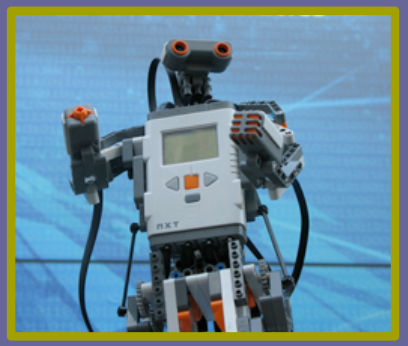
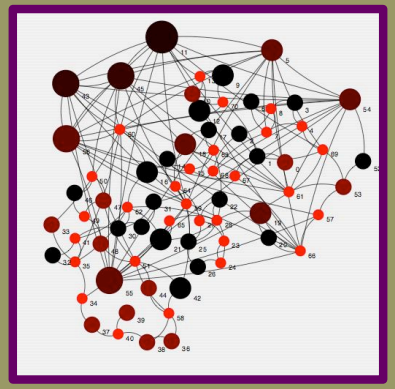
COMP40020 Human Language Technologies

Boundaries, Tokens & Corpora

February 2019





Prof. Julie Berndsen
School of Computer Science
Julie.Berndsen@ucd.ie

HLT8

HLT8

Contents:

- Types of corpora
- Processing raw text
- Text processing using NLTK

Aim:

- To give an overview of the basics of corpus-based techniques and how to apply these in NLTK → with more details at the workshop on Wednesday.

Some Literature

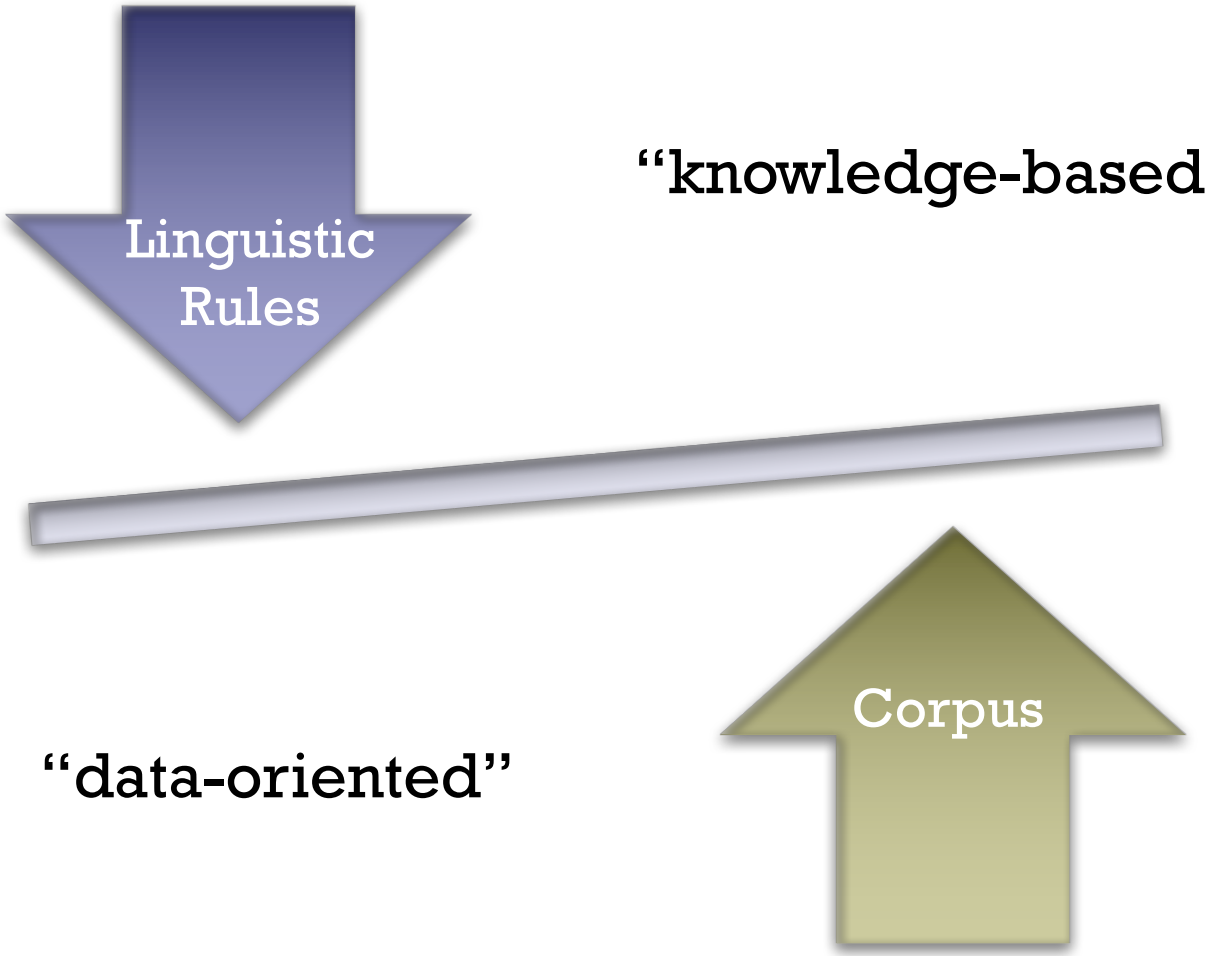
HLT8

- Manning, C. D. & H. Schütze (1999): *Foundations of Statistical Natural Language Processing*. Cambridge Massachusetts: MIT Press.
  - http://nlp.stanford.edu/links/statnlp.html

- Bird, S; E. Klein & E. Loper (2009): *Natural Language Processing with Python*, O'Reilly Media.
  - http://www.nltk.org/book/ch03.html

