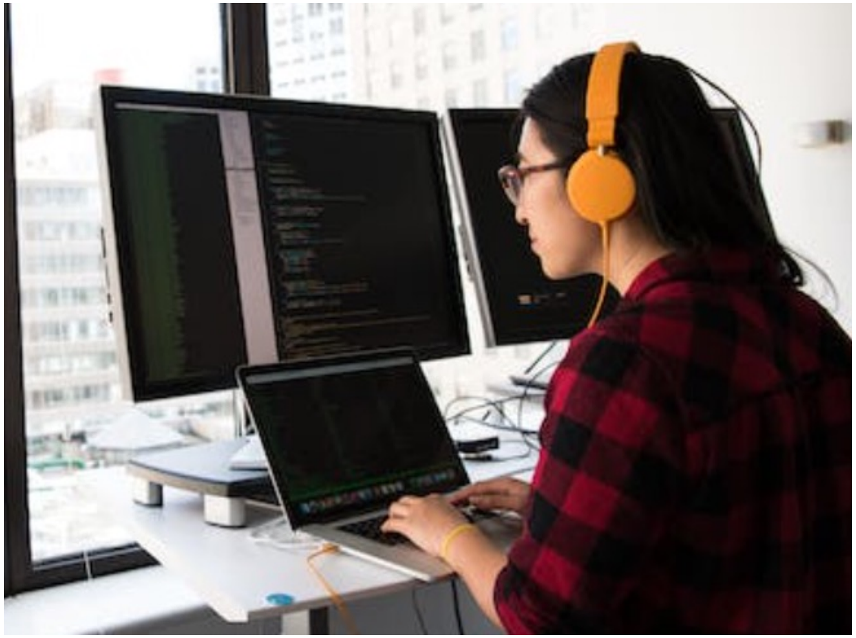
The potential of corpus linguistics

Alice Deignan

School of Education

Two ways of doing linguistic research



Corpus linguistics

To do corpus linguistics, you need two things:

A collection of texts, written and/ or transcribed spoken, in machine readable form (your corpus, plural corpora) and

A software package so that you can interrogate your corpus in various ways

Changes in access to corpus data

Corpus linguistics: from small, expensive collections to massive, freely available corpora.

When I started:

1998: accessing the Bank of English corpus in Birmingham from Leeds via a telnet link. Cost £500 per year, for 59 million words, one of the largest publicly available. Had to learn basic keystroke commands.

Now:

2023: using the Oxford English Corpus, BNC and many others through Sketchengine. Cost: £60 per year for Sketchengine/ or free institutional access, for billions of words, in dozens of languages, sophisticated processing packages, user interfaces etc.

Goals of the discipline

The early days: very much a child of English Language Teaching, English for Specific and Academic Purposes.

Looking inwards at the core structures and lexis of English. Also other languages, but at the beginning, English was by far the most researched- much funding was from ELT.

Corpus linguistics 1980s- 2023

A journey from a fairly young discipline, which was still exploring its own core knowledge base, to a maturing field, encompassing a number of sub specialisms, which now looks outward to the contributions it can make to wider questions.

Many research questions have a language dimension.

In the 2010s and 2020s: Looking outwards

Joining the Social Sciences, with contributions to make to wider issues;
Eg a number of projects based at CASS (Corpus Approaches to Social Sciences) at Lancaster University, an ESRC large centre since 2013, such as 'Corpus approaches to healthcare', 'Changing climates' 'understanding corporate communications'
Leeds has a network of corpus linguists, across the Schools of Computing, Education, English, Linguistics, and others. The group is a satellite of 'Language at Leeds'.

Corpus linguistics and the social sciences

Valuable in any research in which we want to know:

- what the characteristics of particular discourses are: e.g. Berber Sardinha & Pinto (2017) American television and off-screen registers: a corpus-based comparison. *Corpora* 12/1, 85-114;
- how an issue or group of people is represented in (public) discourse: e.g. Baker et al (2013) *Discourse and media attitudes: The representation of Islam in the British press*. CUP.

