Slipping through the cracks of e-lexicography: lessons from

Ana Frankenberg-Garcia

Geraint Rees

Robert Lew

Jonathan C. Roberts

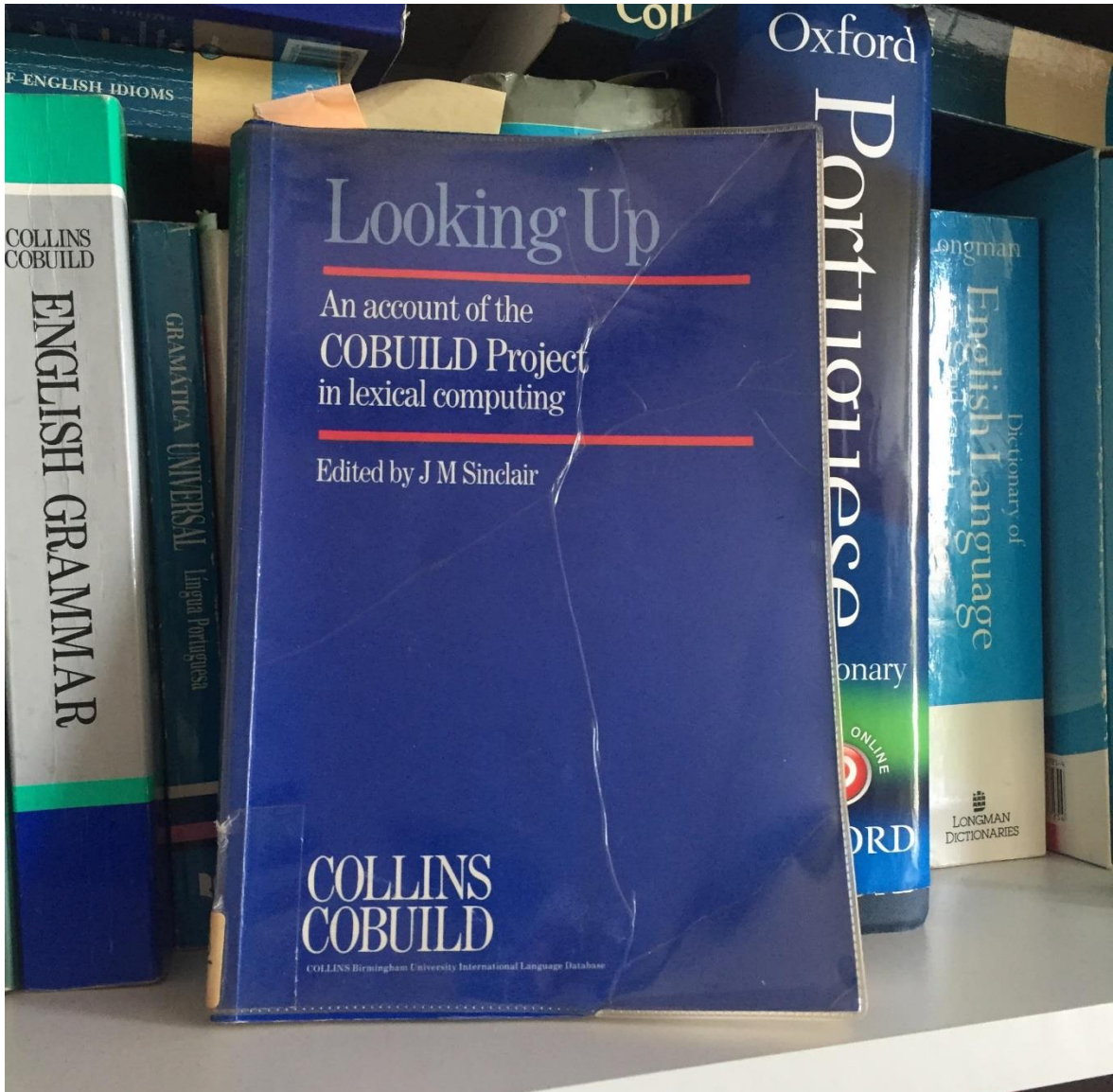
Nirwan Sharma

Peter Butcher

1987

First developments in e-lexicography

- COBUILD project (Sinclair 1987)
- Lexical analyses aided by corpora
  - Concordances
  - Word lists
  - Collocations



Things have evolved…

Word Sketches

- Lexical summary of how words are used with other words (Kilgarriff et al. 2004, 2014)

logDice

- Better statistics for corpus-based lexicography (Rychlý 2008, Gablasova et al. 2017, Frankenberg-Garcia 2018)

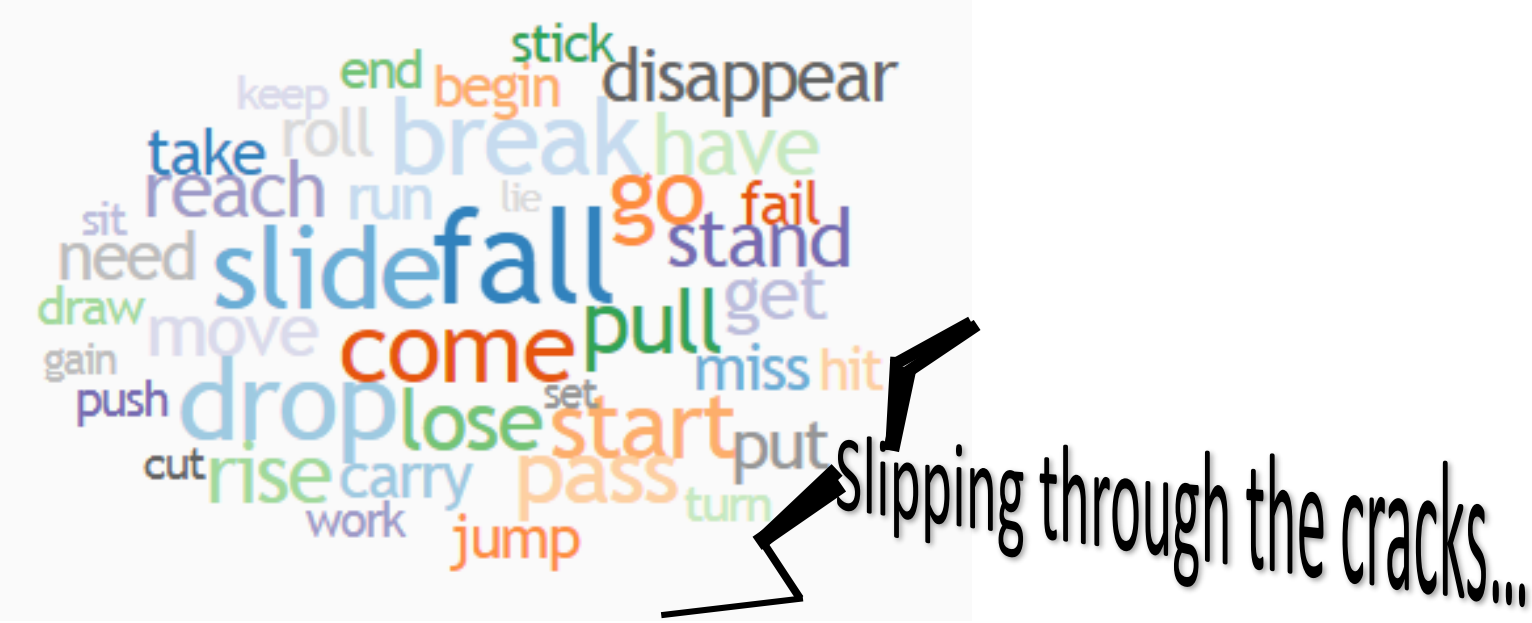
GDEX

- Easier to find good dictionary examples from corpora (Kilgarriff et al. 2008)

e-Lexicography today

- Headword lists & defining vocabularies adjusted to better reflect language use
- Definitions enhanced with linguistic data beyond introspection
- Improved coverage of syntactic patterns, lexical collocations & phraseology
- Abundancy of authentic examples of words in context