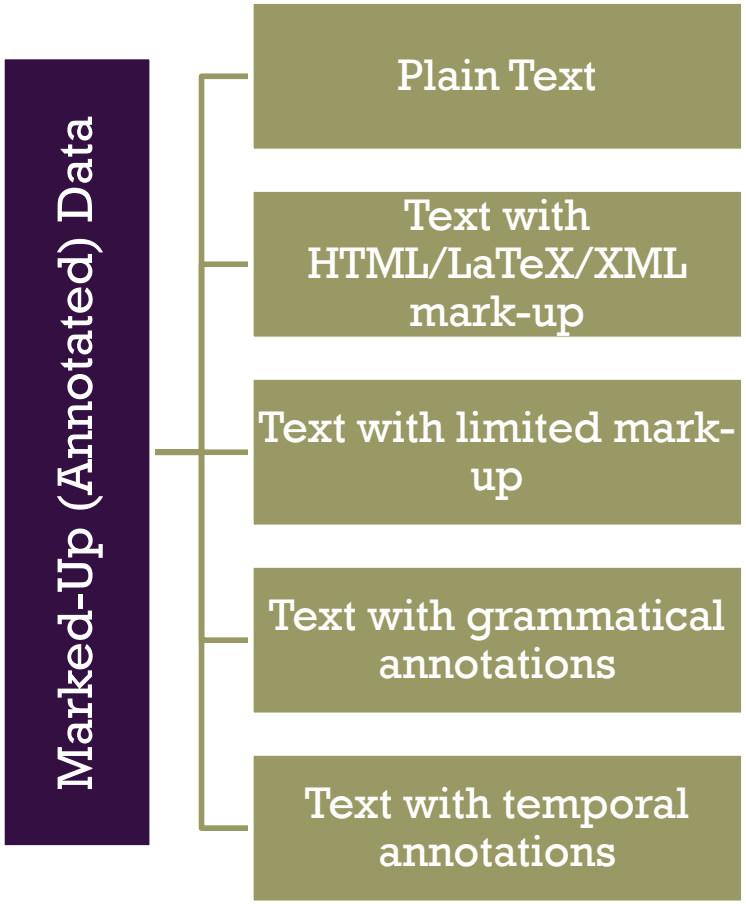
UCD DUBLIN

+ Achieving Coverage…



Marked-Up Data and Corpora

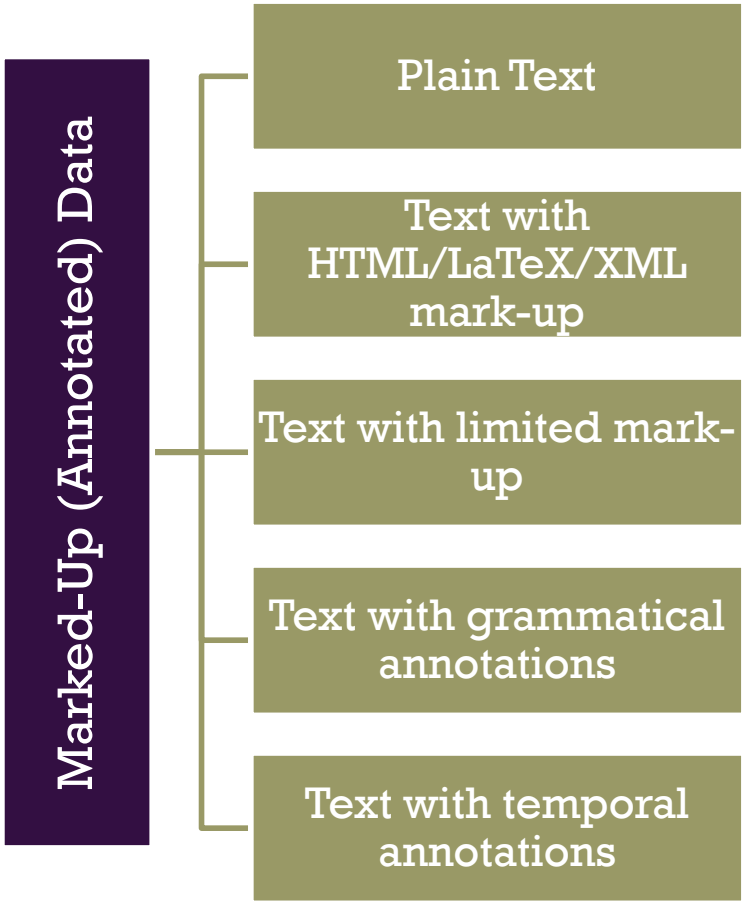
HLT8

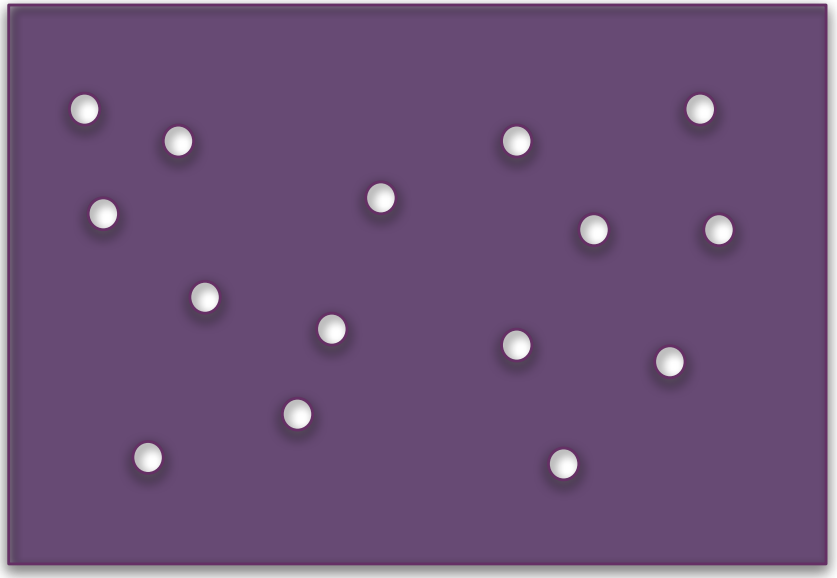


UCD DUBLIN

Marked-Up Data and Corpora

HLT8





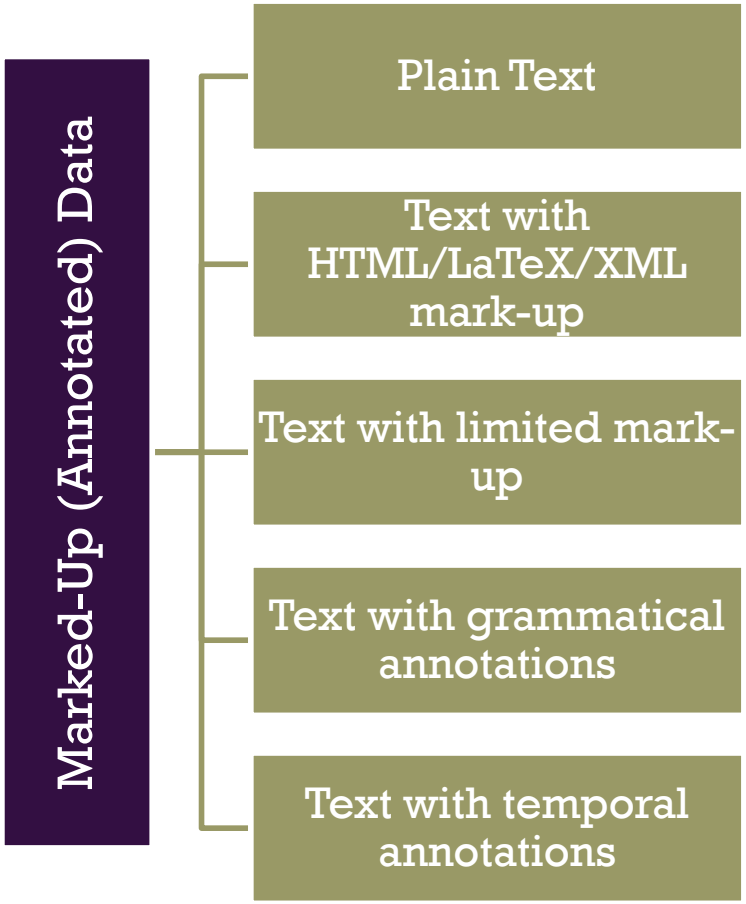
Isolated Corpora

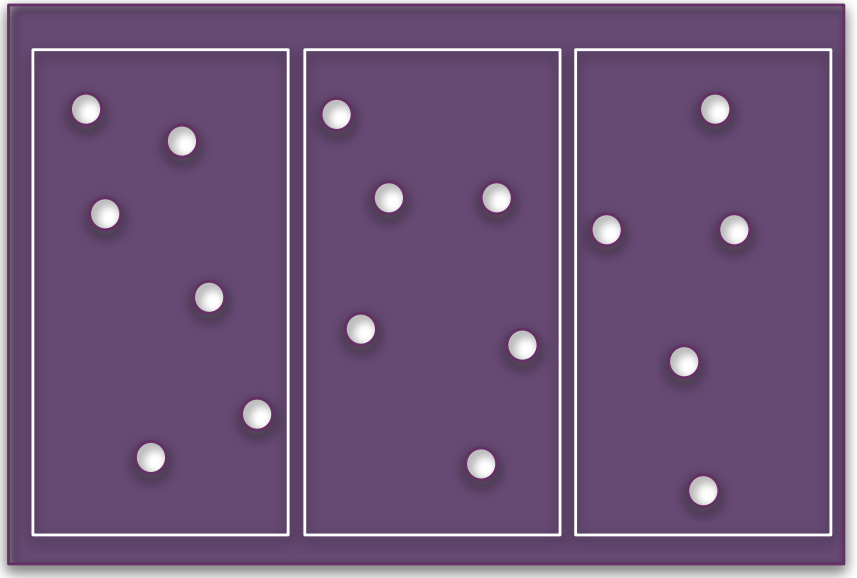
NLTK Book Figure 2.1.3

UCD DUBLIN

Marked-Up Data and Corpora

HLT8





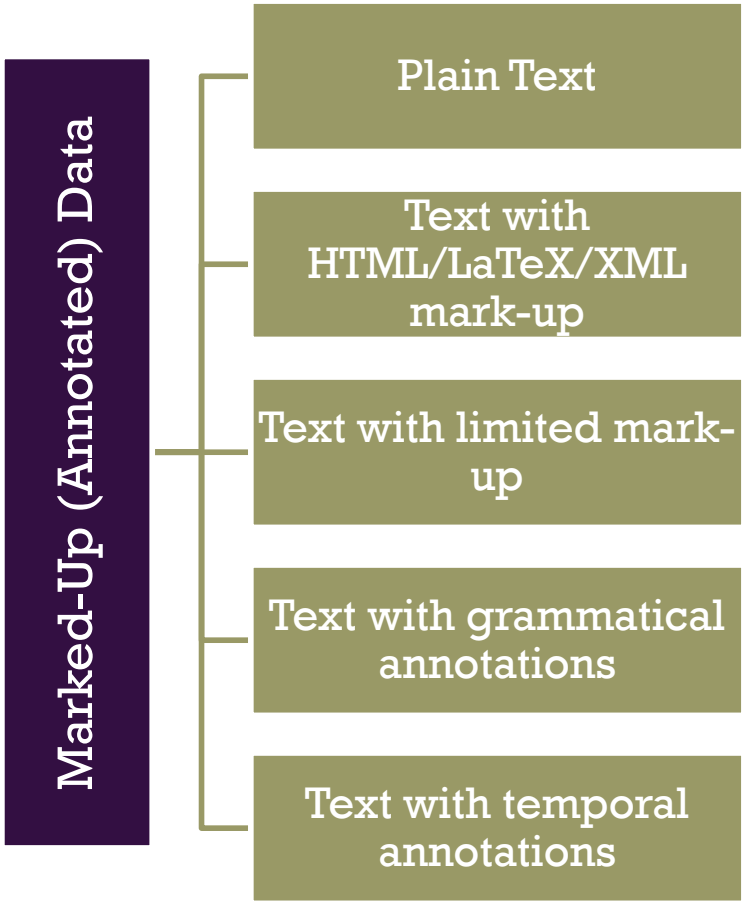
Categorised Corpora