The linguistic challenge of the transition from primary to secondary school

ESRC ES/R006687/1
Sept 2018- Dec 2022

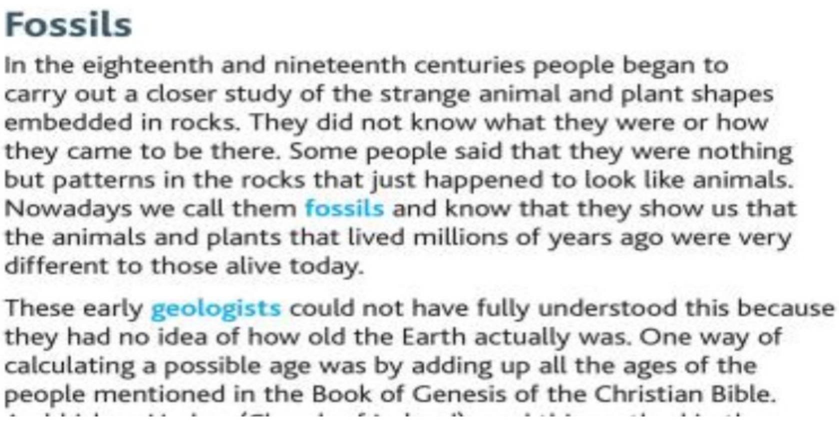
RQs:

How does the language of school change between late primary school and early secondary school?

How is the language of secondary school different from everyday language?

How much variation is there in language between different school subjects?

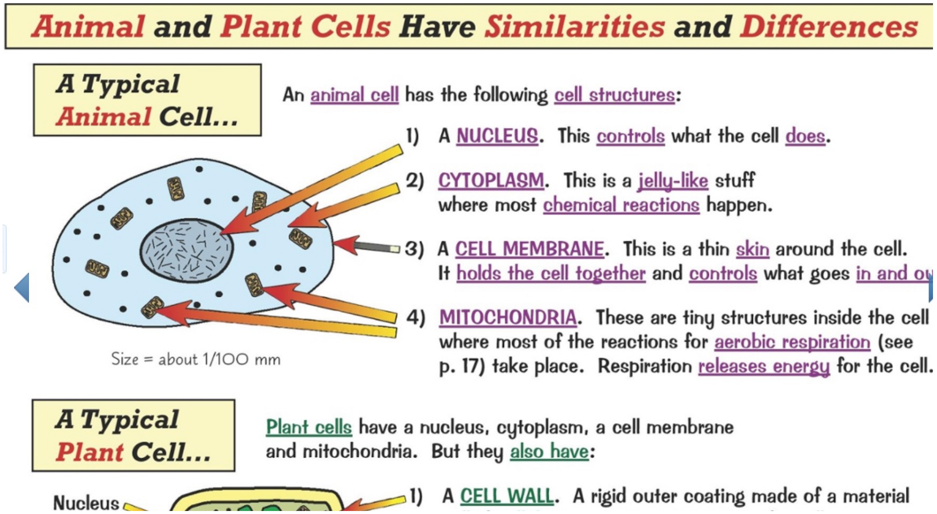
Which is more accessible..?



Fossils

In the eighteenth and nineteenth centuries people began to carry out a closer study of the strange animal and plant shapes embedded in rocks. They did not know what they were or how they came to be there. Some people said that they were nothing but patterns in the rocks that just happened to look like animals. Nowadays we call them fossils and know that they show us that the animals and plants that lived millions of years ago were very different to those alive today.

These early geologists could not have fully understood this because they had no idea of how old the Earth actually was. One way of calculating a possible age was by adding up all the ages of the people mentioned in the Book of Genesis of the Christian Bible.



Animal and Plant Cells Have Similarities and Differences

A Typical Animal Cell...

An animal cell has the following cell structures:

1) A NUCLEUS. This controls what the cell does.

2) CYTOPLASM. This is a jelly-like stuff where most chemical reactions happen.

3) A CELL MEMBRANE. This is a thin skin around the cell. It holds the cell together and controls what goes in and ou

4) MITOCHONDRIA. These are tiny structures inside the cell where most of the reactions for aerobic respiration (see p. 17) take place. Respiration releases energy for the cell.

Size = about 1/100 mm

A Typical Plant Cell...

Plant cells have a nucleus, cytoplasm, a cell membrane and mitochondria. But they also have:

1) A CELL WALL. A rigid outer coating made of a material

The project corpora

With many corpus projects, the data collection (i.e. building the corpus) is a very time-intensive part of the work, and spans the bulk of the project time-line. A skilled RF is essential.