And yet…

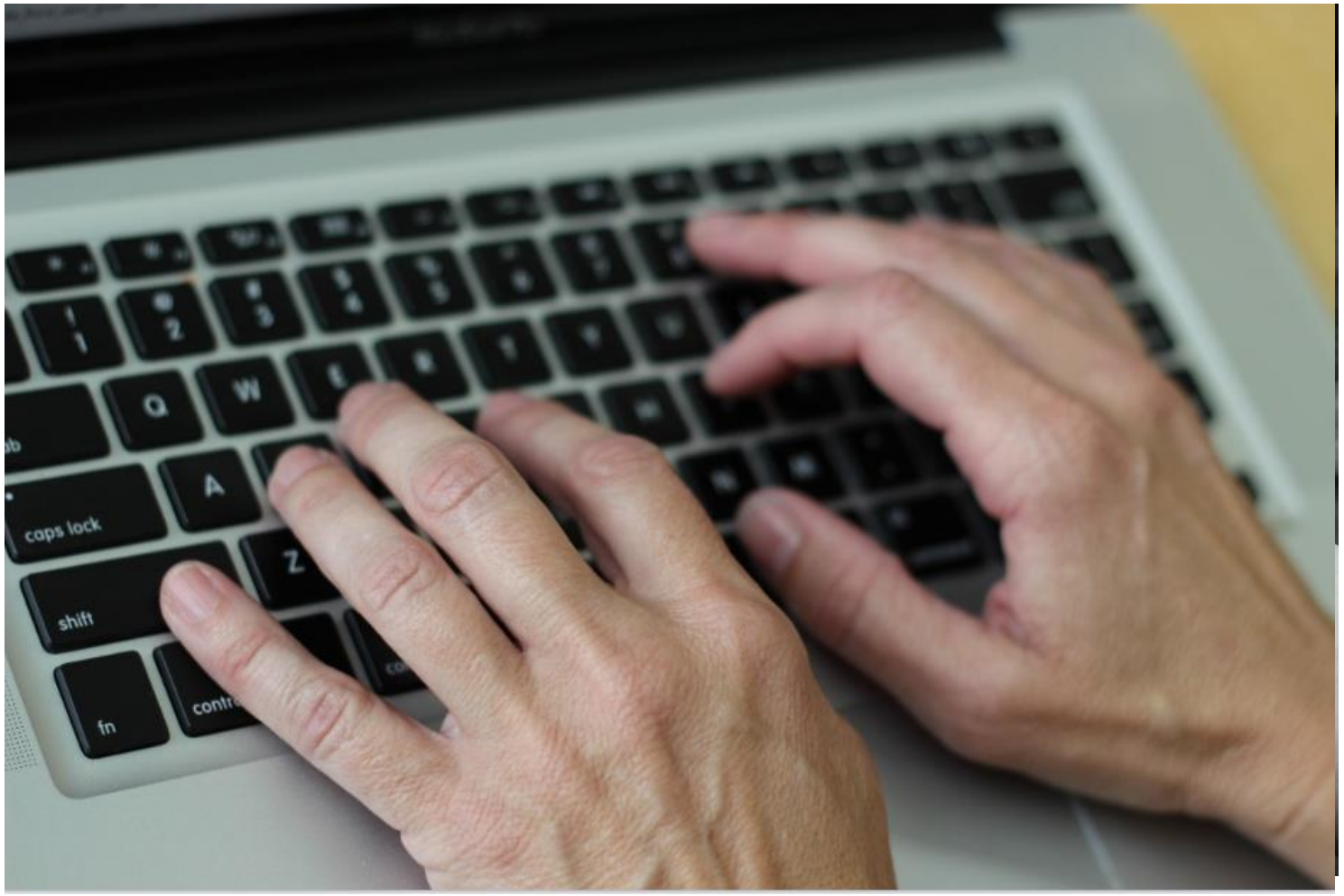


- When compiling the lexical database for ColloCaid, we noted…
- Problems that seem to be slipping through the cracks of e-lexicography

We present next lessons learnt and some examples

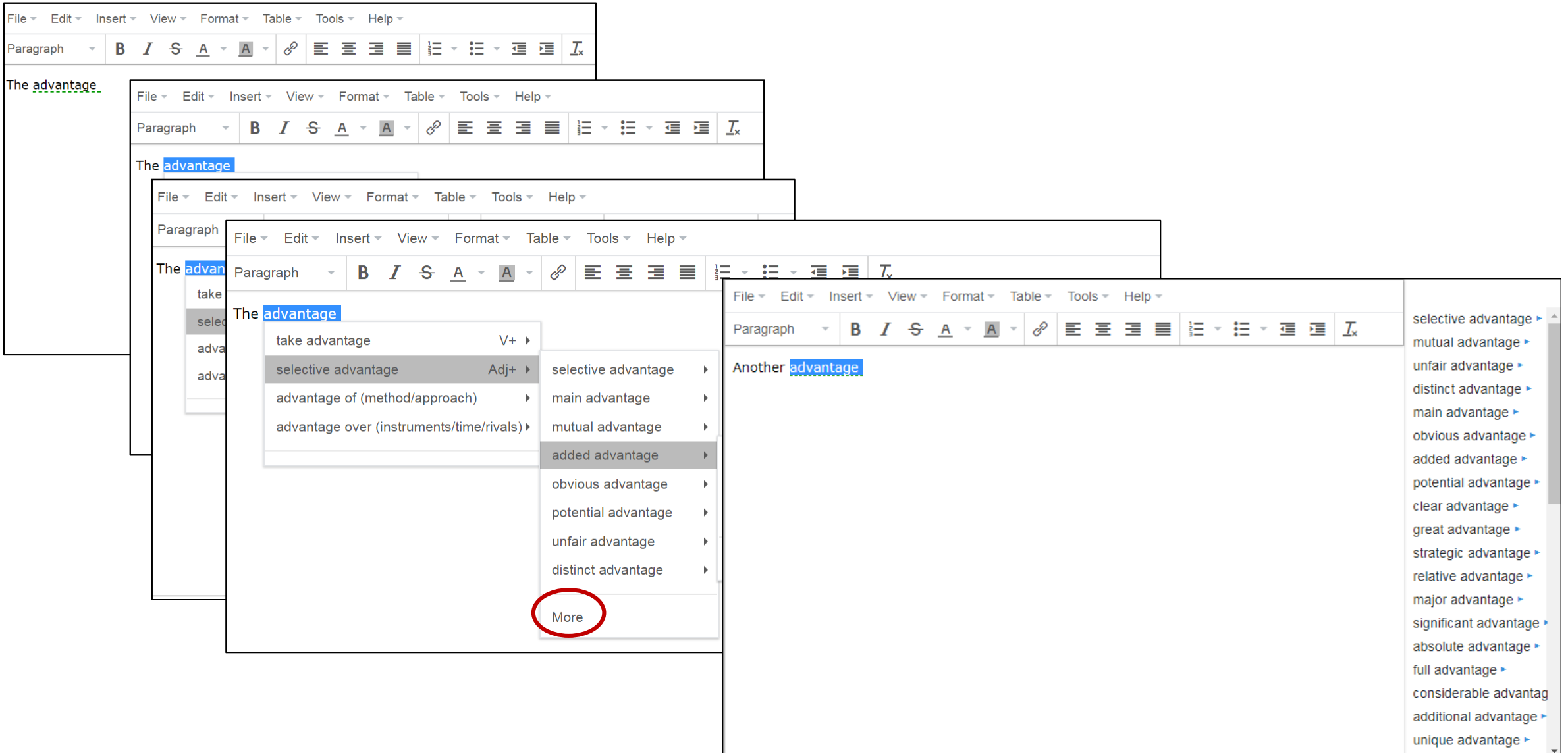
What is ColloCaid?

Academic writing is hard…



ColloCaid is a text editor that assists writers with academic English collocations

How does ColloCaid work?



What's different about ColloCaid?

- Can help writers expand their academic collocation repertoire
  - Real-time collocation reminders
  - Writers can access collocations they might otherwise not remember to use
- Interactive text editor integration
  - Writers don't have to stop writing
  - Collocations suggestions are easy to find
  - But only given when sought, so as not to disrupt writing
- Data-driven learning (Johns 1991)
  - Suggestions shown, not explained
  - Metalanguage kept to a minimum
- Curated lexical data, so that users don't get distracted with
  - Irrelevant or misleading information
  - An overload of information

What lessons did we learn?

Lessons from ColloCaid

1. Issues with node selection
2. Issues with collocation research
3. Issues with corpus examples

Node selection

Node selection

( ) *Where in my text can I fit in conduct research?*

(x) *What verb can I use with research?*

- Starting point is the node, not the collocation
- 500 maximally useful collocation nodes for general academic English
  - 36-month project
- Nouns, verbs and adjectives from 3 well-established word lists of general academic English
  - Extracted using different methods and different corpora
- Build on the strengths of each list

Node selection

1. AKL - Academic Keyword List (Paquot 2010)

- Emphasis on academic keyness

2. AVL-BAWE - Subset of Academic Vocabulary List (Gardner & Davies 2014) that overlaps with BAWE word list (Durrant 2016)