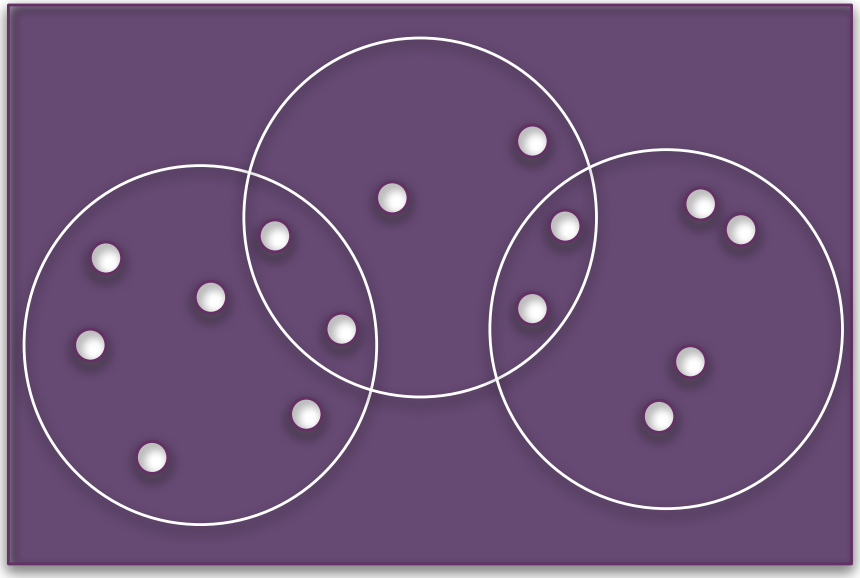
NLTK Book Figure 2.1.3

UCD DUBLIN

Marked-Up Data and Corpora

HLT8





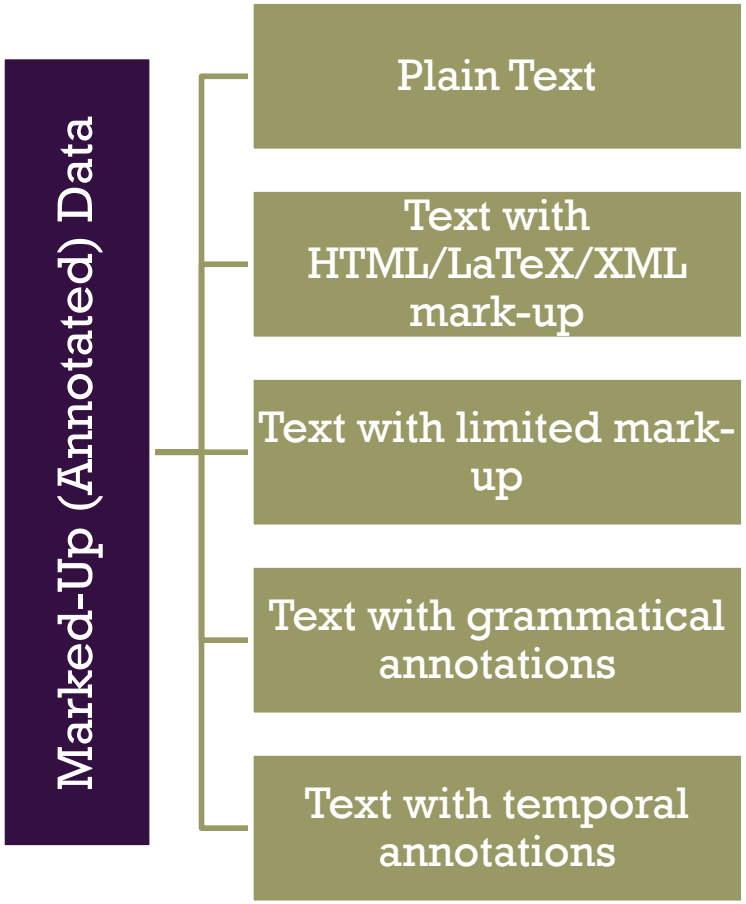
Overlapping Corpora

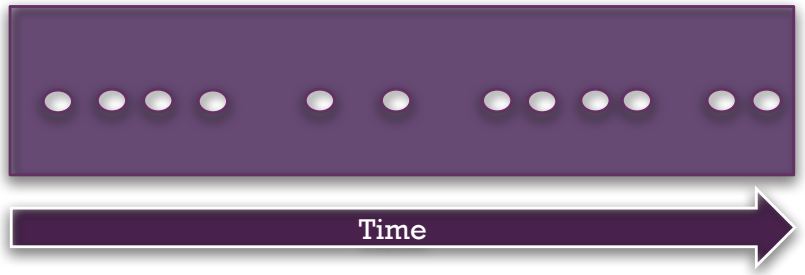
NLTK Book Figure 2.1.3

UCD DUBLIN

Marked-Up Data and Corpora

HLT8





Temporal Corpora

NLTK Book Figure 2.1.3

UCD DUBLIN

Linguistic Corpora

HLT8

- Linguistic Data Consortium:
  https://www.ldc.upenn.edu/
- European Language Resources Association:
  http://www.elra.info/en/
- Oxford Text Archive:
  http://ota.ox.ac.uk/
- British National Corpus:
  http://www.natcorp.ox.ac.uk/
- See also: http://www-nlp.stanford.edu/links/statnlp.html#Corpora

UCD DUBLIN

Identifying Sentences

What is a sentence?

- something ending with **.** **?** or **!**

but actually only 90% of **.** mark the end of a sentence

"Algorithm"

- Place sentence boundaries after all occurrences of . ? ! (and maybe ; : --)
- Move the boundary after following quotation marks, if any.

