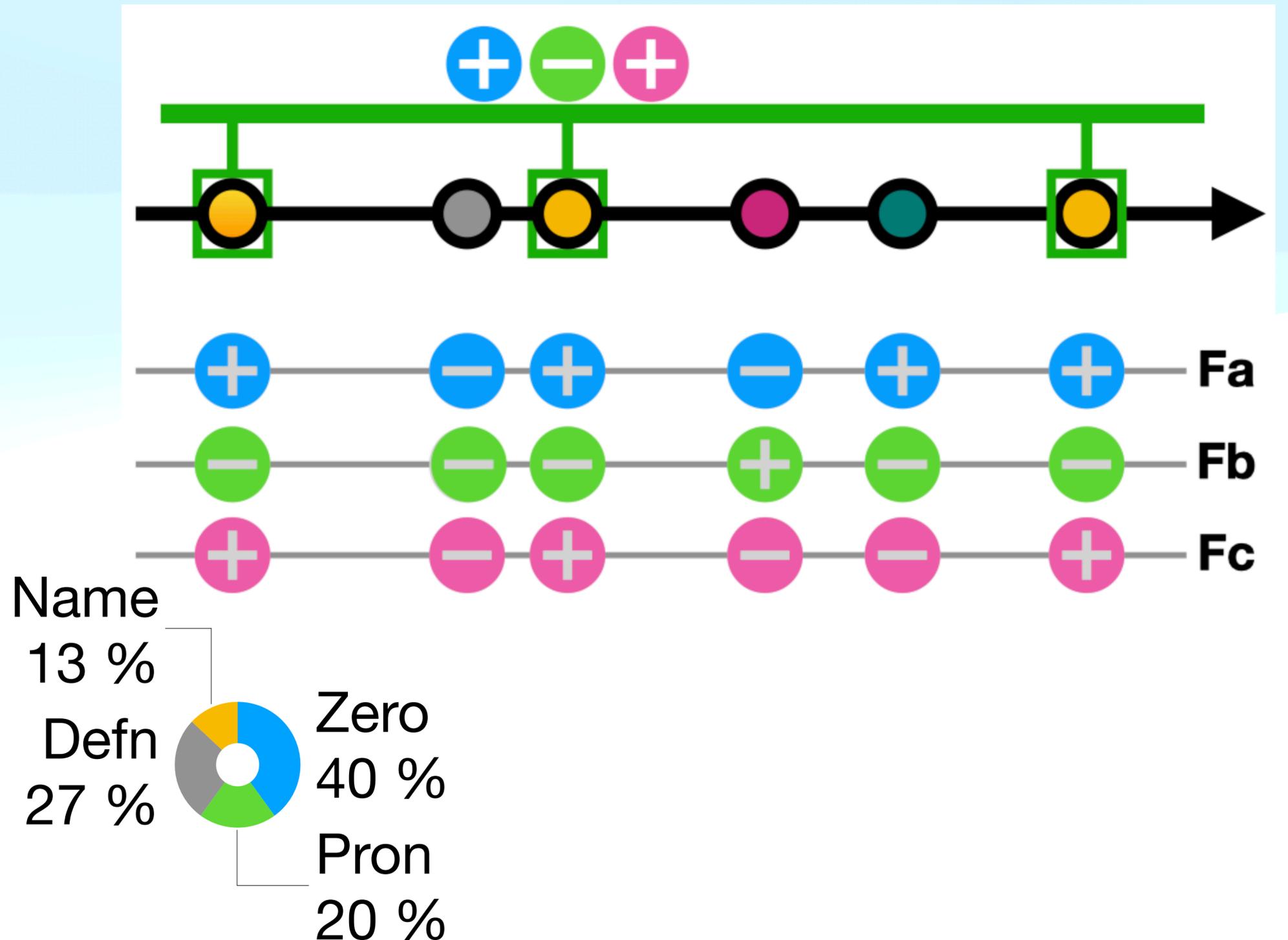
## Distributions of Forms

1. choose features that condition RE form

2. find all REs for each feature value combination