- Emphasis on novice academic writing vocabulary

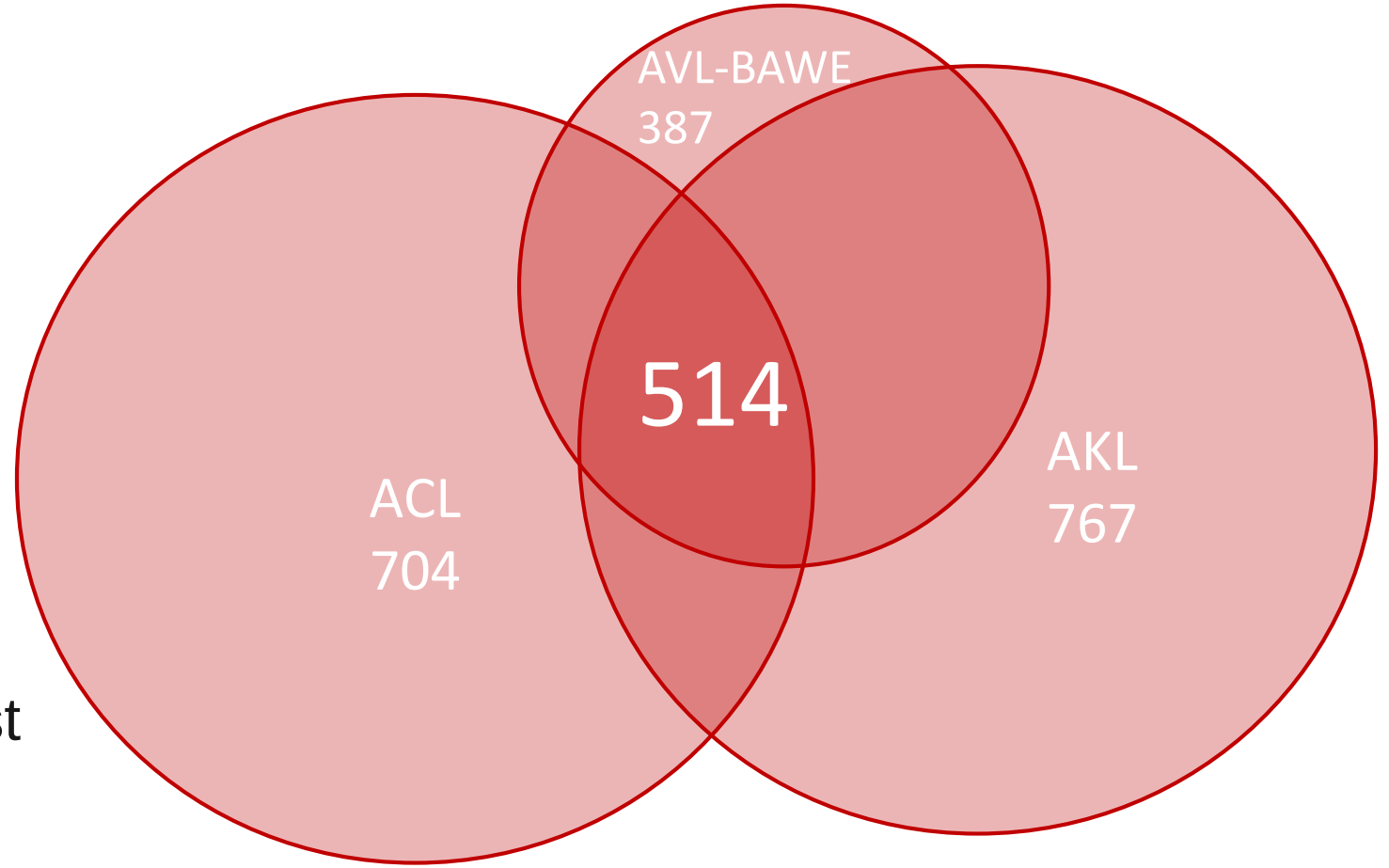
3. ACL - Academic Collocations List (Ackermann & Chen 2013), using headwords in appendix to Longman Collocations Dictionary (Mayor 2013)

- e.g. for *conduct research*, headword = *research*
- Emphasis on strong collocations

Node selection

Word list intersections

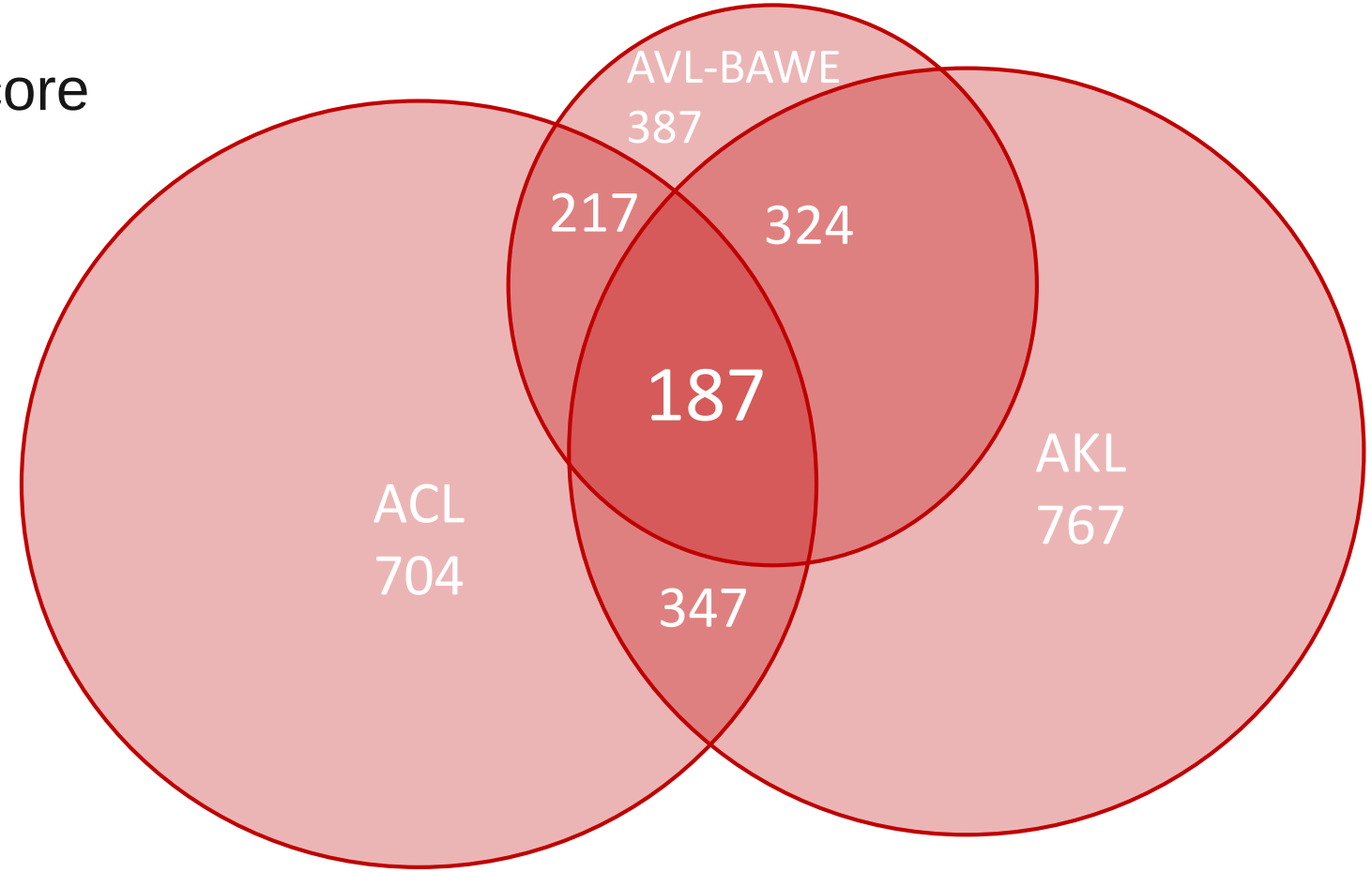
- AKL (biggest, academic keyness)
- AVL-BAWE (smallest, students)
- ACL (biggish, strong collocations)



- 514 lemmas overlapping in at least two lists
- Perfect fit for project scope ☺

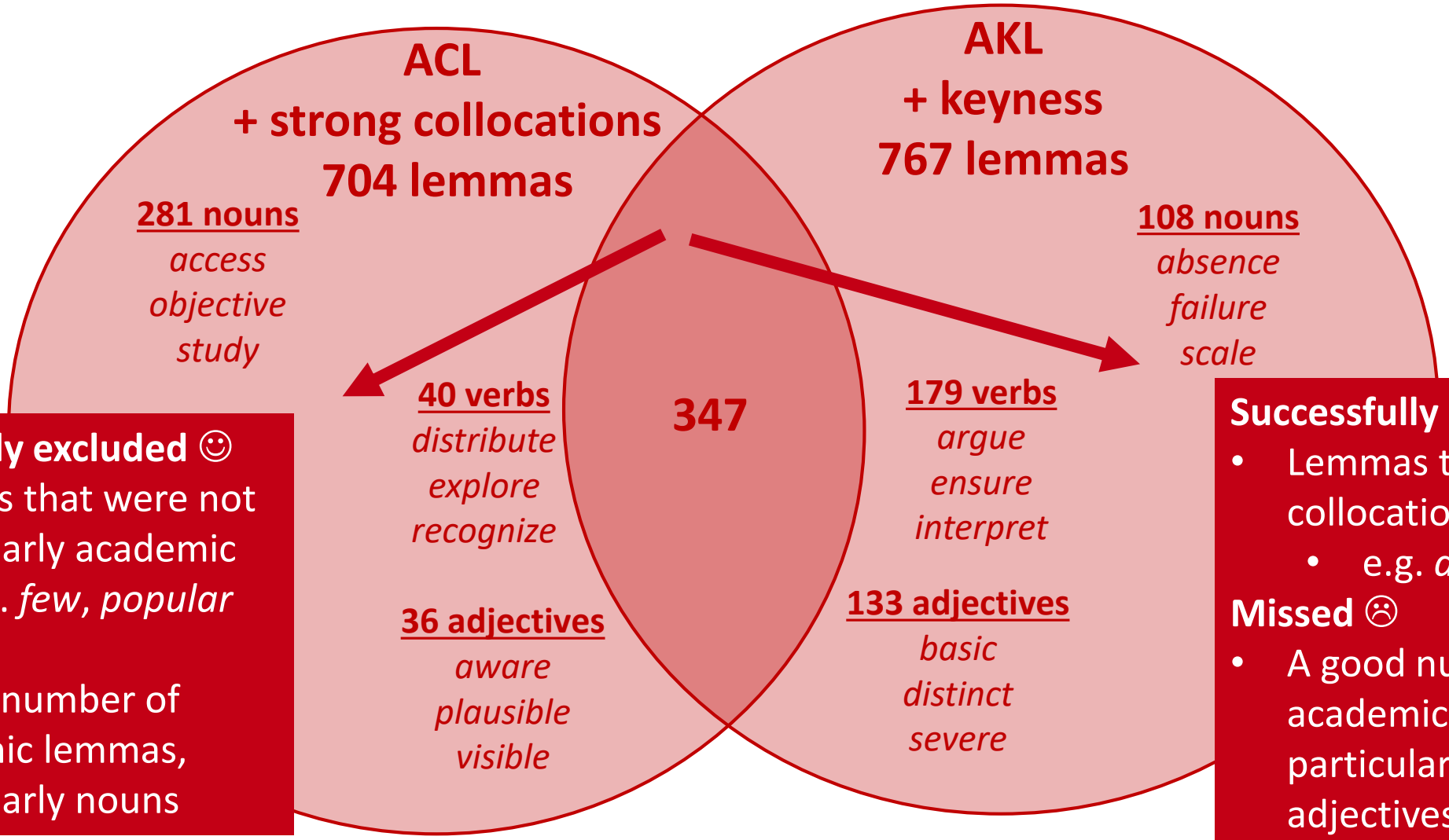
Node selection issues

- Even though all 3 lists are about core general academic English vocabulary…
- They overlap less than expected!