Written data
Worksheets
Textbooks
Exams and assessment tasks
Lesson presentations
Vocabulary/glossary booklets

Spoken data
Audio recordings of lessons (teacher talk only)

Subjects: English, maths, science, history, geography

13 schools contributed data, across the North of England, 5 secondary, 8 primary.

The school materials comprise 1000s of files, currently approx. 2.6 million words, just being finished.

Transcribing speech, removing duplicate files, cleaning up files, organizing, and, usually, converting to txt, is a big job.

Corpora

Also: a ‘reference’ corpus, representing everyday language, which can be used for comparison, to show what is special about the specialized data collected. Many are available.

As building a corpus is so time-intensive; it is much easier to use pre-existing corpora if available.

LexisNexis can be used to build a newspaper corpus around a theme very quickly.

Example findings: maths

In our primary school data, there are around 10,000 unique words in maths lessons.

In our secondary school data, there are around 15,600 unique words. The new words are almost all technical and specific; a few are algebraic symbols.

We can list these by frequency.

Example findings: science

In our primary school data, there are around 11,600 unique words in science classes.

In our secondary school data, there are around 14,000 unique words.

A number of the new secondary school words are common in everyday language, and thus likely to be known to children BUT, they have subtly different meanings. Eg *image*, *substance*, *concentration*. Detecting these differences needs manual analysis as well as the automatic tools.

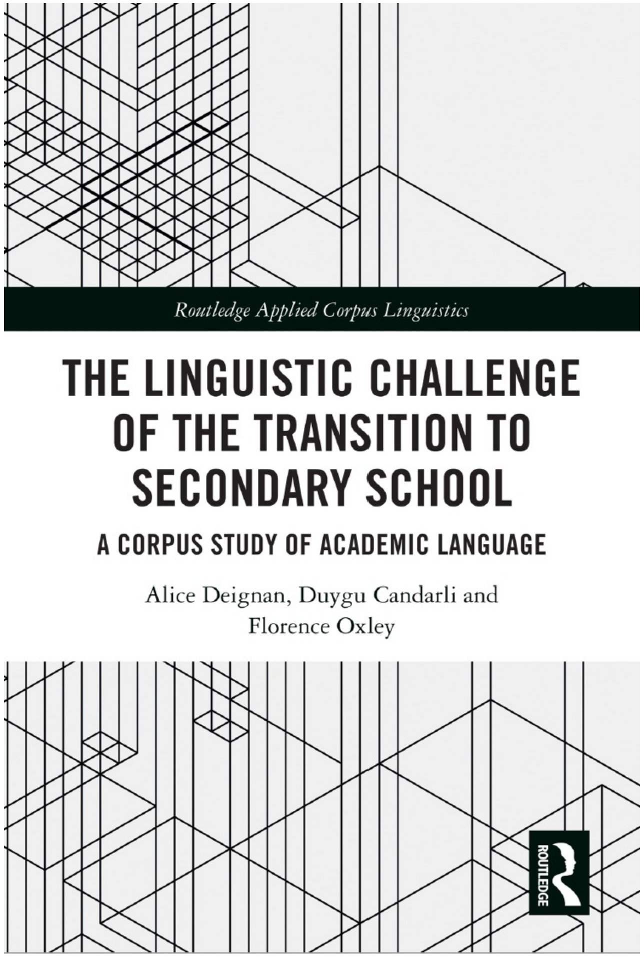
More about the project

Open Access book:

https://www.taylorfrancis.com/books/oa-edit/10.4324/9781003081890/linguistic-challenge-transition-secondary-school-alice-deignan-duygu-candarli-florence-oxley

(can be found on Routledge website)

