You’re not the PoS of Me

Part-of-Speech Tagging as a Markup Problem

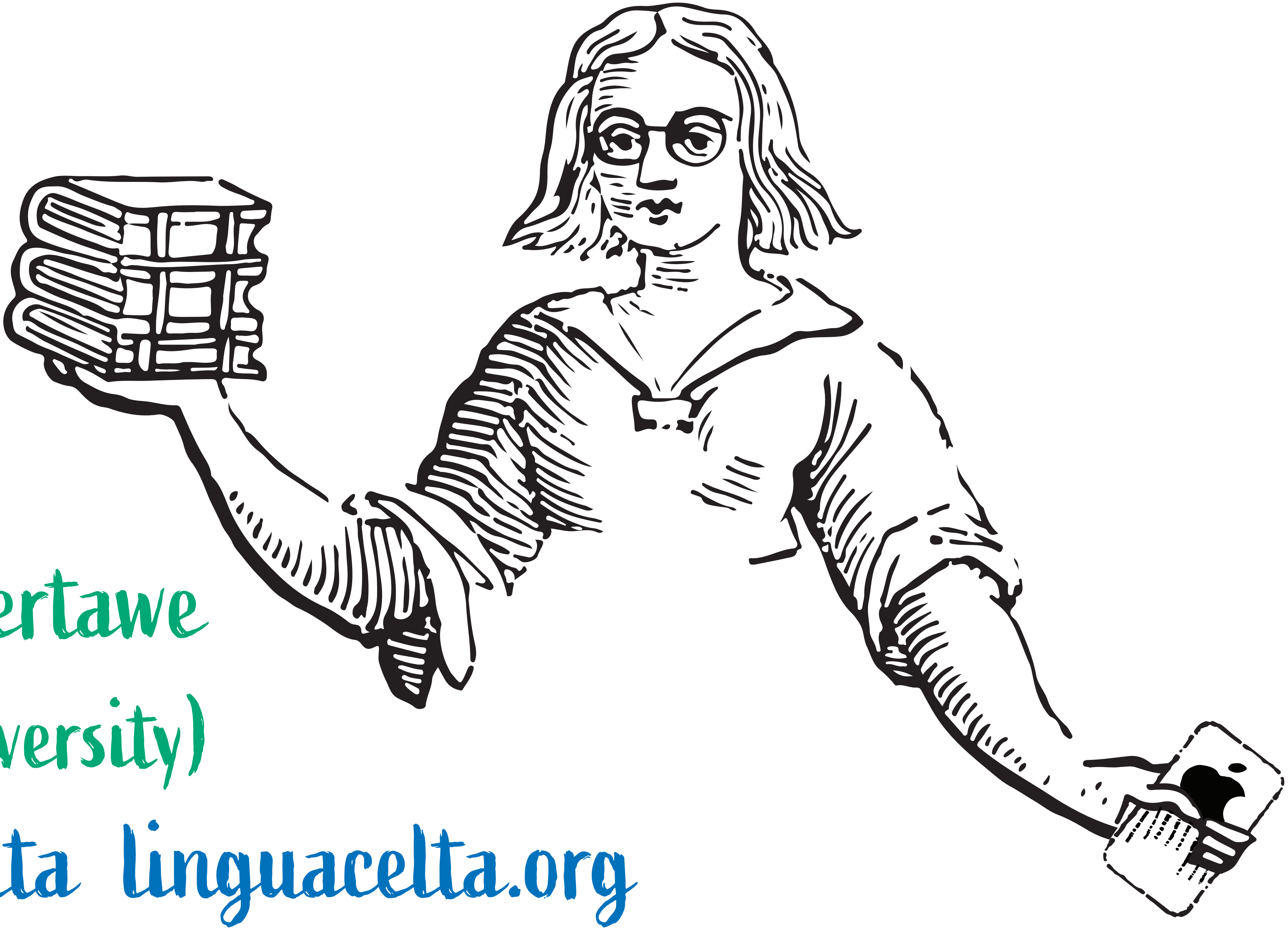
Bethan Tovey

Prifysgol Abertawe / CorCenCC

- Introductions
- Markup Theory…
- …and its Applications
- Conclusions

Introductions

Wherein we learn more about the wheres and the whys and the whats and all that jazz.



Prifysgol Abertawe
(Swansea University)

@LinguaCelta linguacelta.org

bst@bethan.wales

CorCenCC

Corpws Cenedlaethol Cymraeg Cyfoes

(Corpus - National - Welsh - Contemporary)

National Corpus of Contemporary Welsh

www.corcencc.org @CorCenCC

CorCenCC

10M words

4M written - 4M spoken - 2M electronic

Metadata including genre, age, location

My data

Welsh Twitter data:

- colloquial
- code-switched
- assimilated English words

Assimilation

English-origin words:

- morphological assimilation

  (e.g. verb endings, plural endings)