3. build distribution RE forms within each set

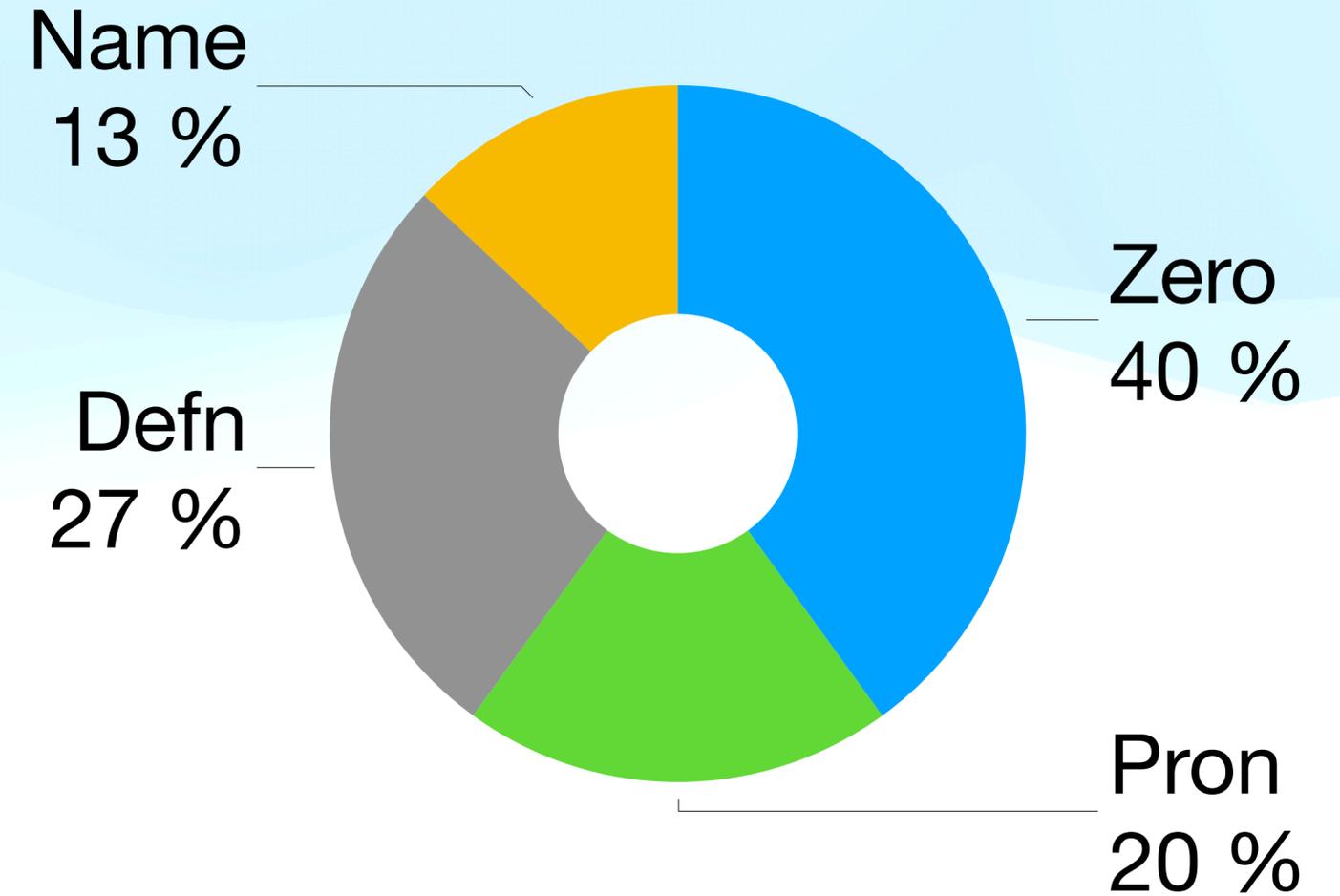
**thus** for each feature value combination, you have a distribution over RE forms