Project website: https://linguistictransition.leeds.ac.uk/



Routledge Applied Corpus Linguistics

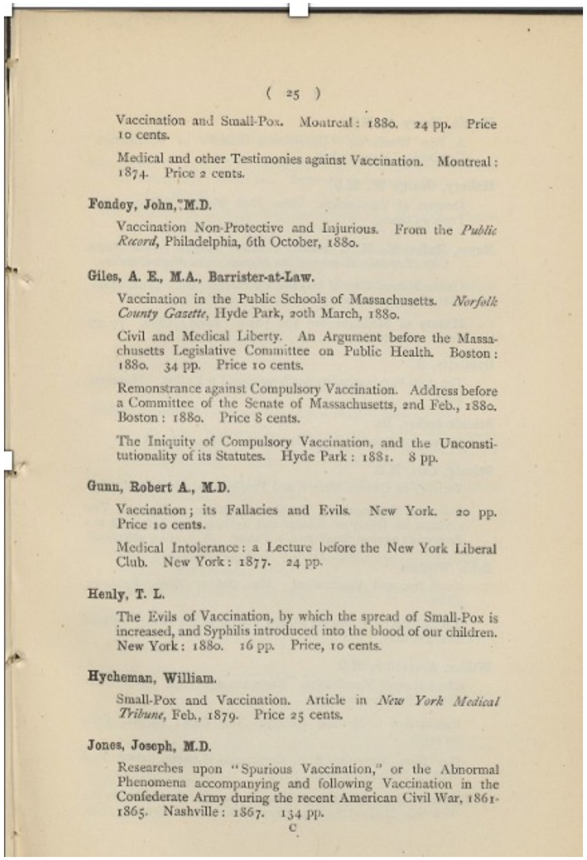
THE LINGUISTIC CHALLENGE OF THE TRANSITION TO SECONDARY SCHOOL

A CORPUS STUDY OF ACADEMIC LANGUAGE

Alice Deignan, Duygu Candarli and Florence Oxley

Questioning Vaccination Discourse (QuoVaDis): A corpus-based study ESRC ES/V000926/1 2021-24

- Has built a number of studies to study discourse around vaccination and to understand motivations for and against being vaccinated.
- Corpora of UK press; Twitter; Reddit; Mumsnet; UK parliamentary debates.
- Also, a corpus of Victorian Anti-Vaccination Discourse (VicVaDis) comprising pamphlets and informal journals of the anti-vaccination movement from the first compulsory smallpox vaccinations.
- https://www.lancaster.ac.uk/vaccination-discourse/



(25)

Vaccination and Small-Pox. Montreal: 1880. 24 pp. Price 10 cents.

Medical and other Testimonies against Vaccination. Montreal: 1874. Price 2 cents.

Fondey, John, M.D.

Vaccination Non-Protective and Injurious. From the *Public Record*, Philadelphia, 6th October, 1880.

Giles, A. E., M.A., Barrister-at-Law.

Vaccination in the Public Schools of Massachusetts. *Norfolk County Gazette*, Hyde Park, 20th March, 1880.

Civil and Medical Liberty. An Argument before the Massachusetts Legislative Committee on Public Health. Boston: 1880. 34 pp. Price 10 cents.

Remonstrance against Compulsory Vaccination. Address before a Committee of the Senate of Massachusetts, 2nd Feb., 1880. Boston: 1880. Price 8 cents.

The Iniquity of Compulsory Vaccination, and the Unconstitutionality of its Statutes. Hyde Park: 1881. 8 pp.

Gunn, Robert A., M.D.

Vaccination; its Fallacies and Evils. New York. 20 pp. Price 10 cents.

Medical Intolerance: a Lecture before the New York Liberal Club. New York: 1877. 24 pp.

Henly, T. L.

The Evils of Vaccination, by which the spread of Small-Pox is increased, and Syphilis introduced into the blood of our children. New York: 1880. 16 pp. Price, 10 cents.

Hycheman, William.

Small-Pox and Vaccination. Article in *New York Medical Tribune*, Feb., 1879. Price 25 cents.

Jones, Joseph, M.D.

Researches upon "Spurious Vaccination," or the Abnormal Phenomena accompanying and following Vaccination in the Confederate Army during the recent American Civil War, 1861-1865. Nashville: 1867. 134 pp.