- orthographic assimilation

Assimilation

fundalŵ

Gwglais

sdwff

ffabiwlys

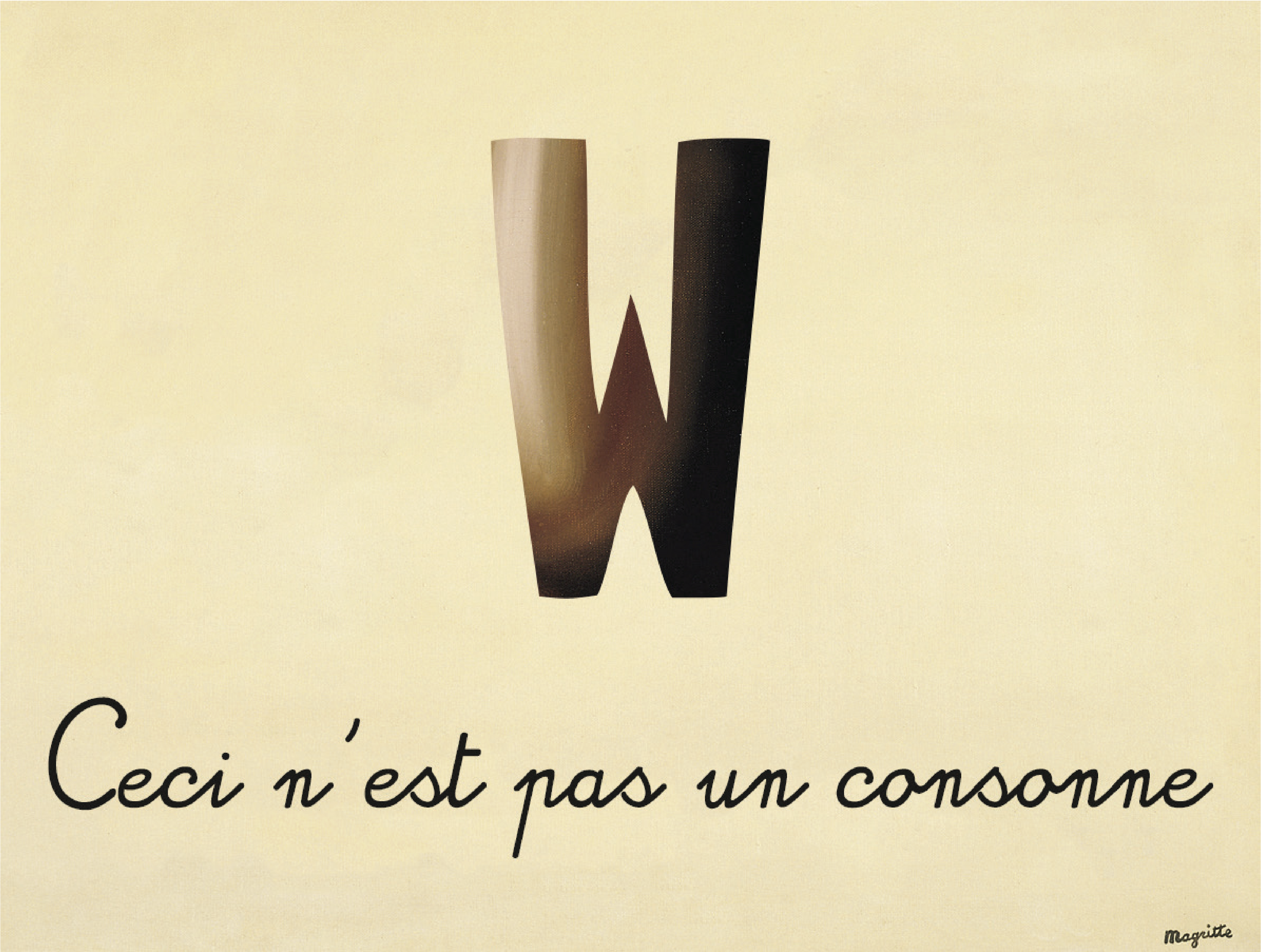
Assimilation

fundalŵ -> vindaloo

Gwglais -> I Googled

sdwff -> stuff

ffabiwlys -> fabulous


Ceci n'est pas un consonne
Magritte

CorCenCC

CyTag: PoS tagger for Welsh texts

Which is where this story starts...

HaHa! A PoS tagger for Welsh, forsooth!

But it taggeth not bilingual texts…

Then I must learn Python and rewrite the tagger to my own ends.

It surely cannot be hard...

Ok, LOL.

Mayhap it is a little hard.

What the actual heck goeth on here?

I truly cannot even!

Thou shouldest introduce a teſt suite...