# Longitudinal Corpora of Variation

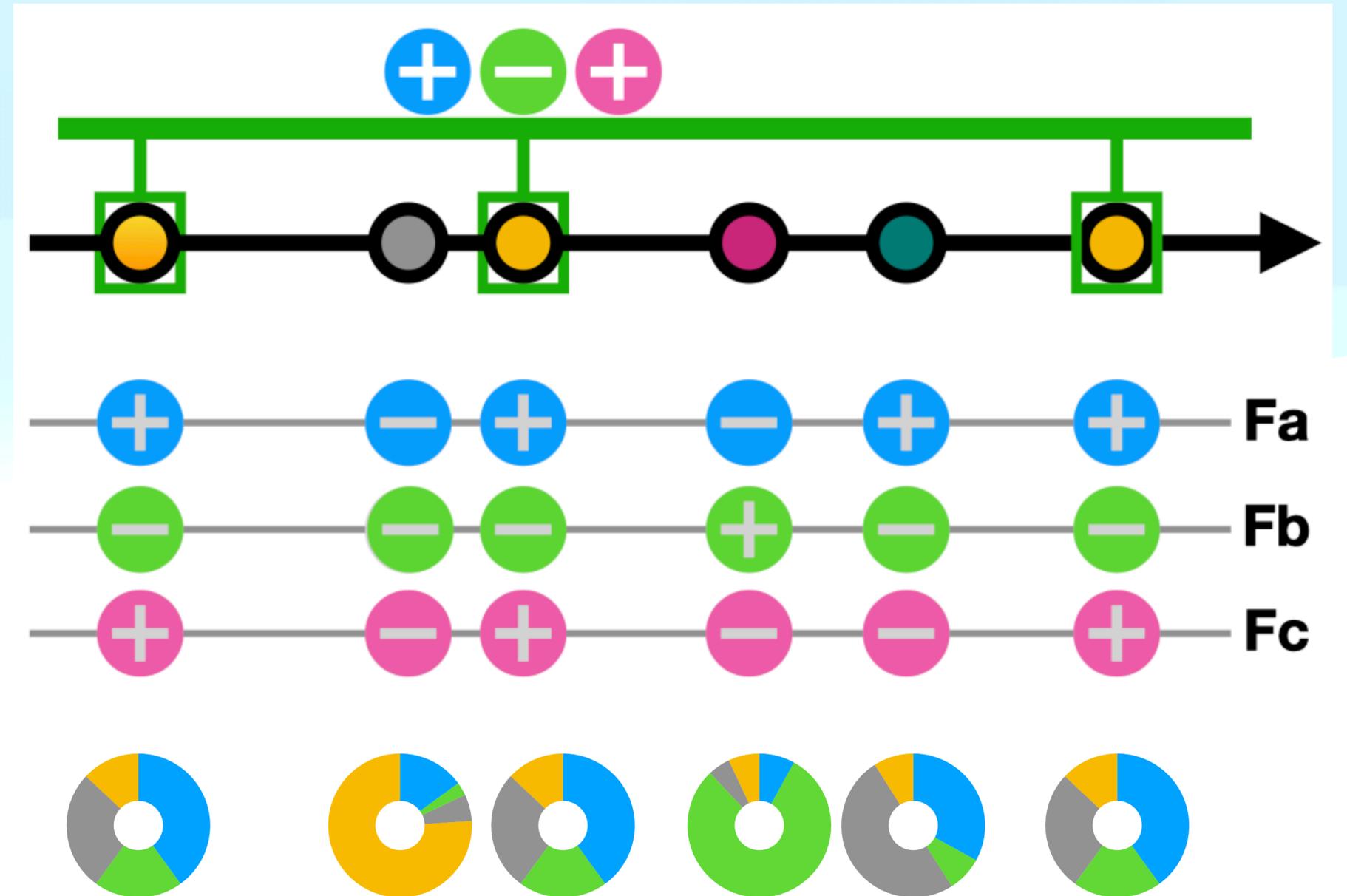
## Distributions of Forms

at each reference, the variation is the **distribution of forms** unpredictable from its feature values



# Longitudinal Corpora of Variation

## Variation



for each feature value combination, you have a distribution over RE forms

# Longitudinal Corpora of Variation

## Conclusion

**Longitudinal corpora of variation** are just corpora that capture at each reference:

how uncertain our best featural model of RE form is

you can ask me why featural models, not e.g. neural net models, in the questions 😊