C

Resources available here

Corpus software

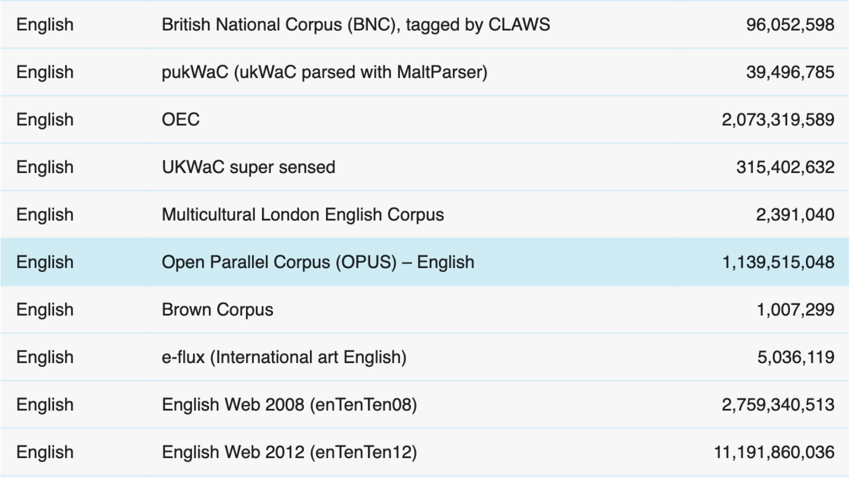
Leeds has an institutional license to Sketch Engine.

There is a huge range of free corpora accessible through their interface.

Lots of online tutorial help.

Kilgarriff, A. et al. 2014. *The Sketch Engine. 10 years on.* Lexicography, 1: 7-36.

http://sketchengine.eu



English	British National Corpus (BNC), tagged by CLAWS	96,052,598
English	pukWaC (ukWaC parsed with MaltParser)	39,496,785
English	OEC	2,073,319,589
English	UKWaC super sensed	315,402,632
English	Multicultural London English Corpus	2,391,040
English	Open Parallel Corpus (OPUS) – English	1,139,515,048
English	Brown Corpus	1,007,299
English	e-flux (International art English)	5,036,119
English	English Web 2008 (enTenTen08)	2,759,340,513
English	English Web 2012 (enTenTen12)	11,191,860,036

Resources

LancsBox is freely available to download.

Stands for Lancaster University corpus toolbox.

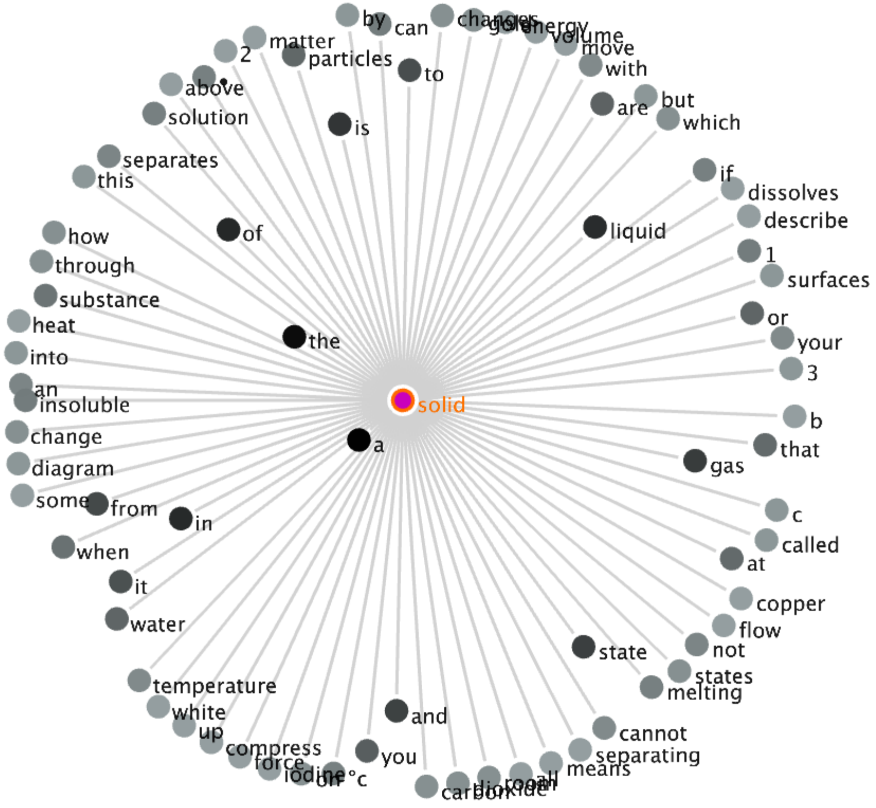
Good for building own corpora, flexible.