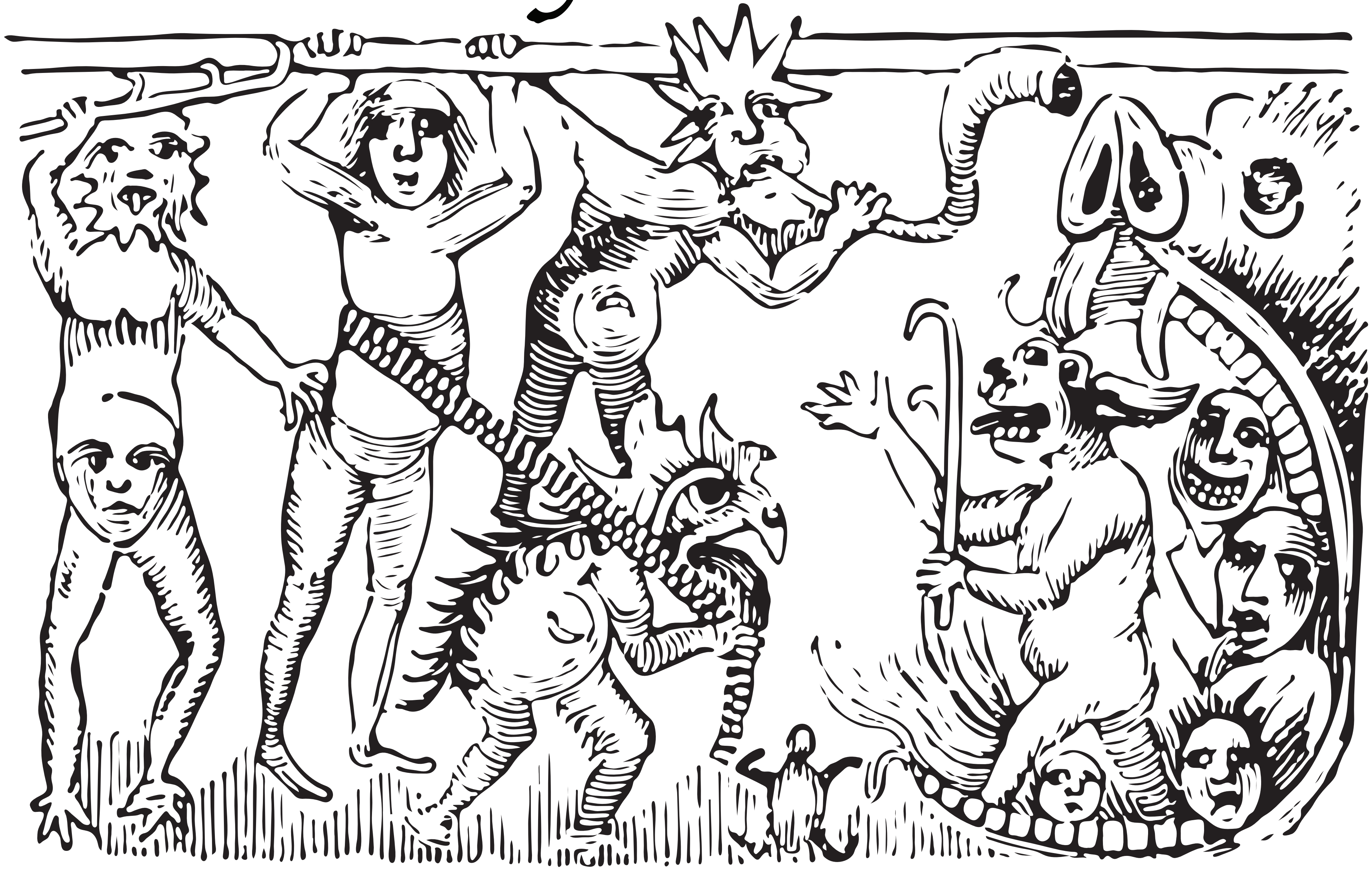
Wherefore saydest thou not this many moons ago?

I am a wretch
who cannot code for toffee.

It is Memorial Day! We shall miss the pool party!

I care nothing for parties when there is code to write!

This is fine.



I have, like, totes got this now.

YASSS! Queen of the Pythons!

DERWen

Derwen = "oak"

Everything is a tree (if you squint hard).

DERWen

Dichell Esboniadol Ramadegol y Wenglish

(Device-Explanatory-Grammatical-the-Wenglish)

Wenglish Grammatical-Explanatory Device

What is PoS tagging?

I am a sentence

What is PoS tagging?

I	am	a	sentence
pronoun	verb	article	noun
1.sg.	pres.1.sg.		sg.
personal		indefinite	

What is PoS tagging?

PoS tagging is markup

Markup Theory

Wherein we encounter some important threads in theoretical approaches to markup.

What is the text?

Sperberg-McQueen (1991)

The text is an abstraction.
The document is a representation.
Documents are never objective representations of the text.

DeRose et al. (1990)

The content elements constitute the document.

What is the vocabulary?

Klein and Hirscheim (1987)

The Universe of Discourse

Realist or Nominalist?