Words lists aiming for similar vocabulary coverage (i.e. core academic vocabulary) can differ substantially depending on corpora and extraction methods used

Node selection issues



Successfully excluded ☺
- Lemmas that were not particularly academic
  - e.g. *few*, *popular*

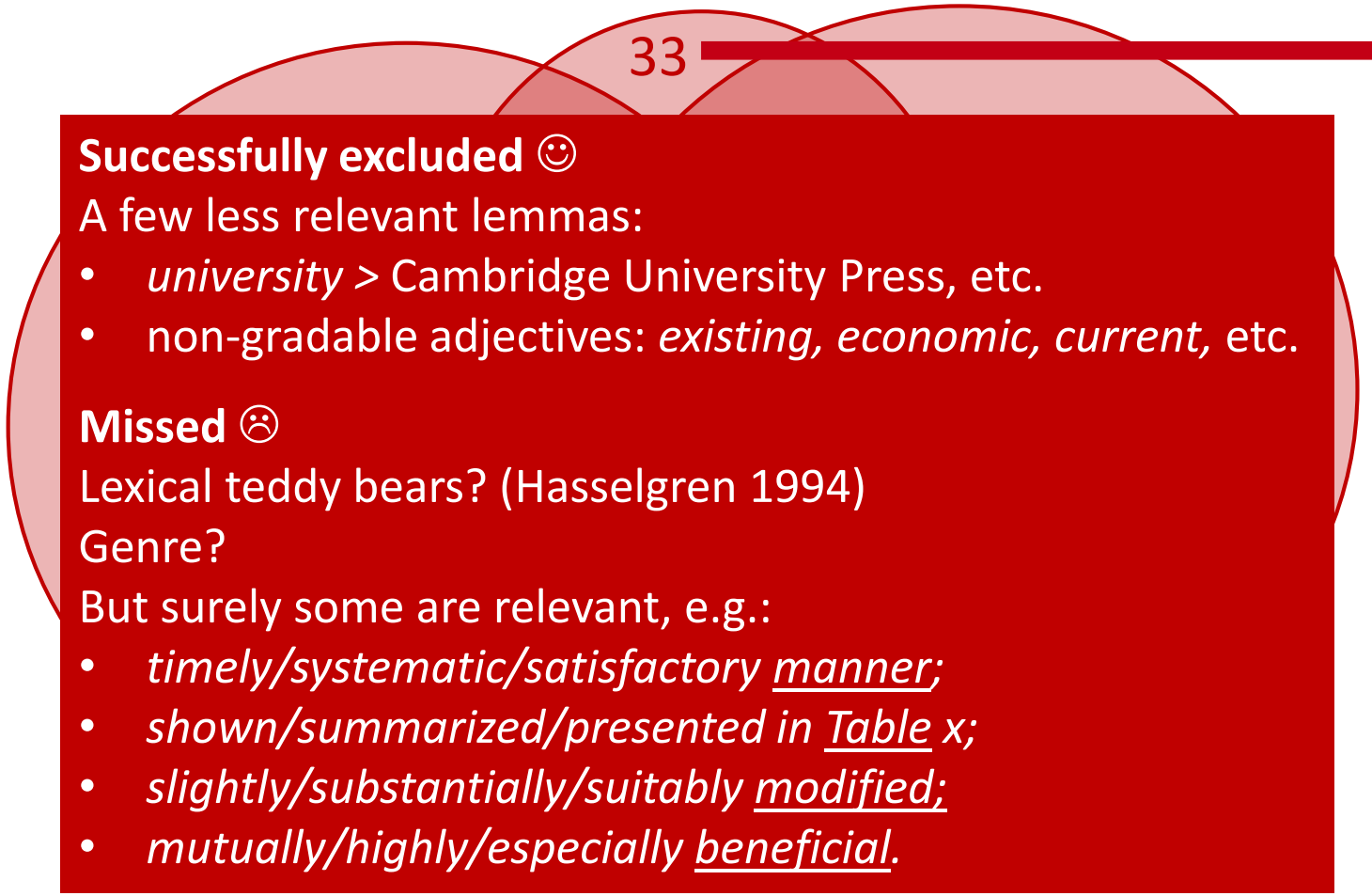
Missed ☹
- A good number of academic lemmas, particularly nouns

Successfully excluded ☺
- Lemmas that were not collocationally productive
  - e.g. *actual*, *prime*

Missed ☹
- A good number of academic lemmas, particularly verbs & adjectives

Node selection issues

AVL-BAWE
+ novice academic
387 lemmas



33

Successfully excluded ☺
A few less relevant lemmas:
- *university* > Cambridge University Press, etc.
- non-gradable adjectives: *existing, economic, current,* etc.

Missed ☹
Lexical teddy bears? (Hasselgren 1994)
Genre?
But surely some are relevant, e.g.:
- *timely/systematic/satisfactory manner;*
- *shown/summarized/presented in Table x;*
- *slightly/substantially/suitably modified;*
- *mutually/highly/especially beneficial.*

9 nouns
appendix
century
flow
inclusion
manner
project
region
table
university

9 verbs
address
calculate
estimate
justify
modify
observe
range
situate
test

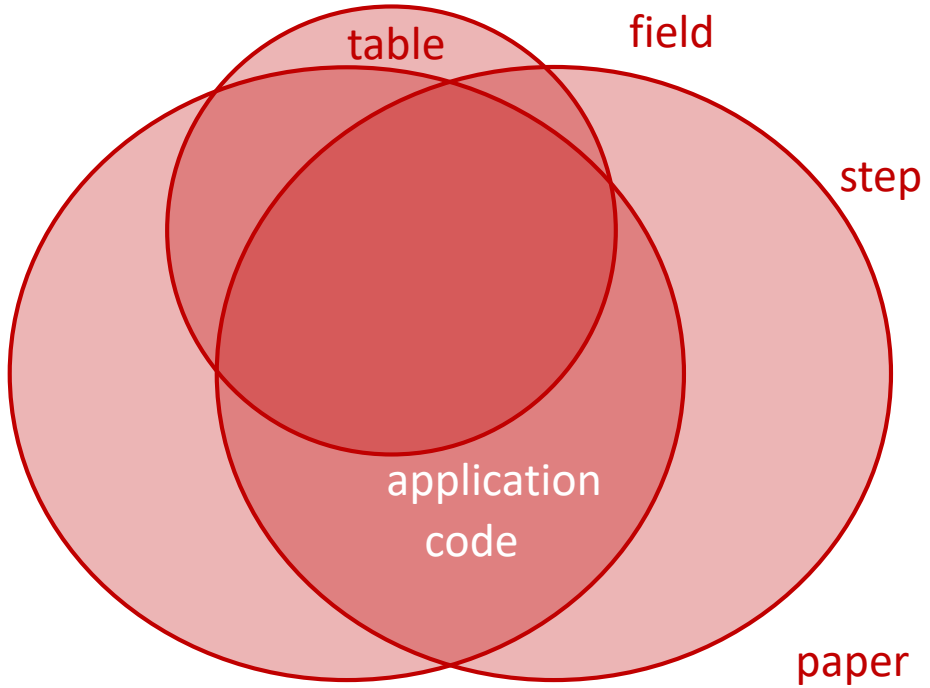
15 adjectives
above
accurate
beneficial
broad
continuous
current
detailed
economic
existing
external
given
global
increased
numerous
varying

Node selection issues

We also noted

1. Basic lemmas with major academic meanings were missing from well-known academic word lists

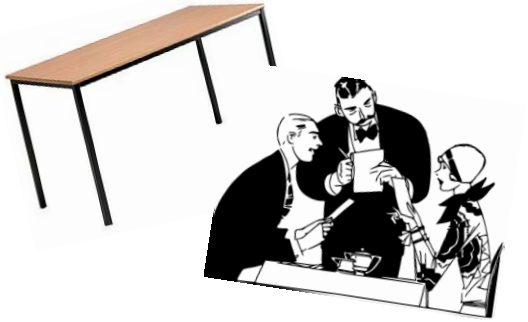
2. Less central lemmas appeared to be overrated in comparison



Node selection issues

Why were basic academic lemmas with major academic meanings missed?

- Words lists used disregard sense distinctions
- Less sensitive to academic senses of lemmas widely used in general language



Important academic senses get diluted and escape frequency thresholds

- Ideally, most obvious missing academic lemmas need to be added

Node selection issues

Why did less central academic lemmas seem overrated in comparison?