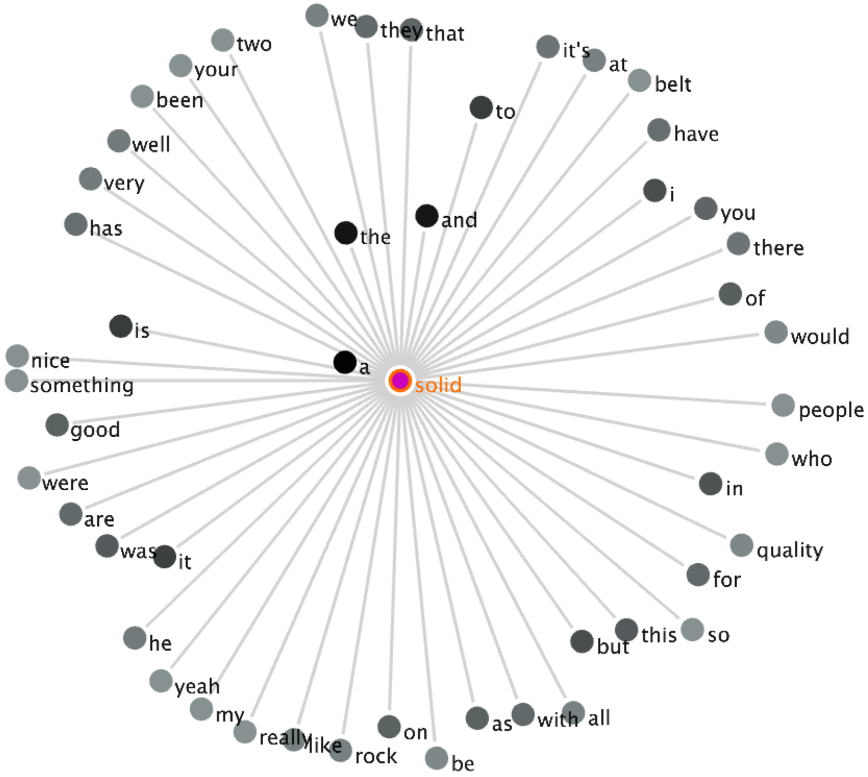
Offers a wide range of statistical choices (eg how dispersion is calculated etc).

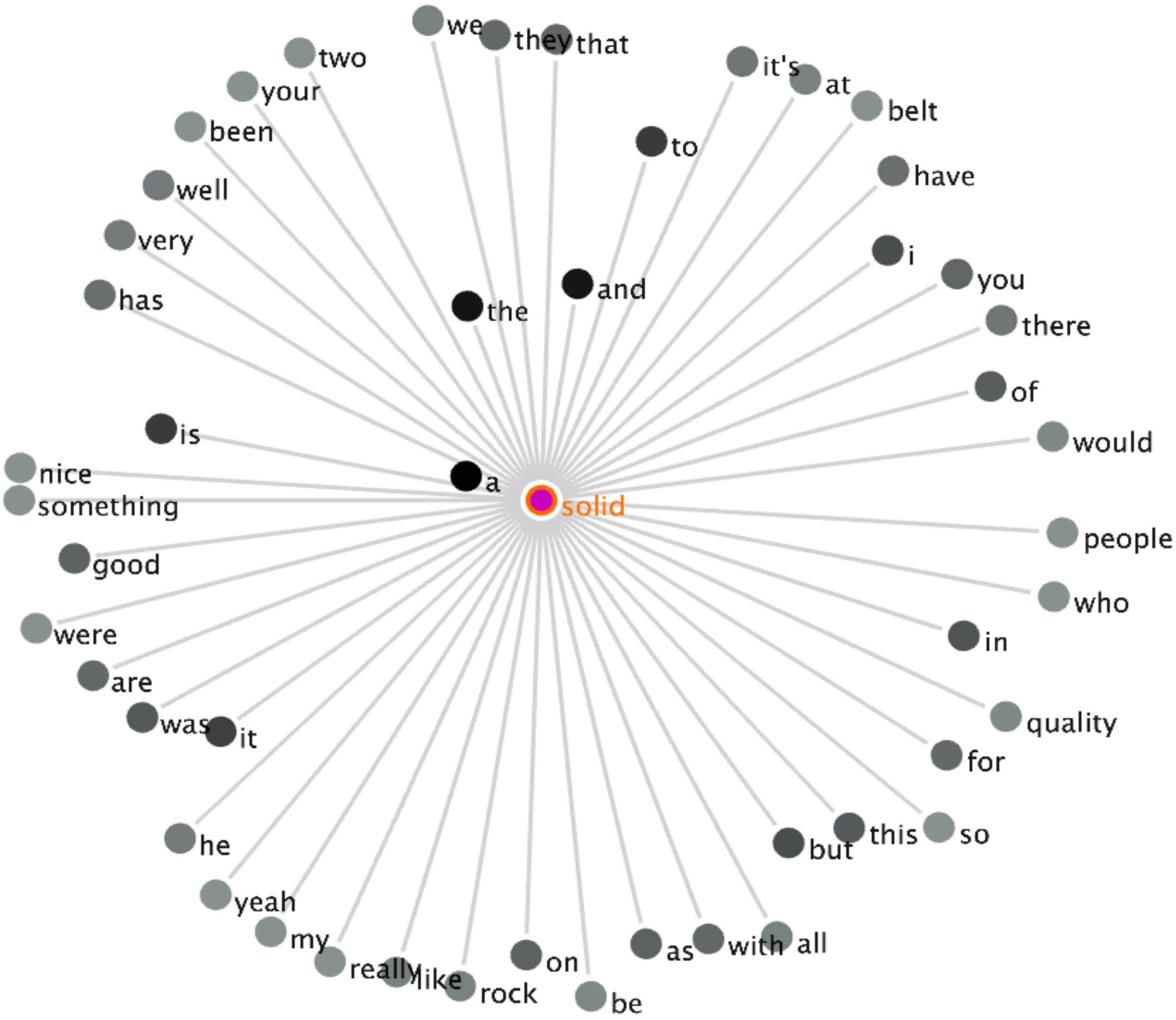
Nice visualisations.