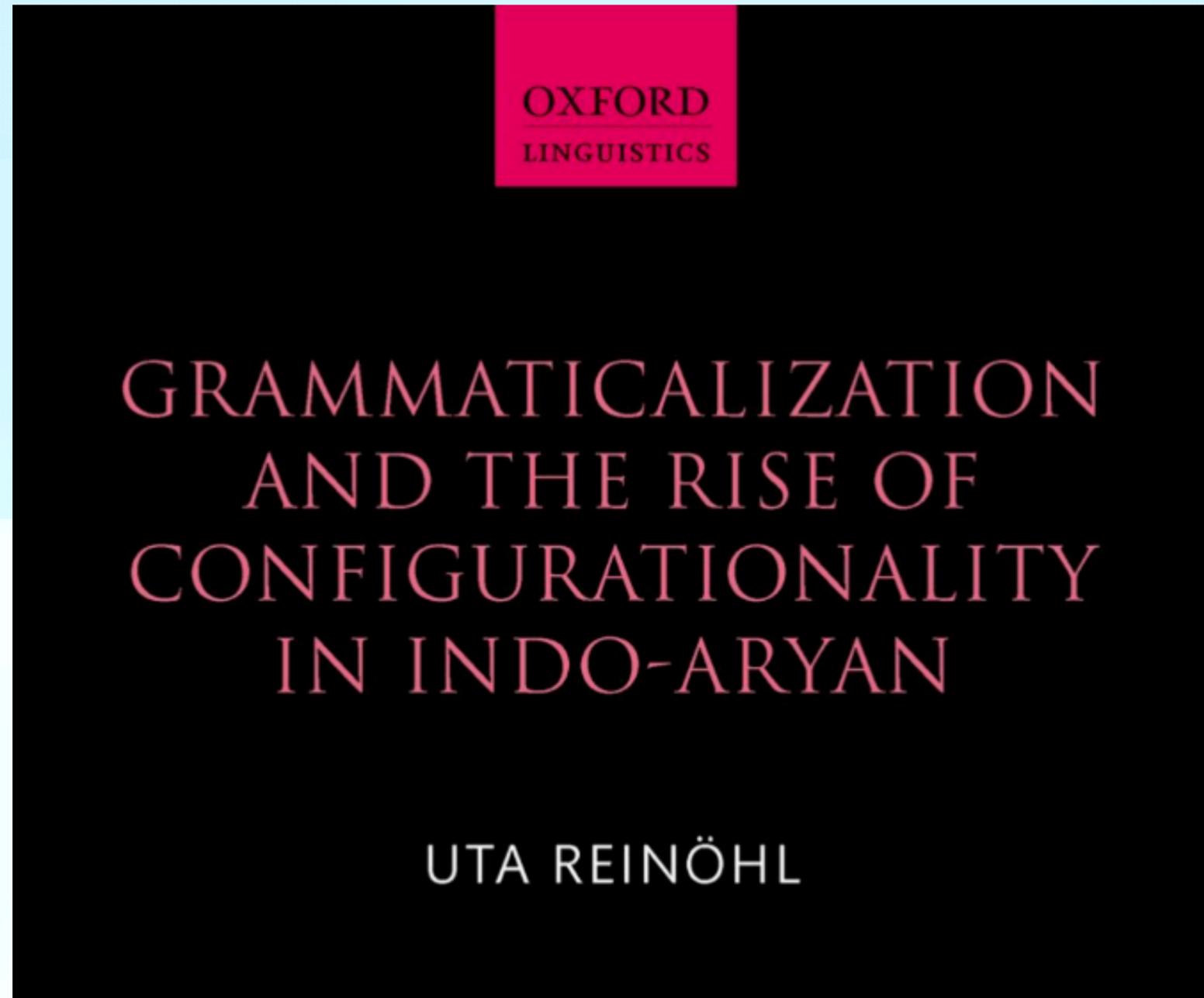
Separate research, looking at distributions of RE forms in this way showed a hitherto unnoticed feature of language, which tells us something about:

- human language processing, and
- language evolution

# Obligatoriness

# Metaphor-Driver Obligatoriness

## The Idea



Reinöhl, U. (2017). *Grammaticalization and the rise of configurationality in Indo-Aryan*. Oxford University Press.