Implications of markup

Piez (2001)

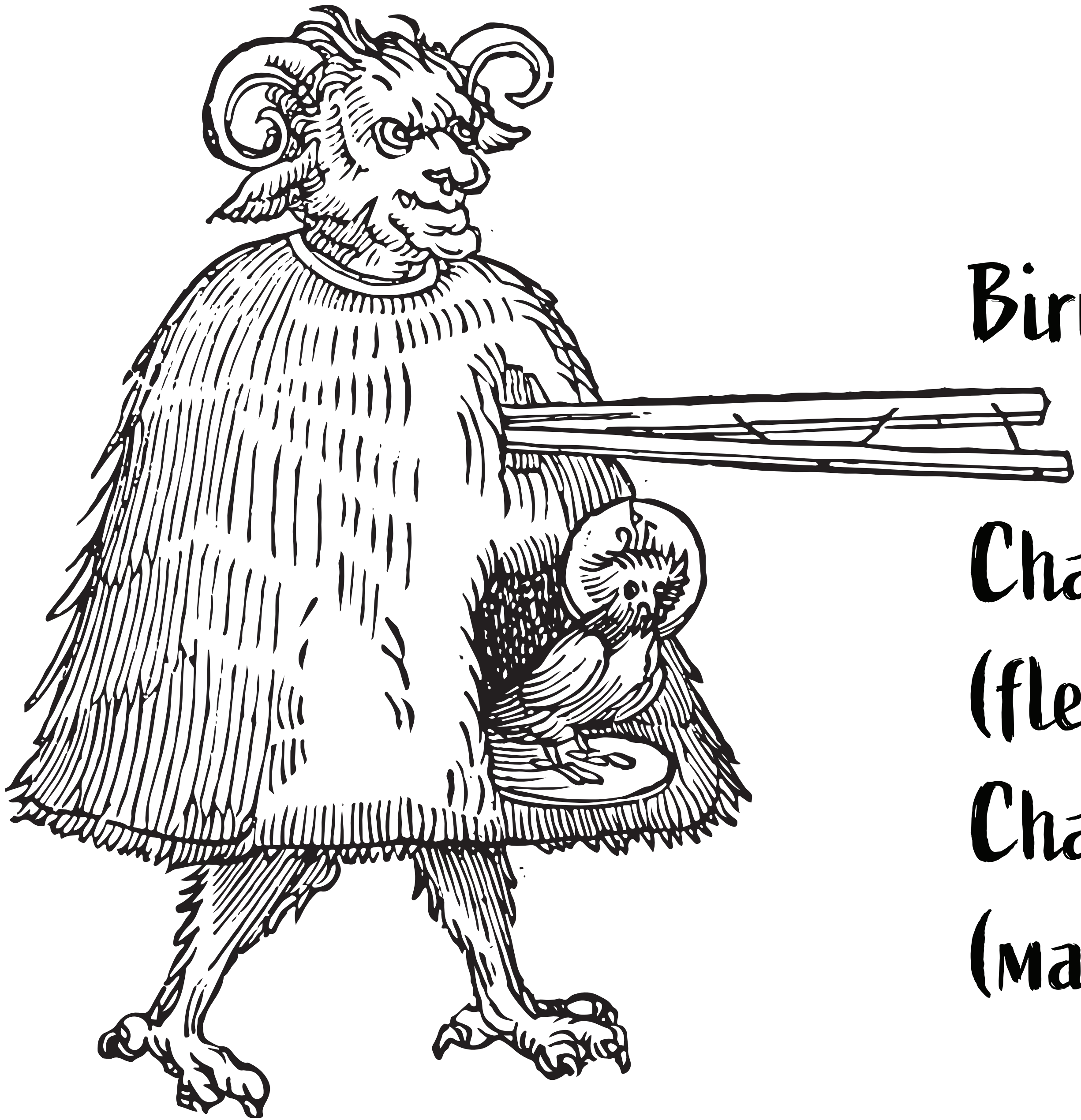
Metaleptic markup
Proleptic markup

Pre-existing features or future uses?

Sperberg-McQueen (1991)

Markup shapes what (we imagine) we can do with the text.

Anomalous Data



Birnbaum and Mundie (1999)

Change the schema?

(flexibility / “escape hatches”)

Change the data?

(make it conform to the schema)

Markup semantics

Renear et al. (2002)

Formal specification of markup semantics to reduce:

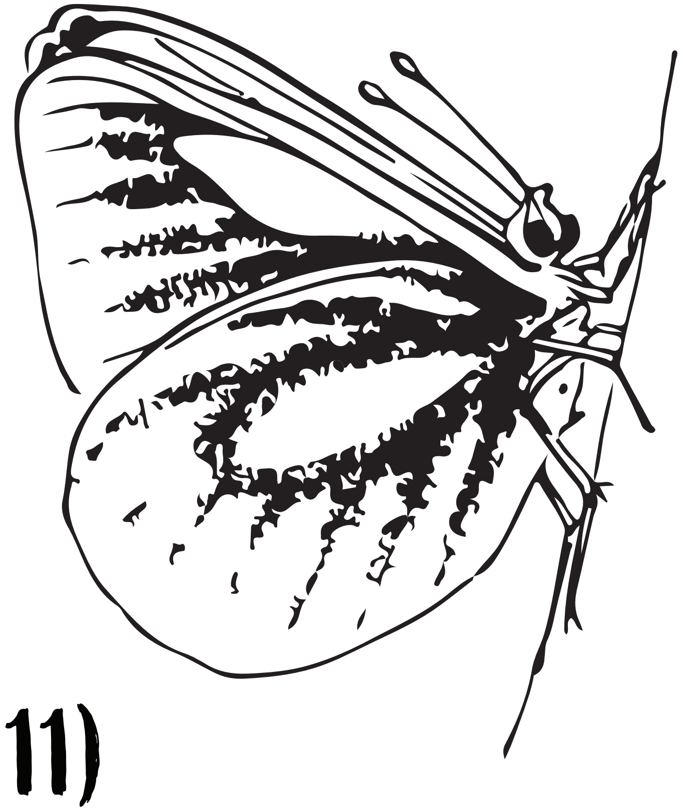
- tag abuse
- ambiguity
- reliance on conjecture