From Manning & Schütze (1999)

Identifying Sentences

- Disqualify a . boundary in the following circumstances:
  - if it is preceded by a known abbreviation of a sort that does not occur word finally, but is commonly followed by a capitalised proper name, such as Prof. or vs.
  - if it is preceded by a known abbreviation and not followed by an uppercase word. This will deal correctly with most usages of abbreviations like etc. or Jr.
- Disqualify a boundary with a ? or ! if it is followed by a lowercase letter or a known name
- Regard remaining sentence boundaries as sentence boundaries.

From Manning & Schütze (1999)

Text: Newspaper (Print)

HLT8

Exposed: the huge class divide in our universities

Katherine Donnelly
Education Editor

THE big class divide in Irish third-level education is exposed in new figures showing how students from better-off families capture most of the places in the country's universities.

It means they are filling the lion's share of the most sought-after honours degree courses, including elite high-points programmes.

Katherine Donnelly
Barriers to equal access to third-level still exist
Analysis, page 10

The first-ever breakdown of the proportion of first-year student grant-holders in each college highlights a wide disparity between the universities and the institutes of technology.

Figures compiled by the Higher Education Authority (HEA), based on data supplied by the student grant agency, SUSI, show that, overall, 46pc of first-years in 2013-14 received a maintenance grant to help cover their living costs.

However, the breakdown by sector reveals 56pc of new entrants to institutes of technology receive a grant, well ahead of 36pc in the universities. Elsewhere, such as teacher-training colleges, the average was 41pc.

In the most extreme example, 71pc of students in Letterkenny Institute of Technology are on a grant, compared with just 24pc in Trinity and 28pc in UCD.

Grants are a good measure of third-level access across the social classes because eligibility is predominantly determined by an assessment of the income of students or their parents.

HEA chief executive Tom Boland said the data would help underpin a new National Strategy on Access to Higher Education.

Full reports: pages 10-11

Irish Independent, Monday, 9th November 2015

Text: Newspaper (Print)

HLT8

THE big class divide in third-level education is exposed in new figures showing how students from better-off families take most of the places in the country's universities.

Expose … class divide in our universities

Katherine Donnelly
Education Editor

THE big class divide in Irish third-level education is exposed in new figures showing how students from better-off families capture most of the places in the country's universities.

It means they are filling the lion's share of the most sought-after honours degree courses, including elite high-points programmes.

Katherine Donnelly
Barriers to equal access to third-level still exist
Analysis, page 10

The first-ever breakdown of the proportion of first-year student grant-holders in each college highlights a wide disparity between the universities and the institutes of technology.

Figures compiled by the Higher Education Authority (HEA), based on data supplied by the student grant agency, SUSI, show that, overall, 46pc of first-years in 2013-14 received a maintenance grant to help cover their living costs.

However, the breakdown by sector reveals 56pc of new entrants to institutes of technology receive a grant, well ahead of 36pc in the universities. Elsewhere, such as teacher-training colleges, the average was 41pc.

In the most extreme example, 71pc of students in Letterkenny Institute of Technology are on a grant, compared with just 24pc in Trinity and 28pc in UCD.

Grants are a good measure of third-level access across the social classes because eligibility is predominantly determined by an assessment of the income of students or their parents.

HEA chief executive Tom Boland said the data would help underpin a new National Strategy on Access to Higher Education.

Full reports: pages 10-11

Irish Independent, Monday, 9th November 2015

UCD DUBLIN

Text: Newspaper (Print)

HLT8

Exposed: the huge class divide in our universities

Katherine Donnelly
Education Editor

THE big class divide in Irish third-level education is exposed in new figures showing how students from better-off families capture most of the places in the country's universities.

It means they are filling the lion's share of the most sought-after honours degree courses, including elite high-points programmes.

Katherine Donnelly
Barriers to equal access to third-level still exist
Analysis, page 10

The first-ever breakdown of the proportion of first-year student grant-holders in each college highlights a wide disparity between the universities and the institutes of technology.

Figures compiled by the Higher Education Authority (HEA), based on data supplied by the student grant agency, SUSI, show that, overall, 46pc of first-years in 2013-14 received a maintenance grant to help cover their living costs.

However, the breakdown by sector reveals 56pc of new entrants to institutes of technology receive a grant, well ahead of 36pc in the universities. Elsewhere, such as teacher-training colleges, the average was 41pc.

In the most extreme example, 71pc of students in Letterkenny Institute of Technology are on a grant, compared with just 24pc in Trinity and 28pc in UCD.

Grants are a good measure of third-level access across the social classes because eligibility is predominantly determined by an assessment of the income of students or their parents.

HEA chief executive Tom Boland said the data would help underpin a new National Strategy on Access to Higher Education.

Full reports: pages 10-11

Irish Independent, Monday, 9th November 2015

Text: Newspaper (Print)

HLT8

the

Exposed: the huge class divide in our universities

Katherine Donnelly
Education Editor

THE big class divide in Irish third-level education is exposed in new figures showing how students from better-off families capture most of the places in the country's universities.

It means they are filling the lion's share of the most sought-after honours degree courses, including elite high-points programmes.

Katherine Donnelly
Barriers to equal access to third-level still exist
Analysis, page 10

The first-ever breakdown of the proportion of first-year student grant-holders in each college highlights a wide disparity between the universities and the institutes of technology.

Figures compiled by the Higher Education Authority (HEA), based on data supplied by the student grant agency, SUSI, show that, overall, 46pc of first-years in 2013-14 received a maintenance grant to help cover their living costs.

However, the breakdown by sector reveals 56pc of new entrants to institutes of technology receive a grant, well ahead of 36pc in the universities. Elsewhere, such as teacher-training colleges, the average was 41pc.

In the most extreme example, 71pc of students in Letterkenny Institute of Technology are on a grant, compared with just 24pc in Trinity and 28pc in UCD.

Grants are a good measure of third-level access across the social classes because eligibility is predominantly determined by an assessment of the income of students or their parents.

HEA chief executive Tom Boland said the data would help underpin a new National Strategy on Access to Higher Education.

Full reports: pages 10-11

Irish Independent, Monday, 9th November 2015

UCD DUBLIN

Text: File (Electronic)

HLT8

Exposed: the huge class divide in our universities

The big class divide in third-level education is exposed in new figures showing how students from better-off families take most of the places in the country's universities.