# Metaphor-Driver Obligatoriness

## The Paper

[Open Access](#) | [Published: 30 July 2022](#)

### Compositionality, Metaphor, and the Evolution of Language

[T. Mark Ellison](#) & [Uta Reinöhl](#) 

[International Journal of Primatology](#) (2022) | [Cite this article](#)

437 Accesses | 4 Altmetric | [Metrics](#)

#### Abstract

One of the great unknowns in language evolution is the transition from unstructured sign combination to grammatical structure. This paper investigates the central — while hitherto overlooked — role of functor–argument metaphor. This type of metaphor pervades modern language, but is absent in animal communication. It arises from the semantic clash between the default meanings of terms. Functor–argument metaphor became logically possible in

Ellison, T. M., & Reinöhl, U. (2022). **Compositionality, Metaphor, and the Evolution of Language.** *International Journal of Primatology*, 1-17.



# Metaphor-Driver Obligatoriness

## The Phenomenon

**Everyone else arrived at the hotel on Thursday. John finally arrived [at it] on Friday night.**

**Everyone else arrived at the conclusion on Thursday. John finally arrived *at it* on Friday night.**

**Everyone else arrived at the conclusion on Thursday. \*John finally arrived on Friday night.**

# Metaphor-Driver Obligatoriness

## The Phenomenon

**Everyone else arrived at the hotel on Thursday. John finally arrived [at it] on Friday night.**

**Everyone else arrived at the conclusion on Thursday. John finally arrived *at it* on Friday night.**

In the metaphorical case, ***at it*** is not optional (without changing the meaning)

Generalisation: **If the referent of an argument is needed to force a metaphorical reading of the functor, that argument must be overt (cannot be zero)**

# Metaphor-Driver Obligatoriness In Corpora

In situations where zero  
would make sense ...

(numbers made up)