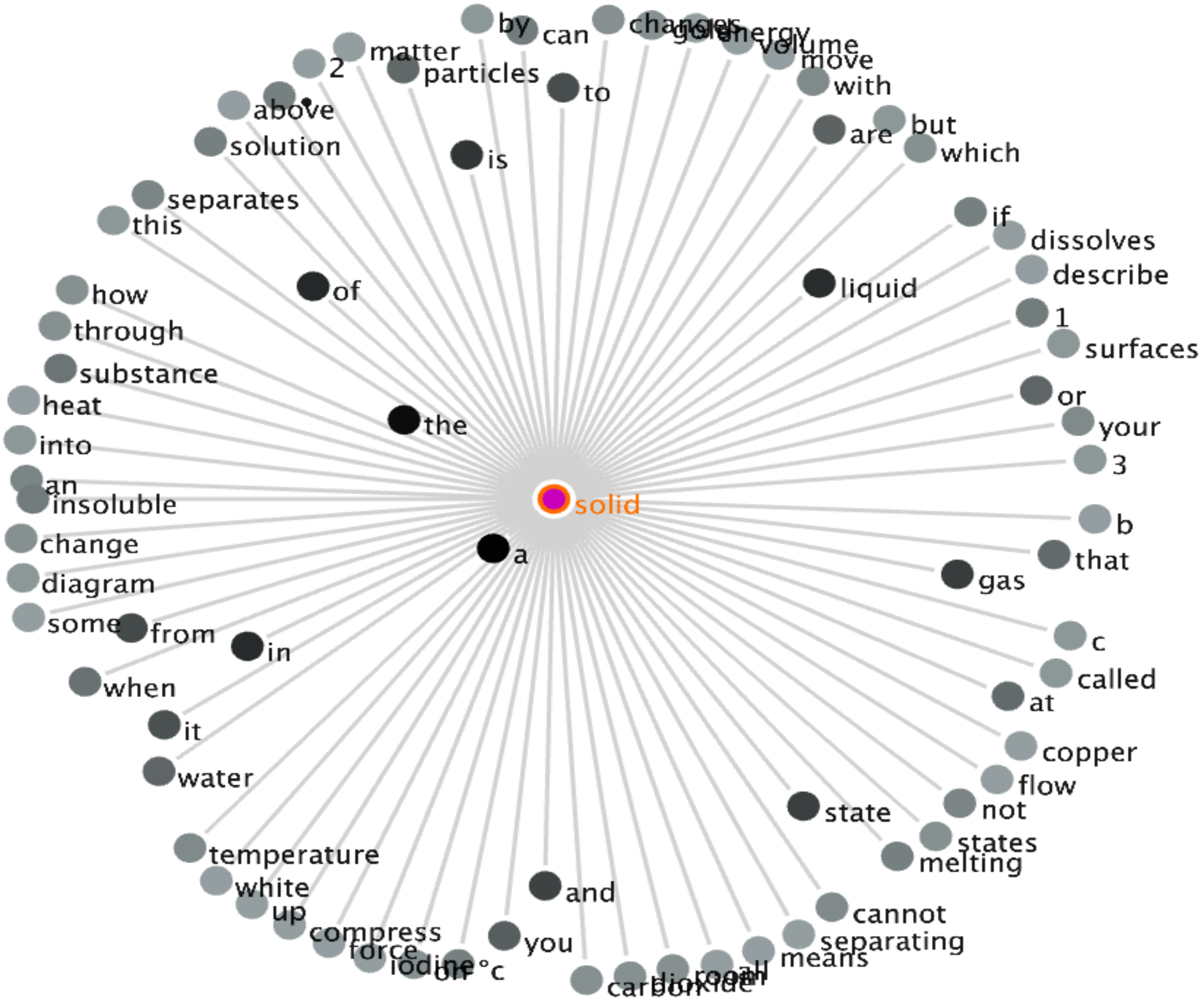
Brezina, V. Weill-Tessier, P. & McEnery, A. 2021. #LancsBox v.6