Two telling examples:

1. application

= request (*a successful patent application*)

= software (*IT applications require significant innovative effort*)

= use *(the application of statistical techniques to test assumptions)*

Lemmas with distinct general academic senses lumped together and jump rank queue

2. code (not interdisciplinary, but multidisciplinary)

In Computing

In Biology

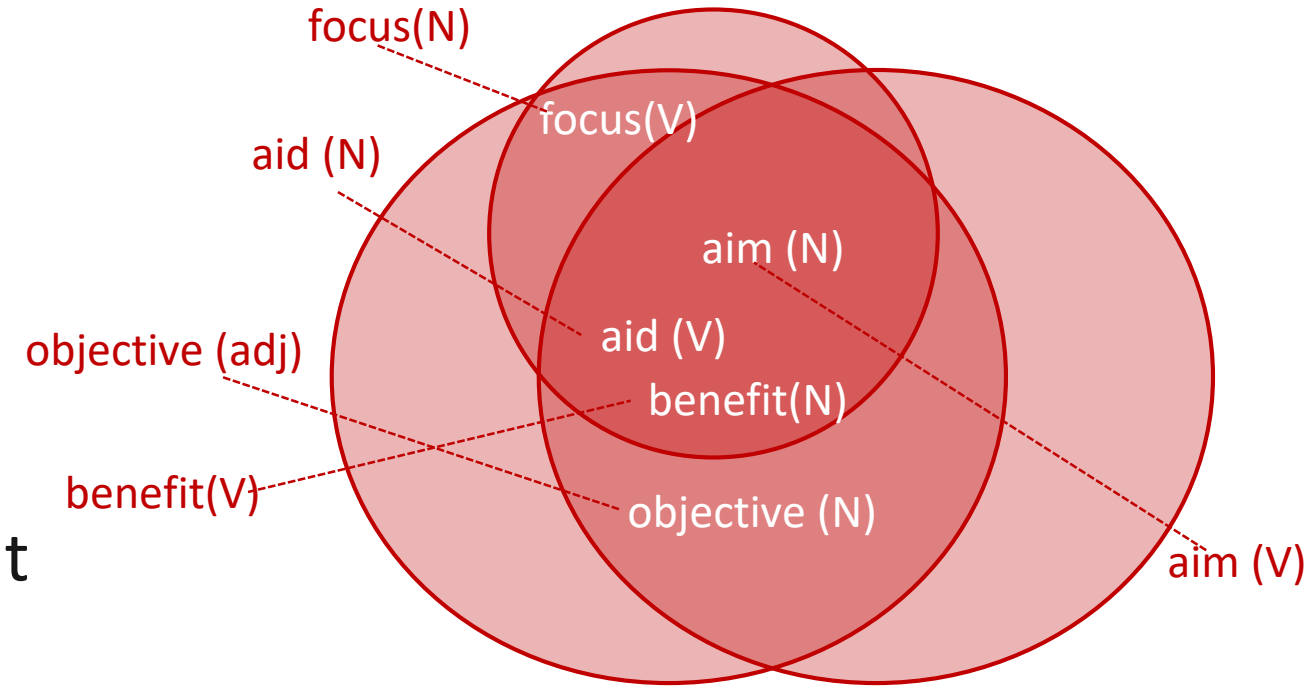
In Language and Linguistics

Terms with distinct discipline-specific senses lumped together and also jump rank queue

Node selection issues

Another issue was word class

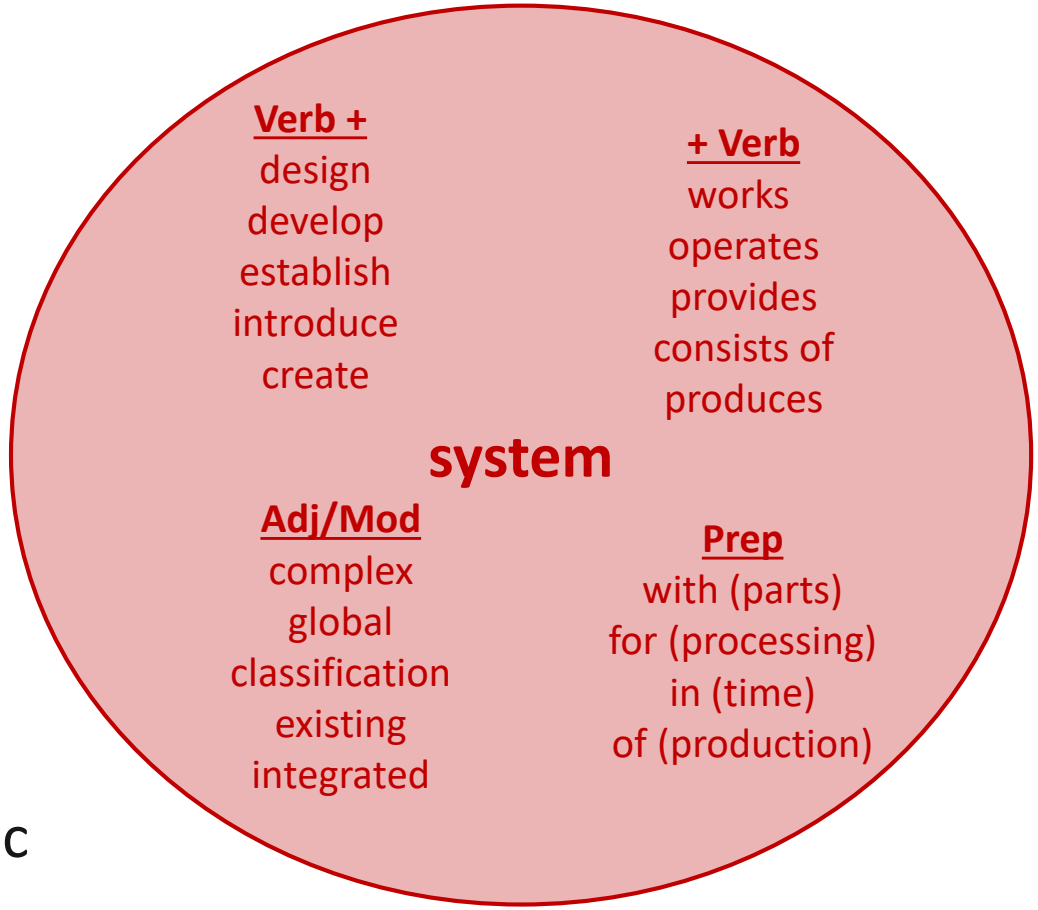
- Some homographs are naturally more frequent in one class than another
- But this can give the impression that lexical coverage is inconsistent
- Decision to include homographs
  - Even if they did not meet our thresholds



Collocation research

Collocation research

- Collocations likely to be looked up for each node
- Interdisciplinary lexical collocations
- Grammatical collocations (syntactic patterns)
- Expert academic English corpora
  - **Oxford Corpus of Academic English**
  - Pearson International Corpus of Academic English
  - COCA academic



Collocation research

Word sketches (Kilgarriff et al. 2004, 2014)

- Collocates sorted per grammar relation

  research + ADJ

  - logDice & co-occurrence thresholds
  - excluded discipline-specific words (dispersion)
  - excluded open-choice combinations
    - e.g. *more research*

qualitative qualitative research	1,522	10.51	•••
future for future research	1,203	9.89	•••
quantitative quantitative research	751	9.53	•••
further further research	873	9.22	•••
previous previous research	693	9.04	•••
empirical empirical research	523	8.91	•••
recent recent research	544	8.52	•••
market market research	472	8.38	•••

Collocation research issues

- Is anything missing?
- *carry out + research* (+5, -5) = 303!
- Space between *carry* and *out* is problematic
- Further examples
  - *set out*
  - *set up*
  - *at stake*
  - *at hand*

conduct	180	10.56	•••
undertake	193	10.42	•••
fund	24	8.08	•••
publish	33	7.69	•••
support	52	7.34	•••
pursue	17	7.19	•••
cite	12	6.77	•••
sponsor	8	6.76	•••
stem	8	6.76	•••
stimulate	9	6.6	•••
summarize	7	6.51	•••

Verbs used with *research* as object in PICAE

Collocation research issues

Another problem…

- Word sketches sort collocates according to word class
- But collocation paradigms don't always evoke a single POS
- For example:

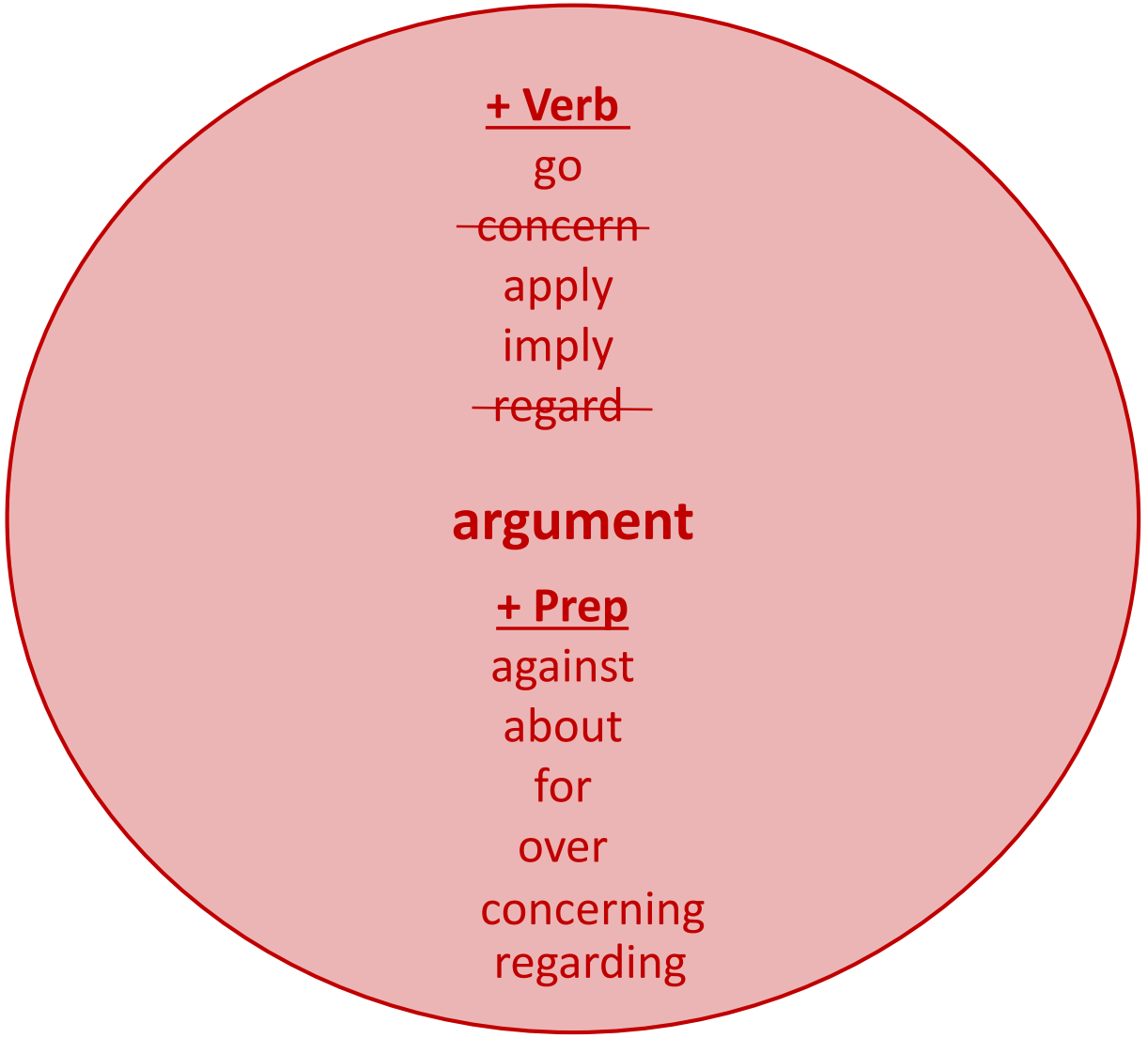
  *Their **argument about** film and dream*

  *Their **argument over** film and dream*

  *Their **argument concern**ing film and dream*

  *Their **argument regard**ing film and dream*

- Decision to move them



Collocation research issues

1. attitude + to

- *Positive attitudes to technology*
- ~~*Convey the attitude to the student that…*~~
- *Convey the attitude to the student that…*

2. community + develop

- *How communities develop through time*
- ~~*Connectedness between the community develops*~~
- *Connectedness between the community develops*

Similar on surface, but grammar relation differs

Collocate misidentified (distorting logDice & co-occurrence figures)

Collocation research issues

1. Late (ADJ) + development

- *the late development of passives*
- *the latest technological developments in the field*

2. Increase (ADJ) + demand

- *A momentarily increased demand for…*
- *To satisfy increasing energy demands …*

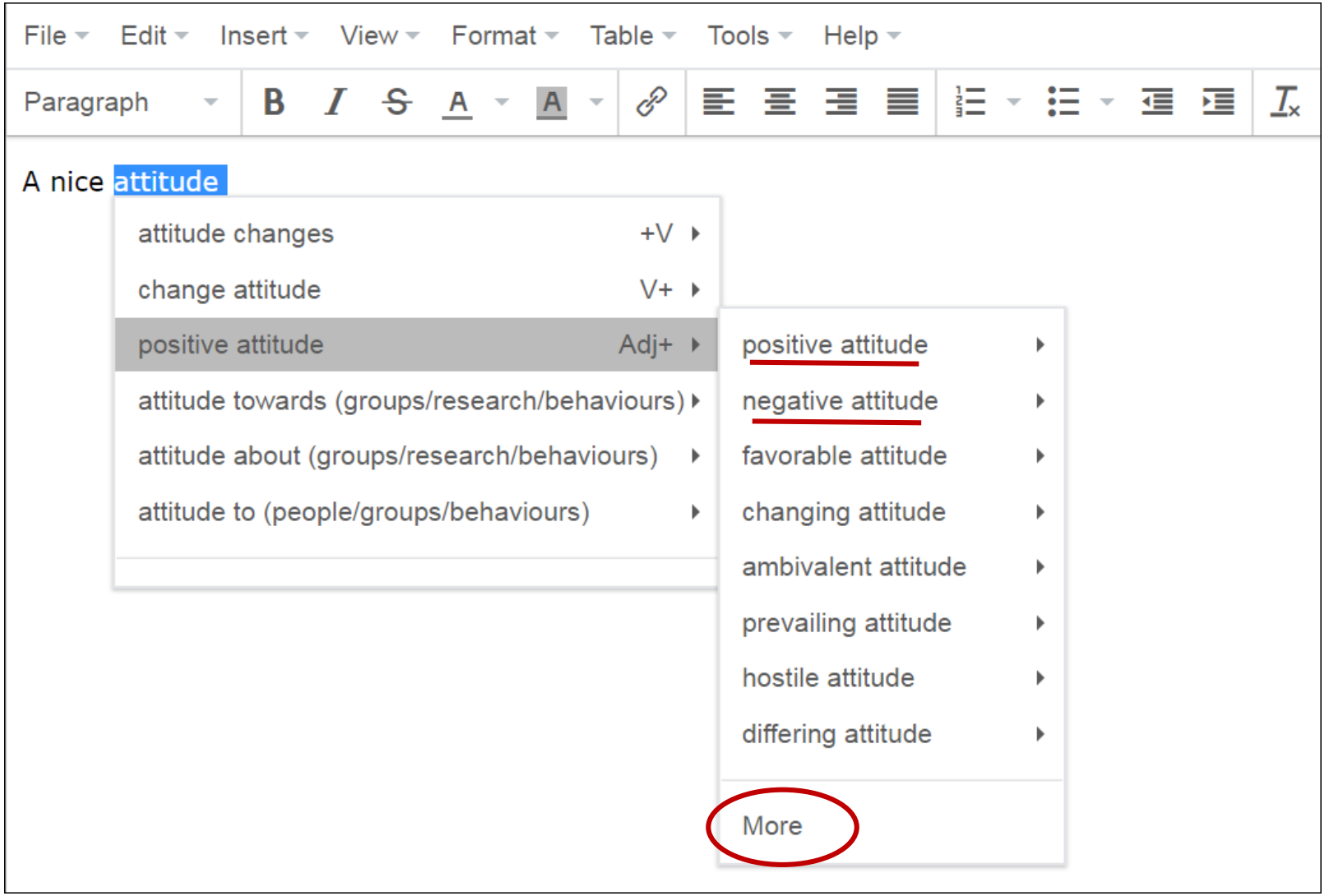
Similar on the surface, but semantically different

- Affects logDice & co-occurrence
- Polysemous collocates get bumped up the queue