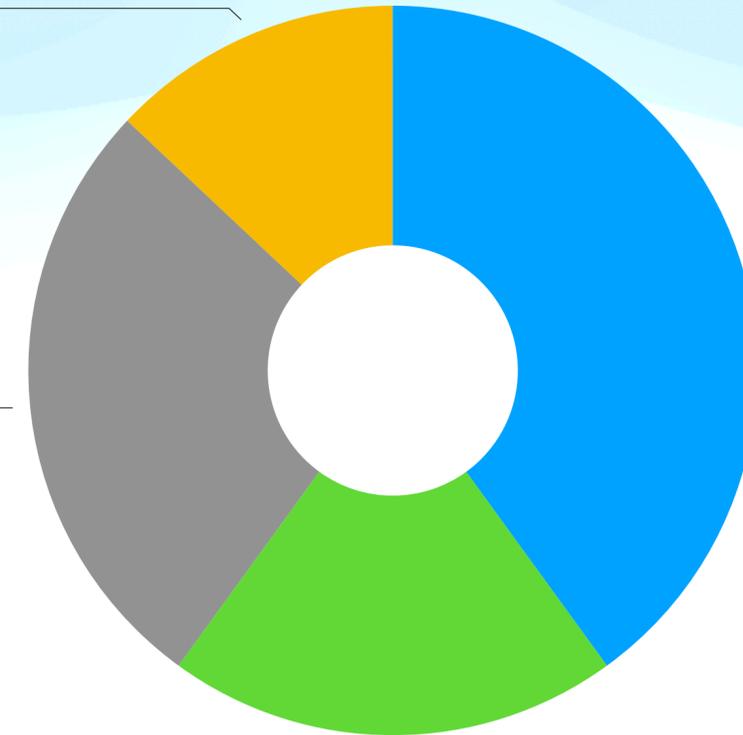
Name  
13 %

Defn  
27 %

**Expectation**

Zero  
40 %

Pron  
20 %



# Metaphor-Driver Obligatoriness In Corpora

In situations where zero would make sense ... it just does not occur in corpora

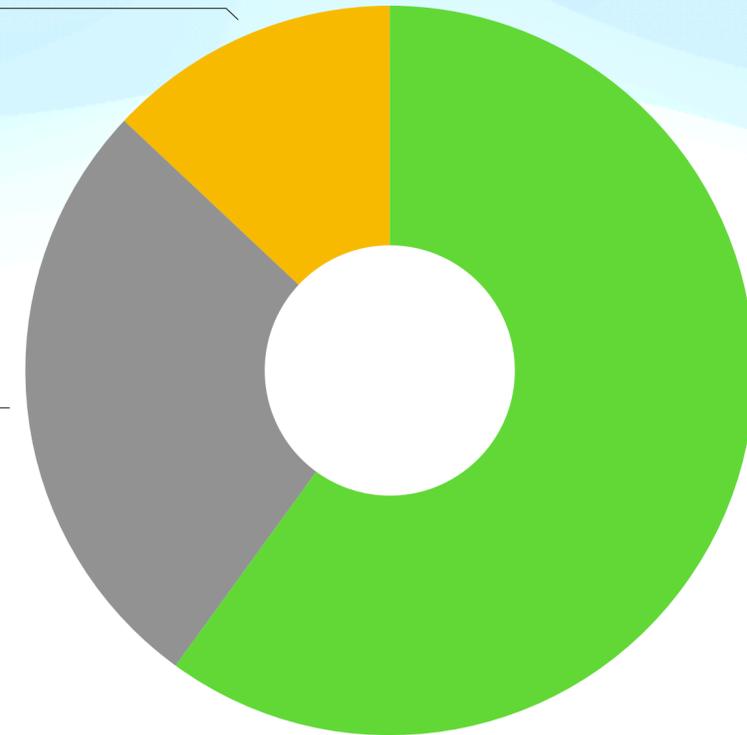
if the referent drives a metaphorical interpretation then it ends up being expressed by a pronoun instead

Name  
13 %

Defn  
27 %

Pron  
60 %

**Reality**



# Metaphor-Driver Obligatoriness In Corpora

In situations where zero would make sense ... it just does not occur in corpora

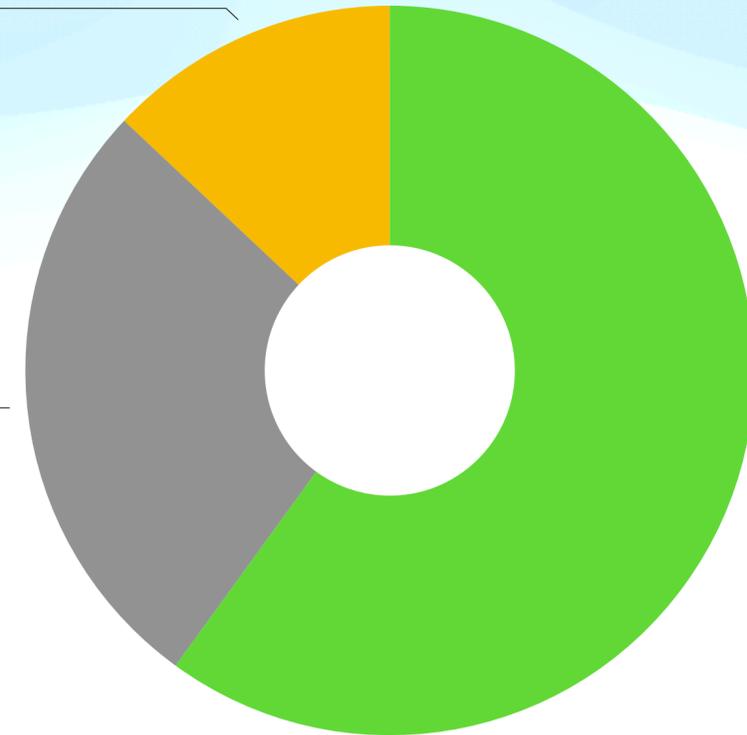
if the referent drives a metaphorical interpretation then it ends up being expressed by a pronoun instead