Vocabulary conversion

Tennison (2002)

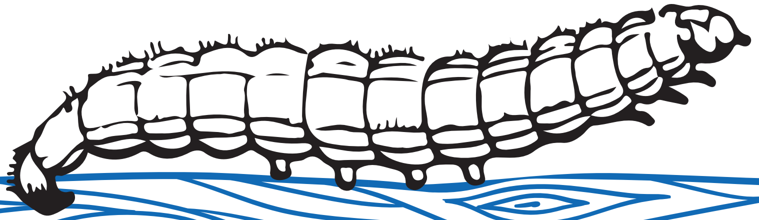
Vocabulary distance:

- what is lost?
- what is gained?



Sperberg-McQueen (2011)

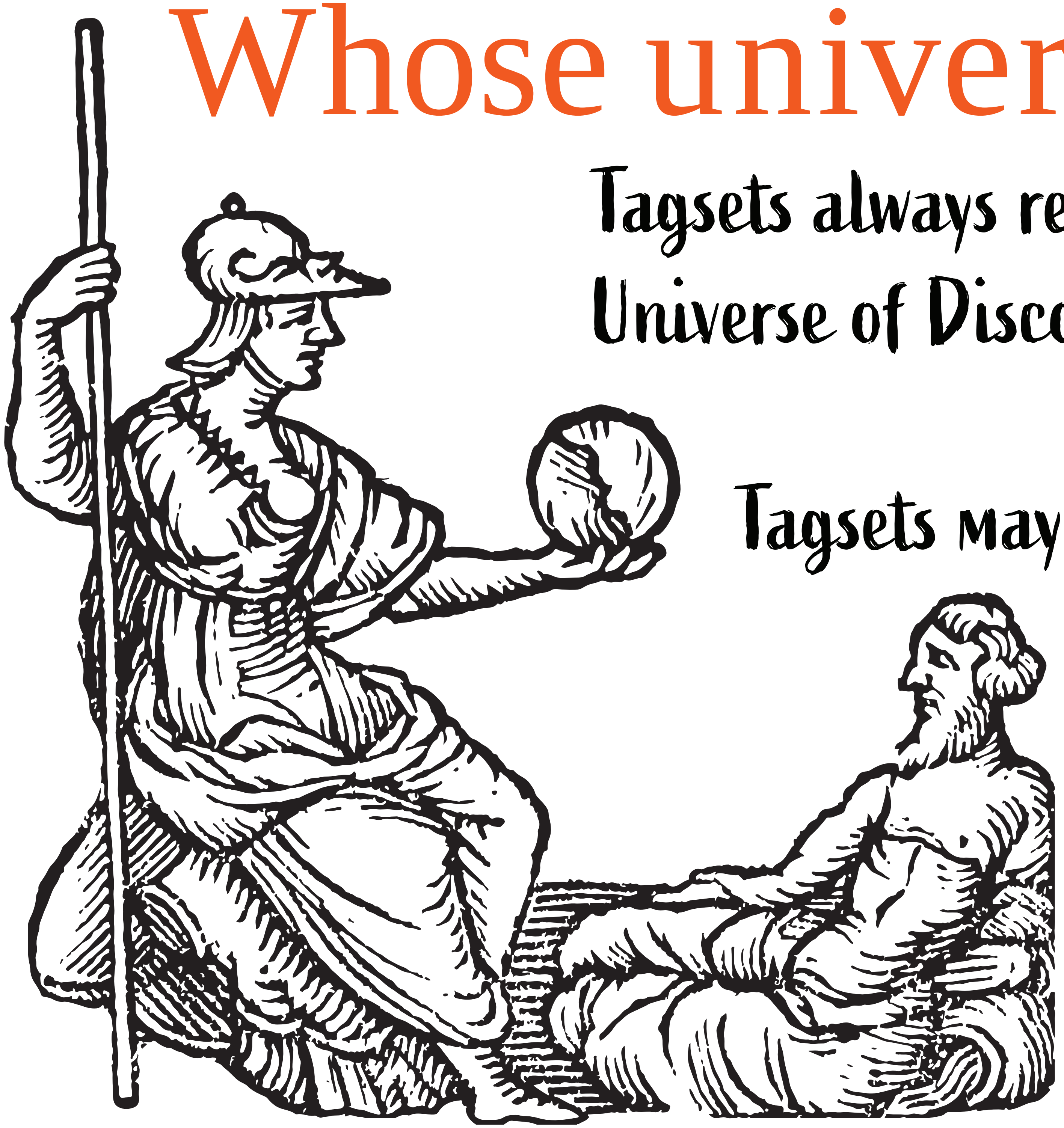
"Noise-free, lossless conversion."