Graphcoll: collocates of ‘solid’ in LancsBox






solid
we
they
that
it's
at
belt
two
your
been
to
have
well
very
i
you
has
the
and
there
of
is
would
nice
a
something
people
good
who
were
in
are
was
it
quality
for
he
but
this
so
yeah
my
really
like
rock
on
be
as
with
all

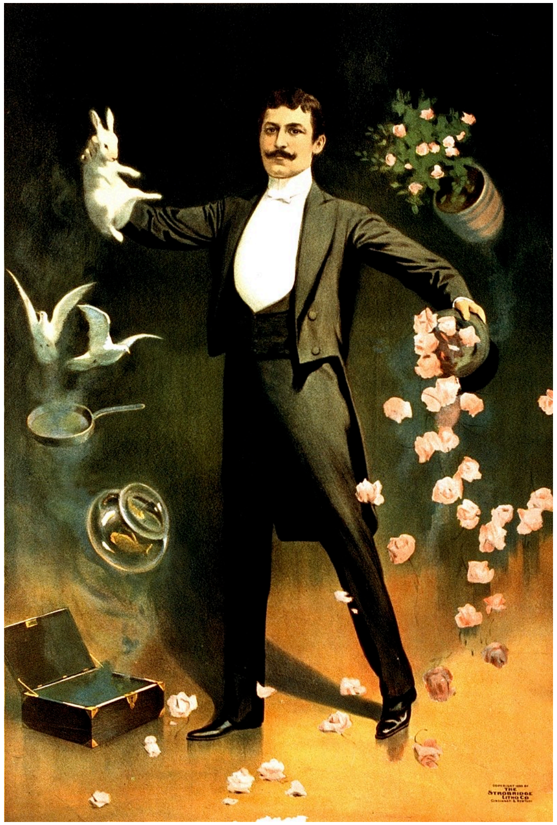

solid
by
can
changes
gold
energy
volume
move
with
to
but
are
which
is
if
dissolves
describe
liquid
1
surfaces
or
your
3
b
that
gas
c
called
at
copper
flow
not
states
melting
state
cannot
separating
means
all
room
dioxide
carbon
and
you
°c
on
iodine
force
compress
up
white
temperature
water
it
when
in
from
some
diagram
change
a
insoluble
an
into
heat
the
substance
through
how
of
this
separates
solution
above
.
2
matter
particles

Do you need to be a specialist to use corpus linguistics?

Yes and no.

It’s a method but also more than a method.

It can look superficially easy, but like every academic discipline, some scholars spend their entire research careers on it.



It is now a developed field, with specialist MAs, a huge literature, 3 dedicated international journals etc.

It has a number of sub-specialisms.

Where to get started?

The basics can be learnt quickly if you have some knowledge of the basic terminology and concepts of language analysis, and basic IT skills. Check out online courses and resources.

For larger projects, a corpus perspective can be brought in through a specialist Co-I or post-doc. There are cohorts of excellent PhD graduates specializing in corpus linguistics.