Collocation research issues

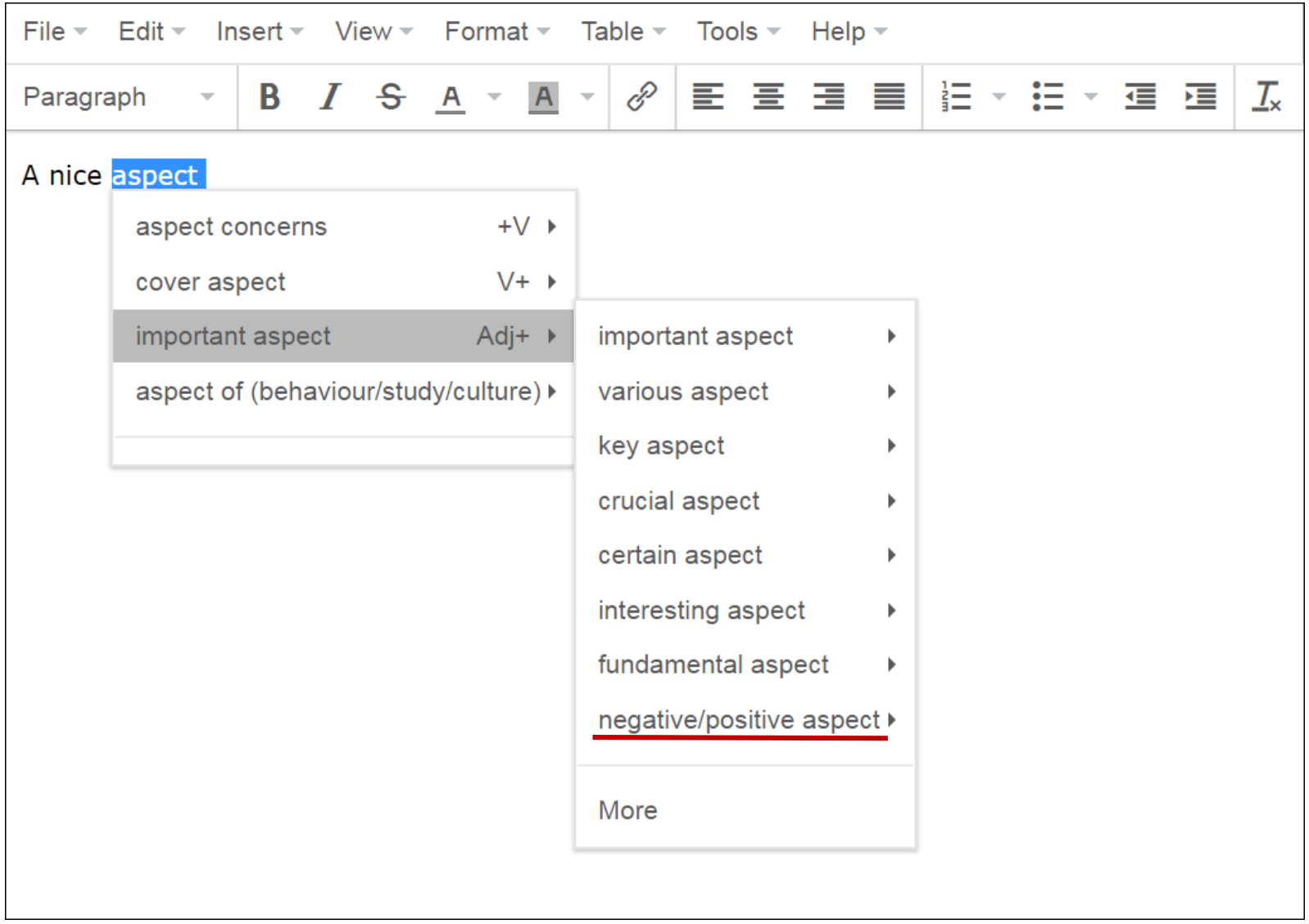
To avoid overwhelming writers with too many collocations…

At this level, we display only the top eight collocates
(sorted by logDice)



Collocation research issues

However, manual curation was necessary in certain cases, to avoid impression of inconsistent coverage



Example selection

Example selection

1. Examples selected from corpora of expert academic writing
   - From word sketch to concordance
   - With editorial intervention (to select, adapt, anonymize, shorten)

2. Are interdisciplinary or transferable to other disciplines

3. Promote data-driven learning (Johns 1991)
   - Three examples (Frankenberg-Garcia 2014)
   - Colligation cues if relevant/possible

4. Brief: less effort to process, less distracting for writers
   - GDEX (Kilgarriff et al. 2008)

Example selection issues

V obj N*
undertake ••• activities undertaken
coordinate •••
regulate •••
perform •••
organize •••
inhibit •••
monitor •••
increase •••
stimulate •••

<s> But by a clever construction of notional prices (called ' shadow prices ' ; Chapters 7-8) , economists have adapted GDP even for economies like Desta 's , where much economic **activity** is **undertaken** in non-market institutions . </s>

<s> Because people can't insure themselves sufficiently against failure , they are reluctant to **undertake** **activities** offering a chance of huge success if there is also an accompanying chance of large failure . </s>

<s> Once all of the goals and indicators for success (for the collaboration , for the school , for the community , for the parents , and for the students) have been identified and it is clear what has to happen first , interventions (**activities** **undertaken** in the community school , such as after-school programs , health services , or parent workshops) are added wherever needed to reach an outcome . </s>

<s> In general , online users showed a greater predilection to expand the set of online **activities** **undertaken** in a year 's time . </s>

<s> This is important to lifelong learning where self and peer assessment provide critical and informal feedback in the workplace and in the myriad of other human **activities** **undertaken** in daily life . </s>

Example selection issues

V obj N*
undertake •••
activities undertaken
perform •••
organize •••
inhibit •••
monitor •••
increase •••
stimulate •••

GDEX

Longest commonest match (colligation) can be less evident

<s> Companies **undertake** marketing **activities** in order to elicit some kind of response from buyers . </s>

<s> Having difficulty **undertaking** a particular **activity** is inherently a subjective determination . </s>

<s> Rather , there has been the specialization of the **activities** **undertaken** by business units and increasing cooperation between them . </s>

<s> Edison 's organizational innovation lay in the range and scale of the research **activities** **undertaken** . </s>

<s> They **undertake** political **activity** , but not of the traditional kind . </s>

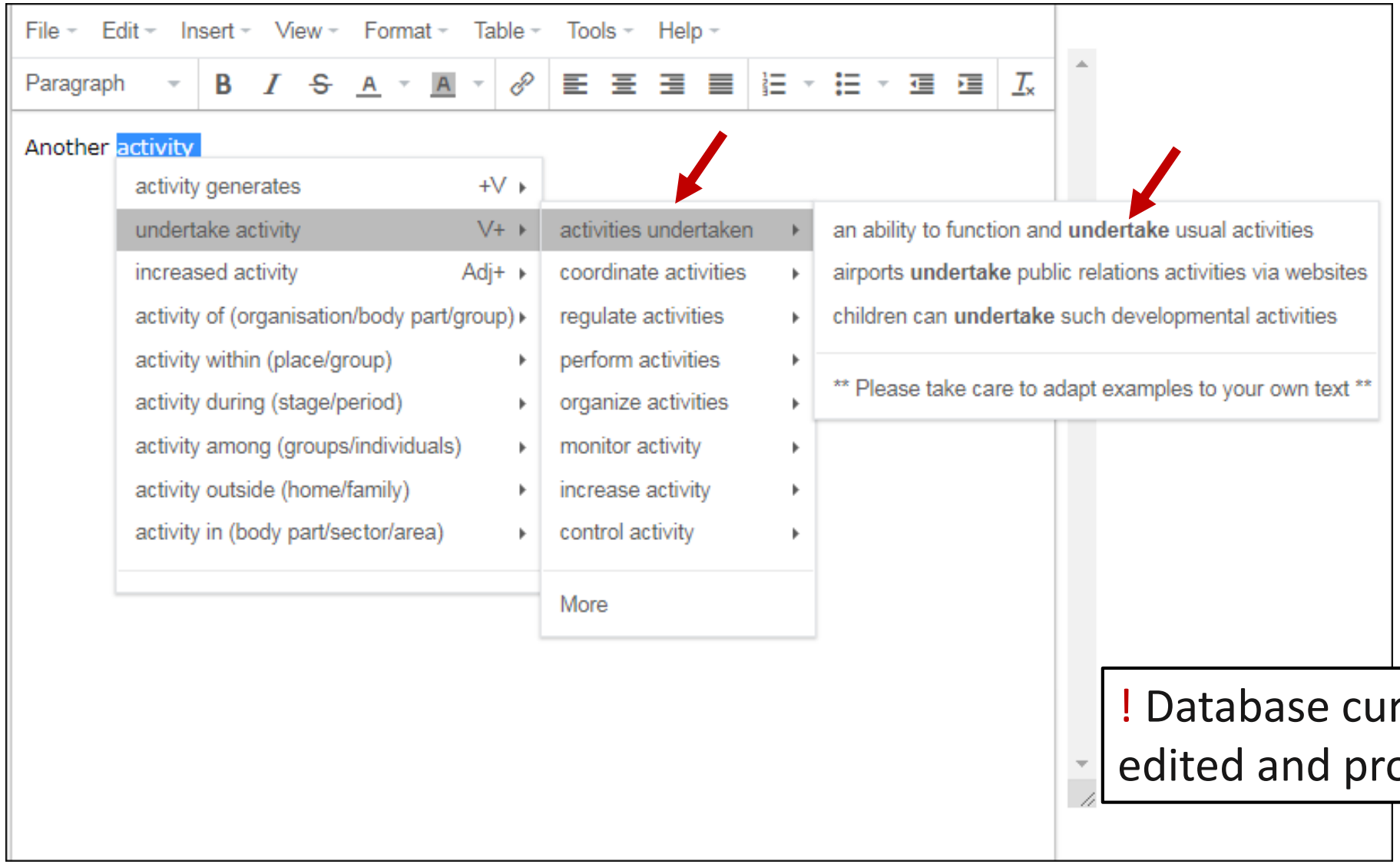
<s> For patients at high risk of a cardiovascular event , **undertaking** sexual **activity** may constitute a significant risk . </s>

<s> When an international firm crosses national borders to **undertake** business **activities** , it may also cross cultural borders . </s>

<s> Applicants must only **undertake** unpaid **activities** related to the purposes of the sponsoring organization . </s>