Applications

Wherein we consider how the foregoing theories can illuminate PoS tagging as a concept.

Whose universe?



Tagsets always represent a socially-constructed Universe of Discourse.

Tagsets may represent a model of predicted end-users and end-uses.

Tagsets make some NLP/Corpus tasks easier than others.

Representing information

Cross-language tagsets assume that different languages consist of the same basic entities.

Conversion between PoS tagsets is rarely lossless or noise-free.

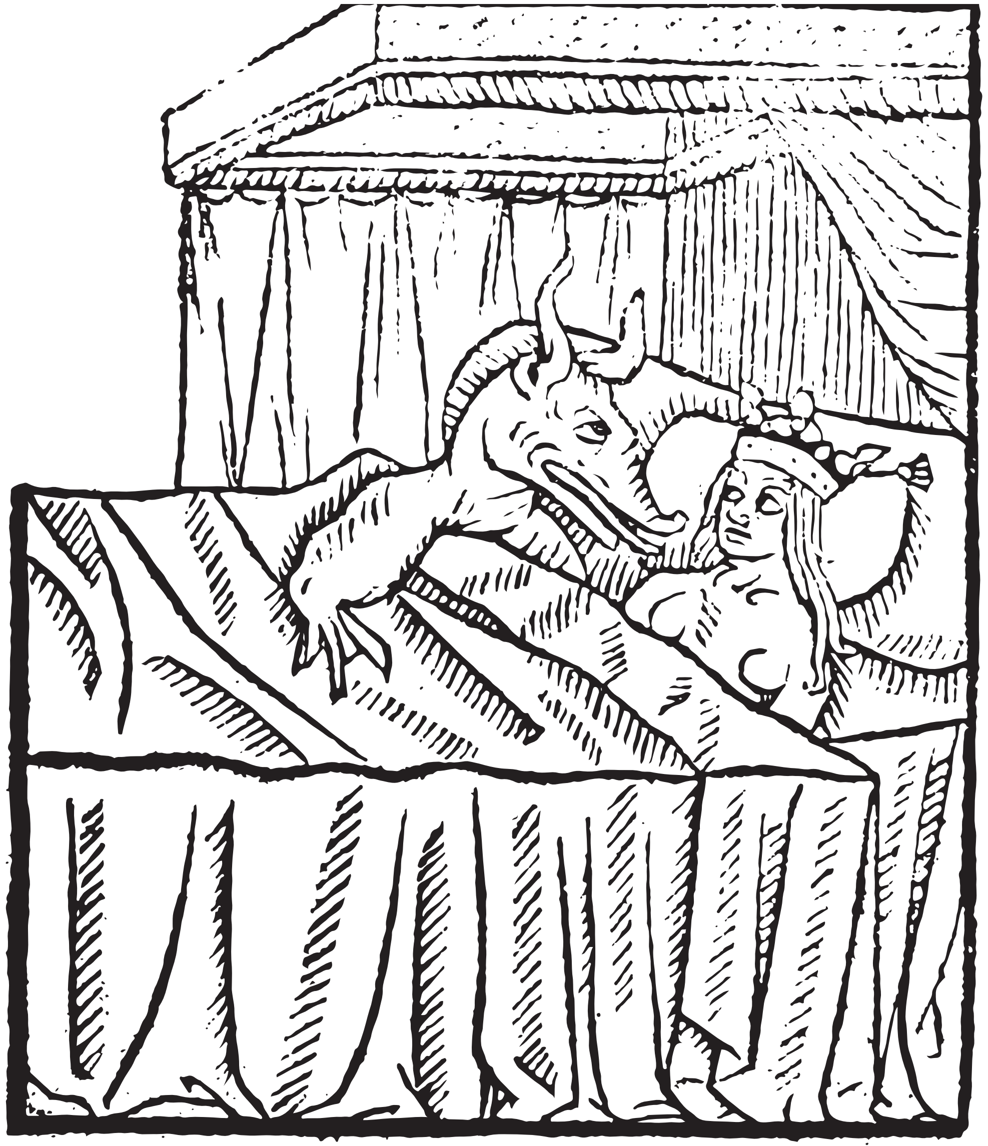
Tag abuse can result from underspecified semantics.

Difficult data

Anomaly exists with respect to the text's ideal/intended structure.

Modelling anomaly implies that there is a way to be non-anomalous.