Name  
13 %

Defn  
27 %

Pron  
60 %

**Reality**



# Metaphor-Driver Obligatoriness

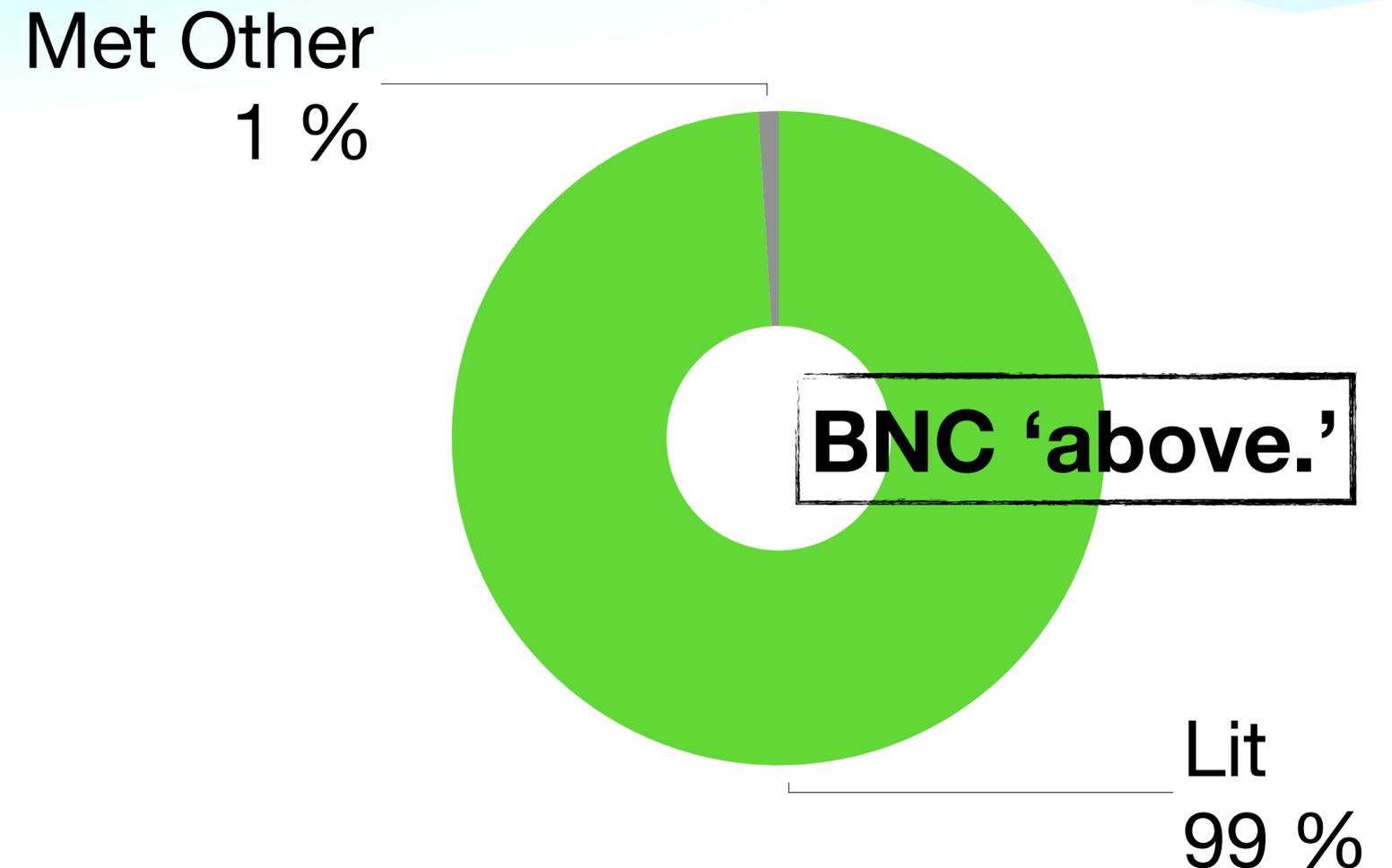
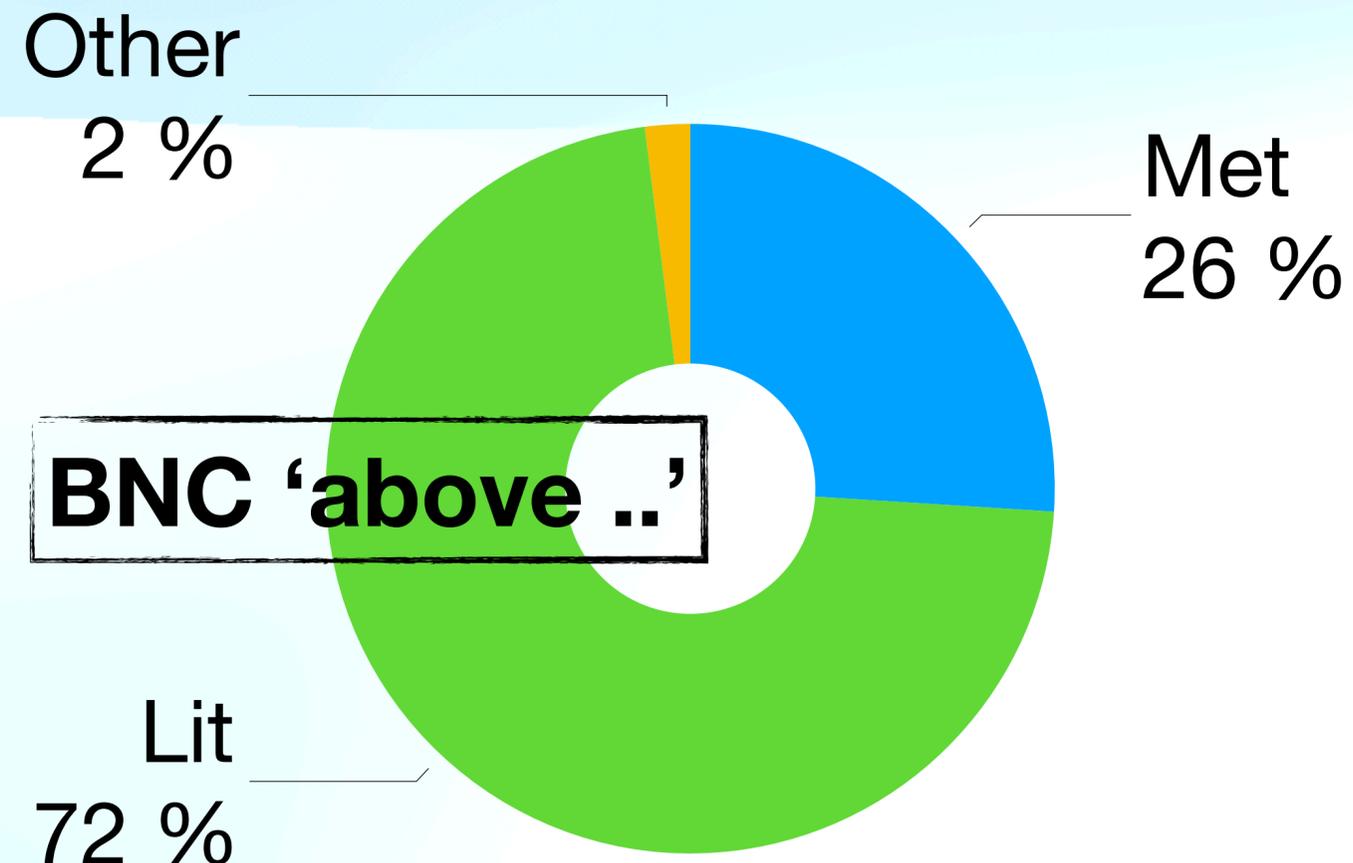
## In the BNC

Only the Queen is **above** the law and could not be subpoenaed. (Met)

Ian was hit just **above** the eye. (Lit)

Is there rank in the wives? ... I think *it comes from* below rather than from **above**. (Met Other)

The Ballachulish bridge overpowering **above** . (Lit)



# Wrapping Up

- Prominence is interesting for linguistics
  - but also important for language technologies, such as NLG
- RE form choice - is about prominence: getting clarity for minimum effort
- The 'right answers' are distributions of free variation, not forms
- *Gold standard* corpora of variation can be best made from existing corpora
- Looking at distributions of RE forms can lead to exciting new discoveries about language

**Thank you for your Attention**