New data shows teenagers from wealthier homes are filling the lion's share of the most sought-after honours degree courses, including elite high-points programmes.

The first ever breakdown of the proportion of first year students in receipt of grants in each college, highlights a wide disparity between the universities and the institutes of technology. The figures will now be used help inform a new strategy to level the playing pitch for access to higher education.

Grants are a good measure of third-level access across the social classes, because eligibility is predominantly determined by an assessment of the income of students or their parents. …

Irish Independent, Monday, 9th November 2015

UCD DUBLIN

Text … Web Pages (Electronic)

HLT8

http://www.independent.ie/irish-news/education/exposed-the-huge-class-divide-in-our-universities-34181536.html

Exposed: the huge class divide in our universities

Katherine Donnelly
EMAIL

PUBLISHED
09/11/2015 | 02:30

SHARE

Just 24pc of first-year full-time students at TCD are in receipt of student grants

The big class divide in third-level education is exposed in new figures showing how students from better-off families take most of the places in the country's universities

New data shows teenagers from wealthier homes are filling the lion's share of the most sought-after honours degree courses, including elite high-points programmes.

The first ever breakdown of the proportion of first year students in receipt of grants in each college, highlights a wide disparity between the universities and the institutes of technology. The figures will now be used help inform a new strategy to level the playing pitch for access to higher education.

Grants are a good measure of third-level access across the social classes, because eligibility is predominantly determined by an assessment of the income of students or their parents.

Grant eligibility income thresholds range from about €40,000 to €65,000 a year, and factors such as the number of dependant children and how many are attending third level also come into play.

Depending on circumstances, grants are worth up to €6,000 a year. Figures compiled by the Higher Education Authority (HEA), based on data supplied by the student grant agency, SUSI, for 2013/14, show that, overall, 46pc of first years received a maintenance grant to help cover their living costs.

However, the breakdown by sector reveals 56pc of new entrants to institutes of technology receive a grant, well ahead of 36pc in the universities. Elsewhere, such as teacher training colleges, the average was 41pc.

Text … Web Pages (Electronic)

HLT8

http://www.independent.ie/irish-news/education/exposed-the-huge-class-divide-in-our-universities-34181536.html

Exposed: the huge class divide in our universities

Katherine Donnelly
EMAIL

PUBLISHED
09/11/2015 | 02:30

SHARE

Just 24pc of first-year full-time students at TCD are in receipt of student grants

The big class divide in third-level education is exposed in new figures showing how students from better-off families take most of the places in the country's universities

New data shows teenagers from wealthier homes are filling the lion's share of the most sought-after honours degree courses, including elite high-points programmes.

The first ever breakdown of the proportion of first year students in receipt of grants in each college, highlights a wide disparity between the universities and the institutes of technology. The figures will now be used help inform a new strategy to level the playing pitch for access to higher education.

Grants are a good measure of third-level access across the social classes, because eligibility is predominantly determined by an assessment of the income of students or their parents.

Grant eligibility income thresholds range from about €40,000 to €65,000 a year, and factors such as the number of dependant children and how many are attending third level also come into play.

Depending on circumstances, grants are worth up to €6,000 a year. Figures compiled by the Higher Education Authority (HEA), based on data supplied by the student grant agency, SUSI, for 2013/14, show that, overall, 46pc of first years received a maintenance grant to help cover their living costs.

However, the breakdown by sector reveals 56pc of new entrants to institutes of technology receive a grant, well ahead of 36pc in the universities. Elsewhere, such as teacher training colleges, the average was 41pc.

Text … Web Pages (Electronic)

HLT8

http://www.independent.ie/irish-news/education/exposed-the-huge-class-divide-in-our-universities-34181536.html

Exposed: the huge class divide in our universities

Katherine Donnelly
EMAIL

PUBLISHED
09/11/2015 | 02:30

SHARE

Just 24pc of first-year full-time students at TCD are in receipt of student grants

The big class divide in third-level education is exposed in new figures showing how students from better-off families take most of the places in the country's universities

New data shows teenagers from wealthier homes are filling the lion's share of the most sought-after honours degree courses, including elite high-points programmes.

The first ever breakdown of the proportion of first year students in receipt of grants in each college, highlights a wide disparity between the universities and the institutes of technology. The figures will now be used help inform a new strategy to level the playing pitch for access to higher education.

Grants are a good measure of third-level access across the social classes, because eligibility is predominantly determined by an assessment of the income of students or their parents.

Grant eligibility income thresholds range from about €40,000 to €65,000 a year, and factors such as the number of dependant children and how many are attending third level also come into play.

Depending on circumstances, grants are worth up to €6,000 a year. Figures compiled by the Higher Education Authority (HEA), based on data supplied by the student grant agency, SUSI, for 2013/14, show that, overall, 46pc of first years received a maintenance grant to help cover their living costs.

However, the breakdown by sector reveals 56pc of new entrants to institutes of technology receive a grant, well ahead of 36pc in the universities. Elsewhere, such as teacher training colleges, the average was 41pc.

Processing Text

Depending on the source of the text there may be various formatting and content (e.g. document headers and separators, typesetter codes, tables and diagrams) that is just "junk" that needs to be filtered out.