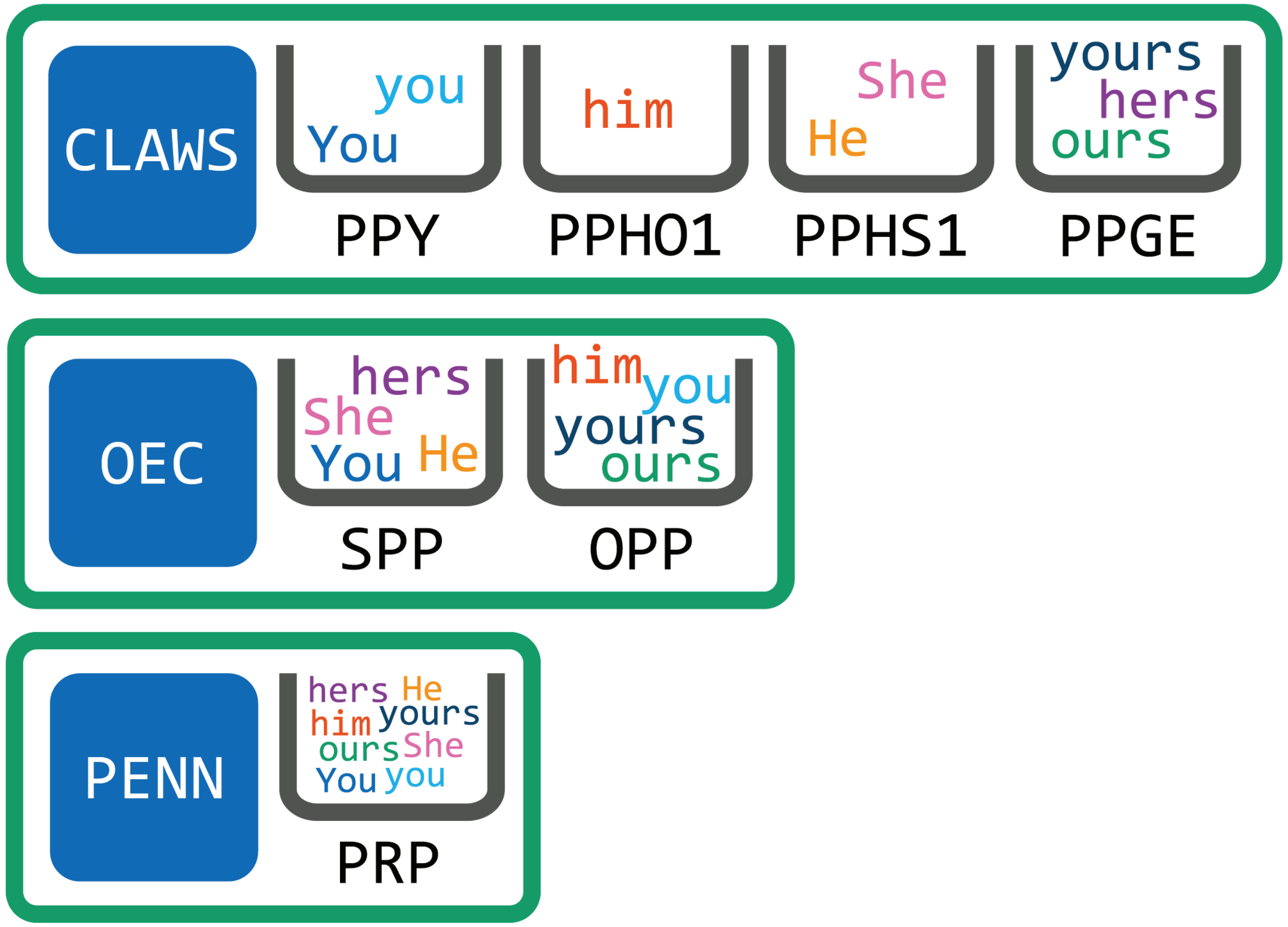
Tagsets based on specific dialects are inappropriate for other dialects.



You are different from him

He needs five of you

She is ours and hers is yours



CLAWS

you
You

PPY

him

PPHO1

She
He

PPHS1

yours
hers
ours

PPGE

OEC

hers
She
You He

SPP

him you
yours
ours

OPP

PENN

hers He
him yours
ours She
You you

PRP

Person

Number

Subject/Object

Gender

Possessive?