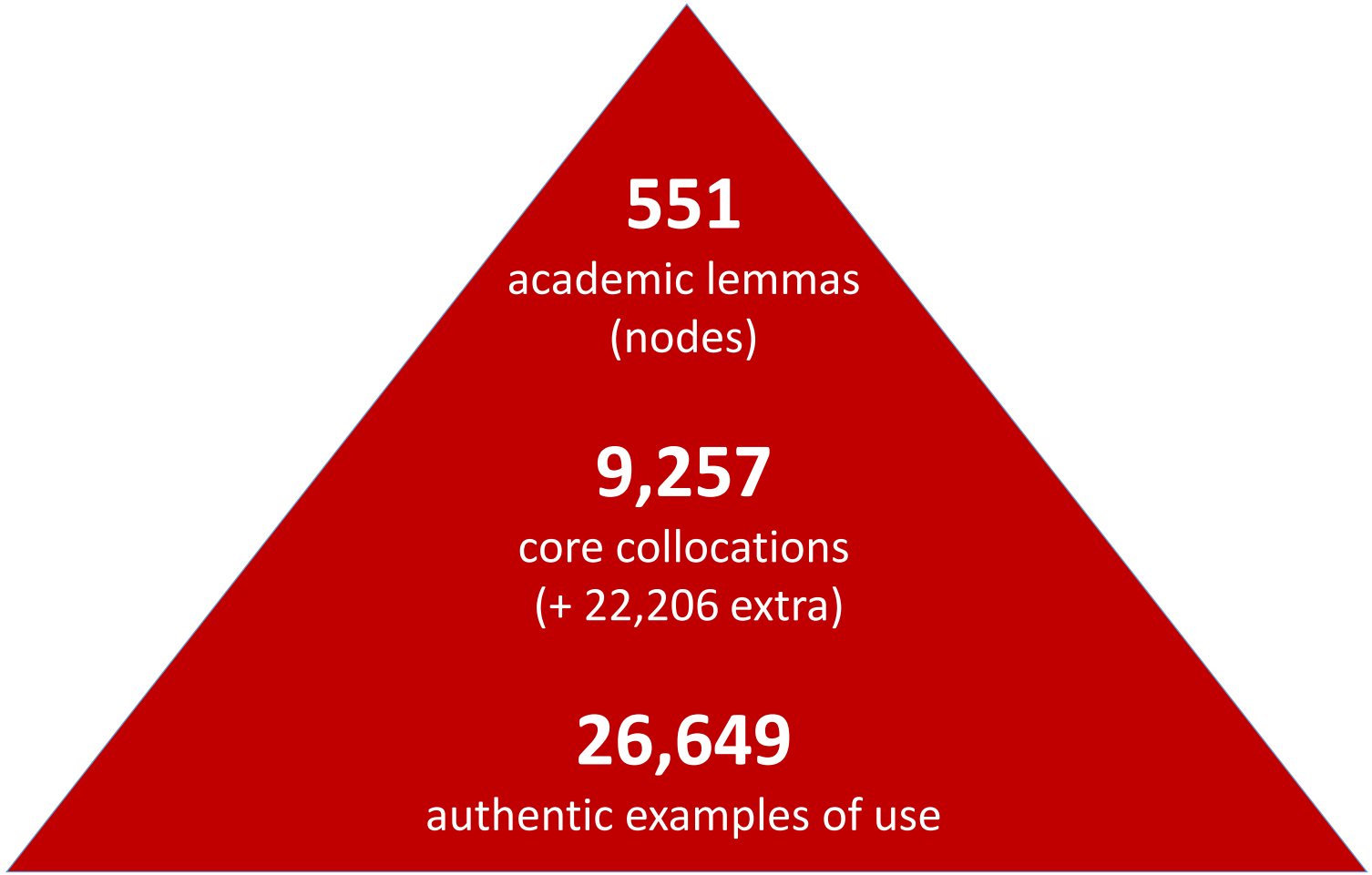
ColloCaid 0.4

Version 0.4



! Database currently being edited and proofread

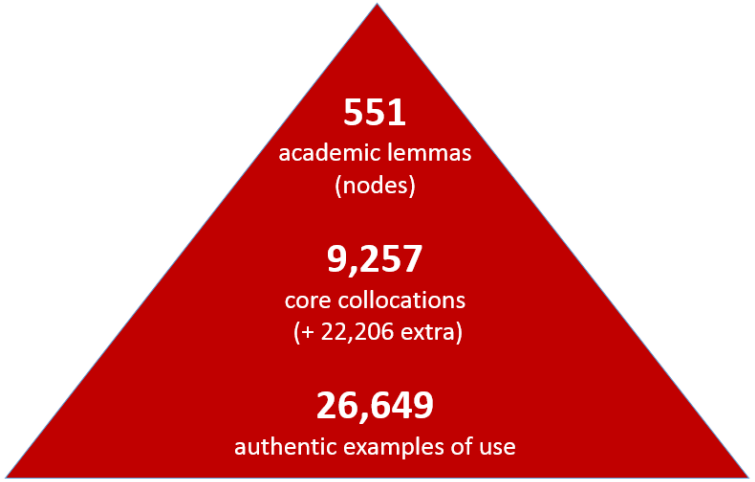
Version 0.4


551
academic lemmas
(nodes)
9,257
core collocations
(+ 22,206 extra)
26,649
authentic examples of use

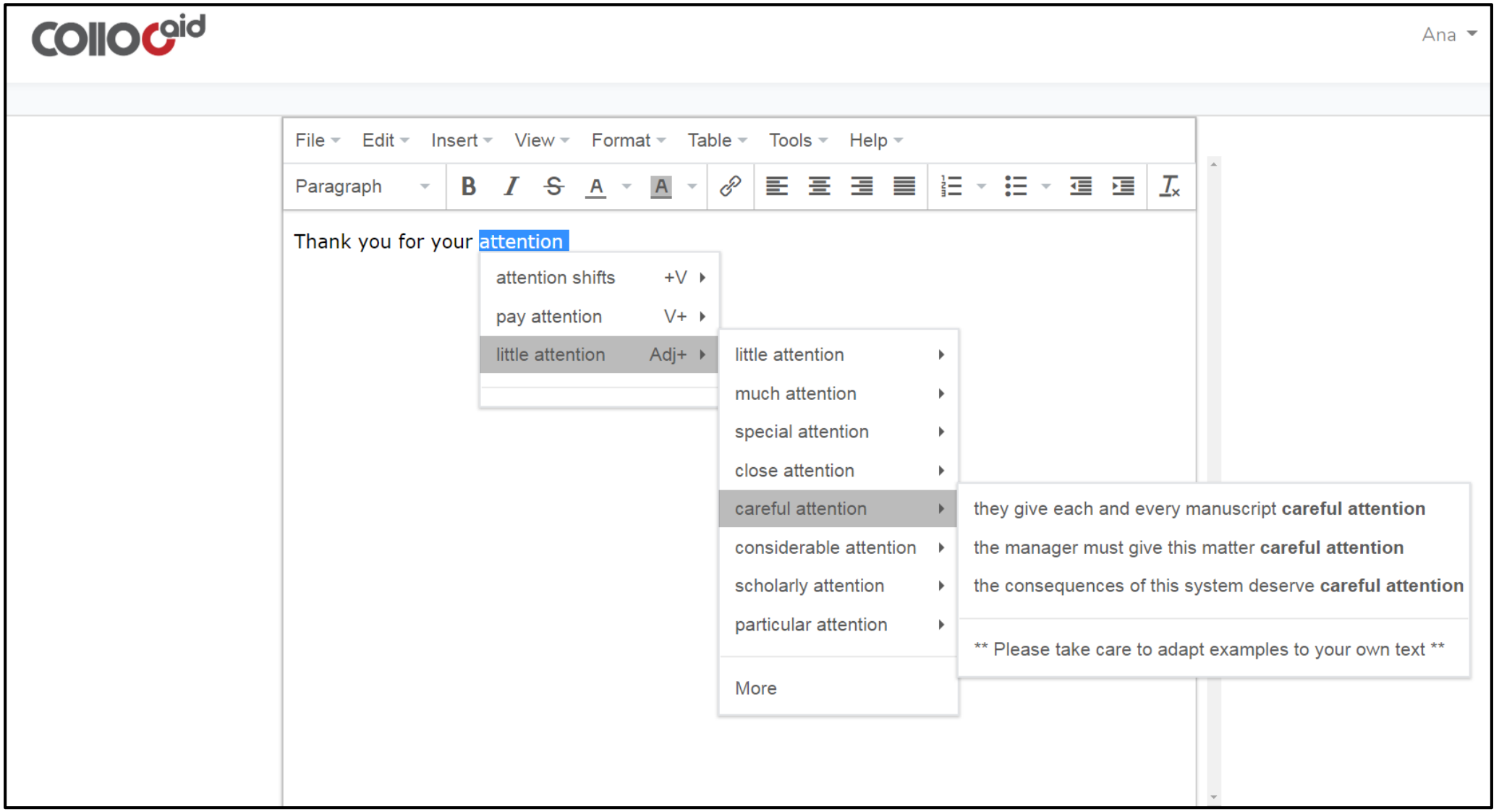
Conclusion

Conclusion

- We couldn’t have achieved anything near what we have managed so far without e-lexicographic tools & resources
- But spent a lot of time curating the data manually
- We hope our analysis of what needs editing/manual curation can contribute to further advances in e-lexicography
  - Collocation Workshop 30/09/2019 ☺
    - e.g. Rychlý & Jakubíček; Fuhrmann et al.; Krek et al.
- And we draw attention to the dangers of uncritical use of existing tools and resources
  - Trust the text (Sinclair 2004), but beware of the rest



www.collocaid.uk



Further information

www.collocaid.uk

AHRC - AH/P003508/1

Principal Investigator Dr Ana Frankenberg-Garcia (Surrey)

Co-investigators Prof Robert Lew (Poznan), Prof Jonathan Roberts (Bangor)

Researchers Dr Geraint Rees (Surrey), Dr Nirwan Sharma (Bangor)

& Peter Butcher (Bangor)

Come to our demo tomorrow!