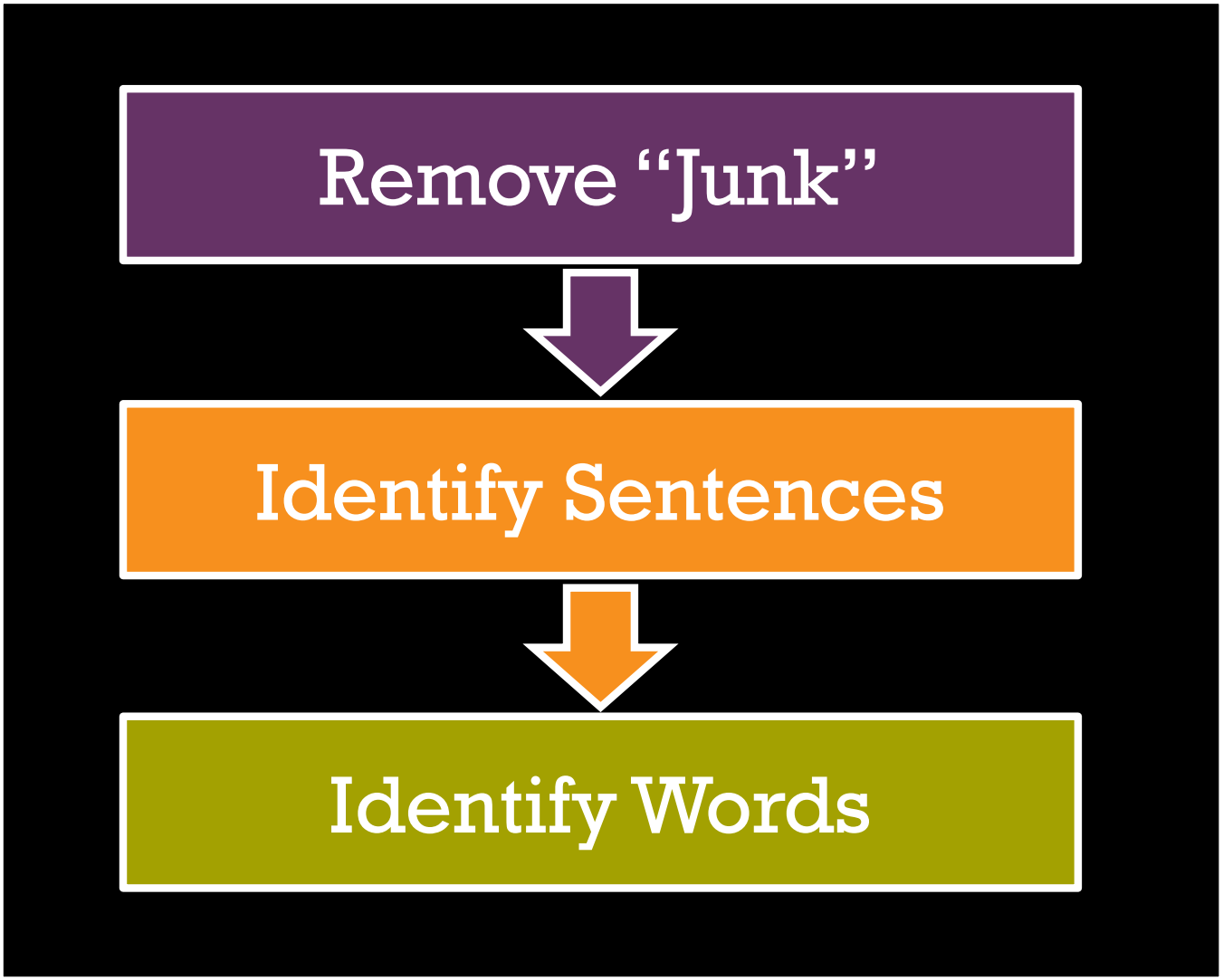
Uppercase vs. Lowercase:

- e.g. THE, The, the

Convert all uppercase to lowercase but:

- Richard Brown vs brown; Institute of Technology vs. technology → proper names
- HEA vs. hea, UCD vs. ucd

Processing Text



Identifying Words

HLT8

What is a word?

- Graphic word (Kuçera & Francis, 1967): "a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks".
  - But cf. the £6.3bn project
- The main clue used in English is the occurrence of white space - a space or tab or the beginning of a new line between words
  - But cf. 42pc

→ not completely reliable

UCD DUBLIN

Identifying Words

What is a word?

- Words are not always surrounded by white space
- Often punctuation marks such as commas, semicolons and full stops attach to words
  - E.g. universities., agency,
- Full stops not just at the end of a sentence
  - E.g. etc., 8.30 a.m.
  - If etc. appears at the end of a sentence, then only one . occurs → haplology
- Contractions such as *I'll*, *isn't* count as one graphic word according to the definition but intuitively there are 2 words in each case.

Identifying Words

HLT8

What is a word?

- Normally orthographic-word-final single apostrophes represent the end of a quotation and so should not be part of a word
  - but after **s,** they may be plural possessive
  - E.g. The students' laptops…
- Sequences with a hyphen may count as one word or two, or indicate correct groupings
  - E.g. better-off, sought-after, 2013-14, teacher-training, first-ever
  - E.g. teacher-training colleges, first-year student grant-holders, first-ever breakdown

UCD DUBLIN

Identifying Words

HLT8

What is a word?