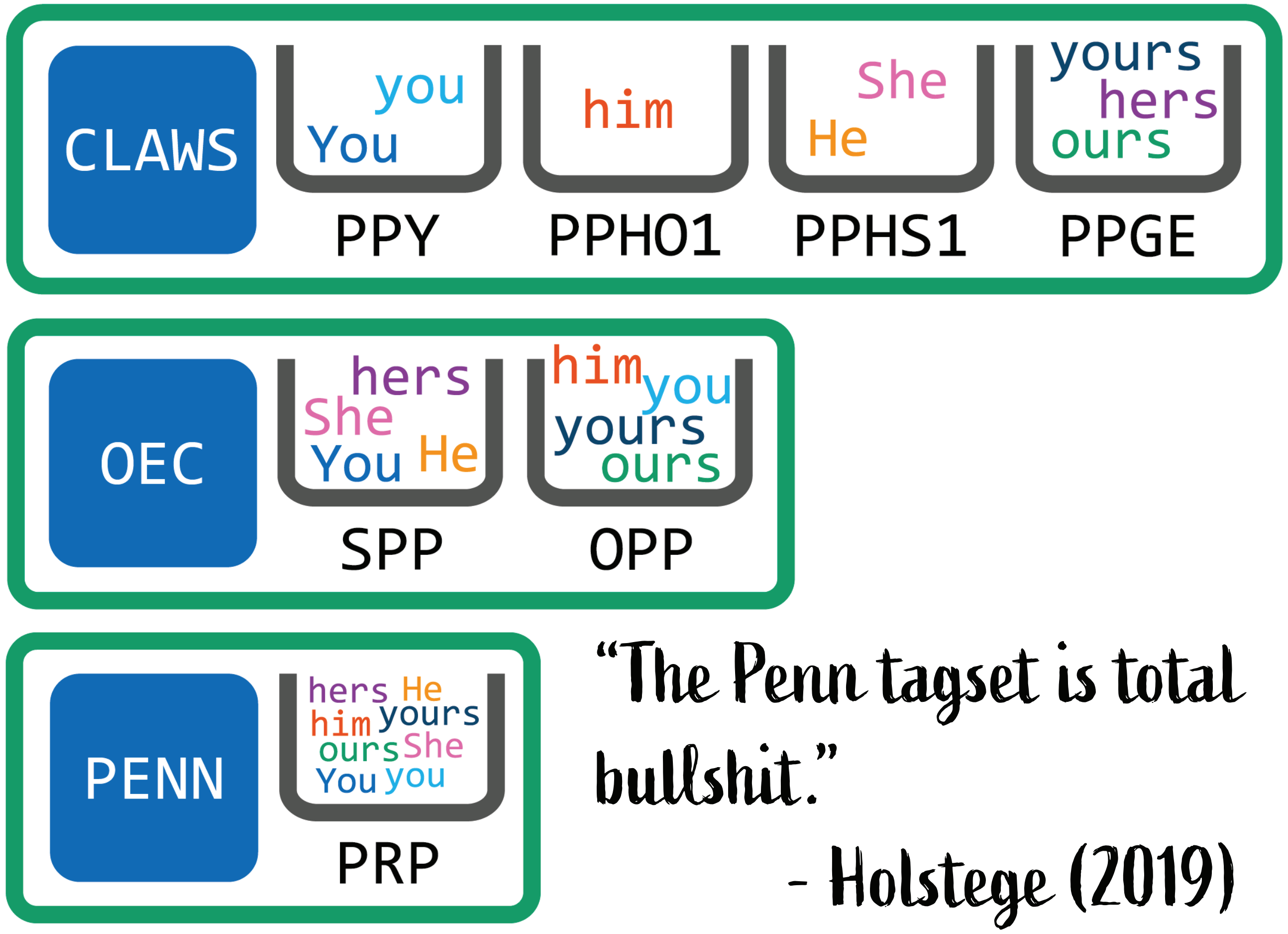
“[T]he Penn Treebank tagset does not distinguish subject pronouns from object pronouns even in cases where the distinction is not recoverable from the pronoun’s form, as with you, since the distinction is recoverable on the basis of the pronoun’s position in the parse tree in the parsed version of the corpus.”

- Taylor, Marcus, and Santorini (2003)

You are different from him

He needs five of you

She is ours and hers is yours



CLAWS

you You — PPY

him — PPHO1

She He — PPHS1

yours hers ours — PPGE

OEC

hers She You He — SPP

him you yours ours — OPP

PENN

hers He him yours ours She You you — PRP

"The Penn tagset is total bullshit."

- Holstege (2019)

Person

Number

Subject/Object

Gender

Possessive?

“[W]hereas many POS tags in the Brown Corpus tagset are unique to a particular lexical item, the Penn Treebank tagset strives to eliminate such instances of lexical redundancy.”

- Taylor, Marcus, and Santorini (2003)

pron-2-pl: ["you"]

How can I help ___ ?

pron-2-pl: [“you”, “y’all”, “yourselves”, “yez”, “yins”, “ye”]

The "verbnoun"

Byddaf i yn mynd

be[future 1 sg.] I in go

I will be going

Mae hi wedi canu

be[present 3 sg.] she after sing

She has sung

Dwi'n hoffi nofio

be[pres. 1 sg.]-I-in like swim

I like swimming

Conclusions

Wherein we conclude. (The clue's in the name.)

Emergent principles

- Be precise (and honest) about what the tagset represents

Emergent principles

- Be precise (and honest) about what the tagset represents
- Be careful when modelling anomaly

Emergent principles

- Be precise (and honest) about what the tagset represents
- Be careful when modelling anomaly
- Specify the semantics of tags (not their contents)

Emergent principles

- Be precise (and honest) about what the tagset represents
- Be careful when modelling anomaly
- Specify the semantics of tags (not their contents)
- Don't throw away distinctions needlessly

Emergent principles

- Be precise (and honest) about what the tagset represents
- Be careful when modelling anomaly
- Specify the semantics of tags (not their contents)
- Don't throw away distinctions needlessly
- Train and test with the weirdest instances you can find

Emergent principles

- Be precise (and honest) about what the tagset represents
- Be careful when modelling anomaly
- Specify the semantics of tags (not their contents)
- Don't throw away distinctions needlessly
- Train and test with the weirdest instances you can find
- Learn Welsh

Thanc-iw feri mytsh!