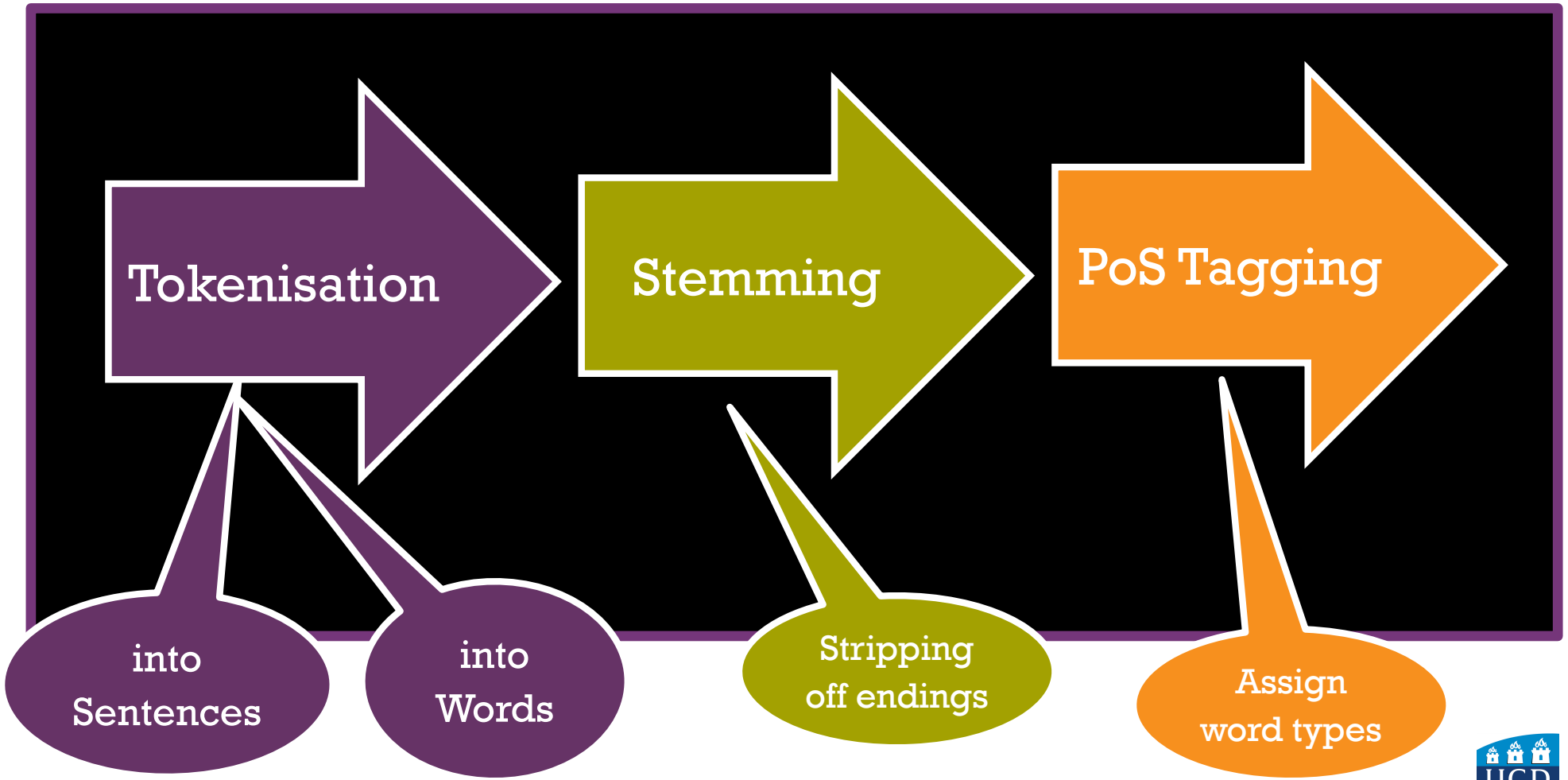
- Same form representing multiple 'words'?:
  - Do we want to regard variant sequences of characters as really the same word?
  - Homographs E.g. saw (V) vs. saw (N)

- Apparently different 'words'?:
  - E,g. 46pc, 56pc, 24pc, 28pc

- Note of course that other languages have different problems identifying words (e.g. those with a more complicated morphology or no white space between words)

UCD DUBLIN

Word and Sentence Identification



Tokenisation

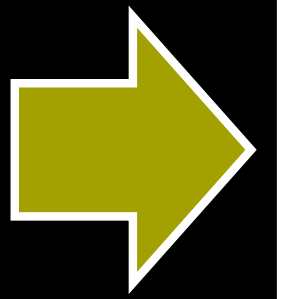
Divide the input text into units called tokens where each is either a word or something else like a number or punctuation mark.

The treatment of punctuation varies:

- keep only sentence boundaries
- keep sentence-internal punctuation for disambiguation → commas and dashes give some clues about the macro structure of the text.

Tokenisation

THE big class divide in third-level education is exposed in new figures showing how students from better-off families take most of the places in the country's universities.



THE	students
big	from
class	better-off
divide	families
in	take
third-level	most
education	of
is	the
exposed	places
in	in
new	the
figures	country's
showing	universities.
how	

Stemming and Lemmatisation

- Should **shows, showed, showing** be treated as individual words or should they be collapsed into a single lexeme?

Stemming :
- stripping off affixes to leave a stem e.g. lie from lies

Lemmatisation:
- attempting to find the lemma (or lexeme) to which the word belongs e.g. lying as a realisation of lie

- Note that a full-form lexicon for languages with a richer morphology than English would be too large

Stemming and Lemmatisation

- Should **shows, showed, showing** be treated as individual words or should they be collapsed into a single lexeme?

But cf. definition from morphology

Stemming :
- stripping off affixes to leave a stem e.g. lie from lies

Lemmatisation:
- attempting to find the lemma (or lexeme) to which the word belongs e.g. lying as a realisation of lie

- Note that a full-form lexicon for languages with a richer morphology than English would be too large

UCD DUBLIN

Stemming

E.g. Porter Stemmer

the orange rabbit flies rapidly through the night

fli rapidli Stems

not ***fly (+s)*** as we might expect

not ***rapid (+ly)*** as we might expect

Stemming

E.g. Porter Stemmer

the orange rabbit flies rapidly through the night

fli rapidli — Stems

E.g. Porter Stemmer

the women chased the children around the park

wom chas childr — Stems

Lemmatisation

Lemmatiser

the orange rabbit flies rapidly through the night

fly rapid

Lemmas

Lemmatiser

the women chased the children around the park

woman chase child

Lemmas

Parts of Speech

Categories

the orange rabbit flies rapidly through the night

DET	N ADJ	N V	N V	ADV	PREP	DET	N

cf. Parsing

PoS Tagger

the orange rabbit flies rapidly through the night

DT	JJ	NN	VBZ	RB	IN	DT	NN

Tags (→ Tag Set)

PoS Tag Sets

Standardly a tag set encodes:

- the target feature of classification, telling the user the useful information about the grammatical class of the word
- the predictive features, encoding features that will be useful in predicting the behaviour of others words in the context

Parts of Speech can be motivated by different aspects:

- semantic (notional)
- syntactic distributional
- morphological

Working with Real Text

HLT8

→ See TextProcessing2019.pdf

UCD DUBLIN