Aston Corpus Summer School 2011

CORPORA IN LEXICOGRAPHY (PART ONE)

Iztok Kosem

Trojina, Institute for Applied Slovene Studies
Ljubljana, Slovenia

Contact: iztok.kosem@trojina.si

Lexicography

- The art of compiling, writing and editing dictionaries (Wikipedia)
- Lexicographer: "writer of dictionaries; a harmless drudge" (Johnson)
- "LEXICOGRAPHER, n. A pestilent fellow who, under the pretense of recording some particular stage in the development of a language, does what he can to arrest its growth, stiffen its flexibility and mechanize its methods." (Ambrose Bierce, *The Devil's Dictionary*, 1911)
- First discipline to fully utilize corpus data
- High demands of lexicographers pushing the functionality of corpus tools

- **Exercise 1**

- In the Sketch Engine, select the CAJA corpus.
- Do the simple search for the lemma ***authority***.
- Make a sample of 50 concordances (use the function Sample in the Sketch Engine).
- Analyse the concordances. If you were making a dictionary entry, how many different senses of ***authority*** would you record?

authority /ɔːˈθɒrɪti/ *noun*

WORD FORMS:
authority, authorities

MENU

1. power to control people or activities
2. government department
3. **(the authorities)** organizations in charge of a country
4. expert
5. important written work
6. person with power
7. official permission
8. personal quality
9. **Computing** type of internet page

Corpora in lexicography

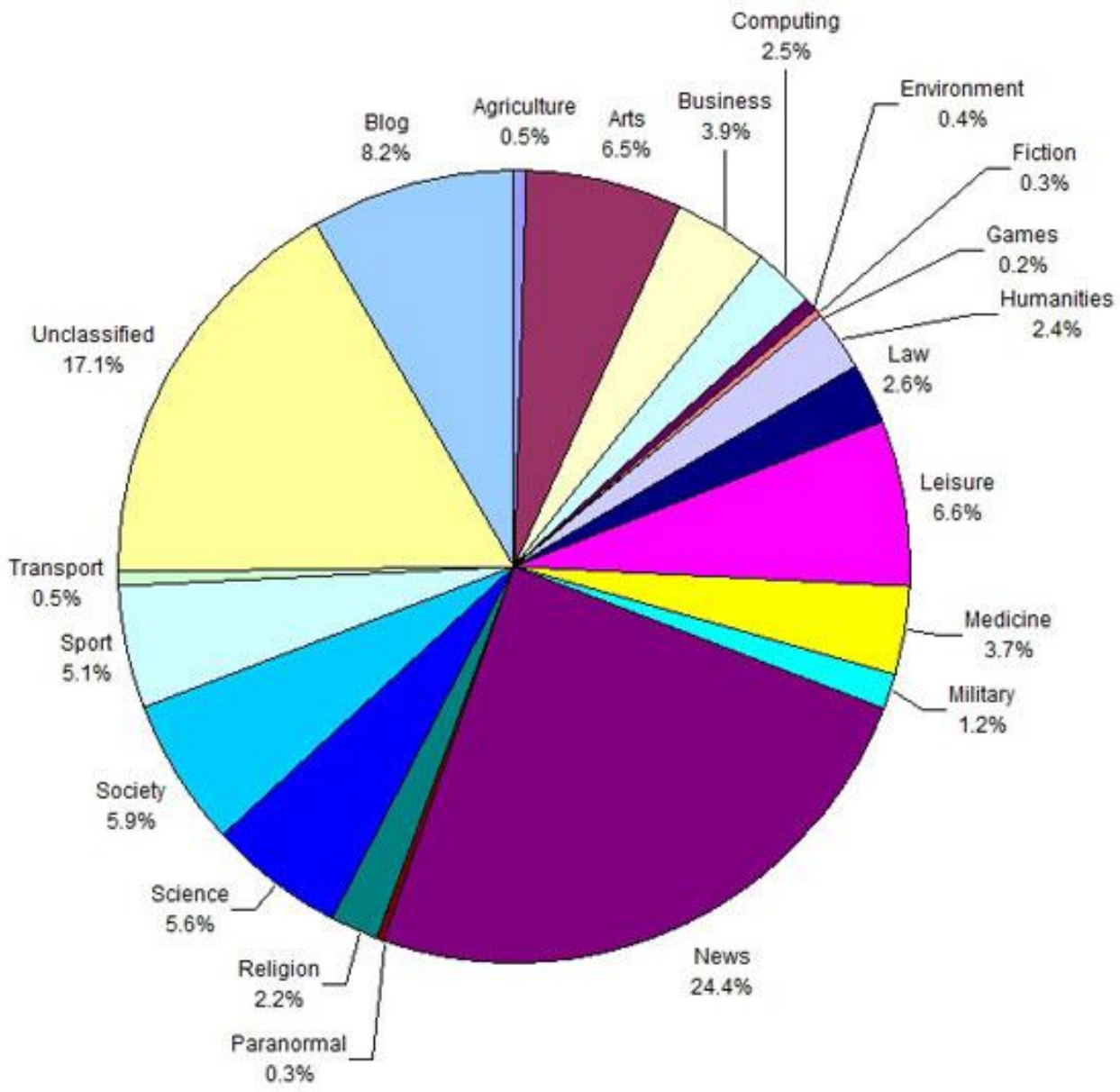
- Pre-computer age:
  - Samuel Johnson's dictionary (1755), OED (1928), Noah Webster's dictionary (1828)
  - Index cards (e.g. OED had 20 million index cards, 5 million citations)
  - Gathering citations by hand, bias towards atypical
- 1980s – COBUILD
  - Corpus: 18M words (initially 7M), written & spoken (UK & US)
  - Corpus-driven approach!
  - Dictionaries, grammars, word bank
  - Others followed: Longman, Cambridge, Oxford, Macmillan

Corpora in lexicography

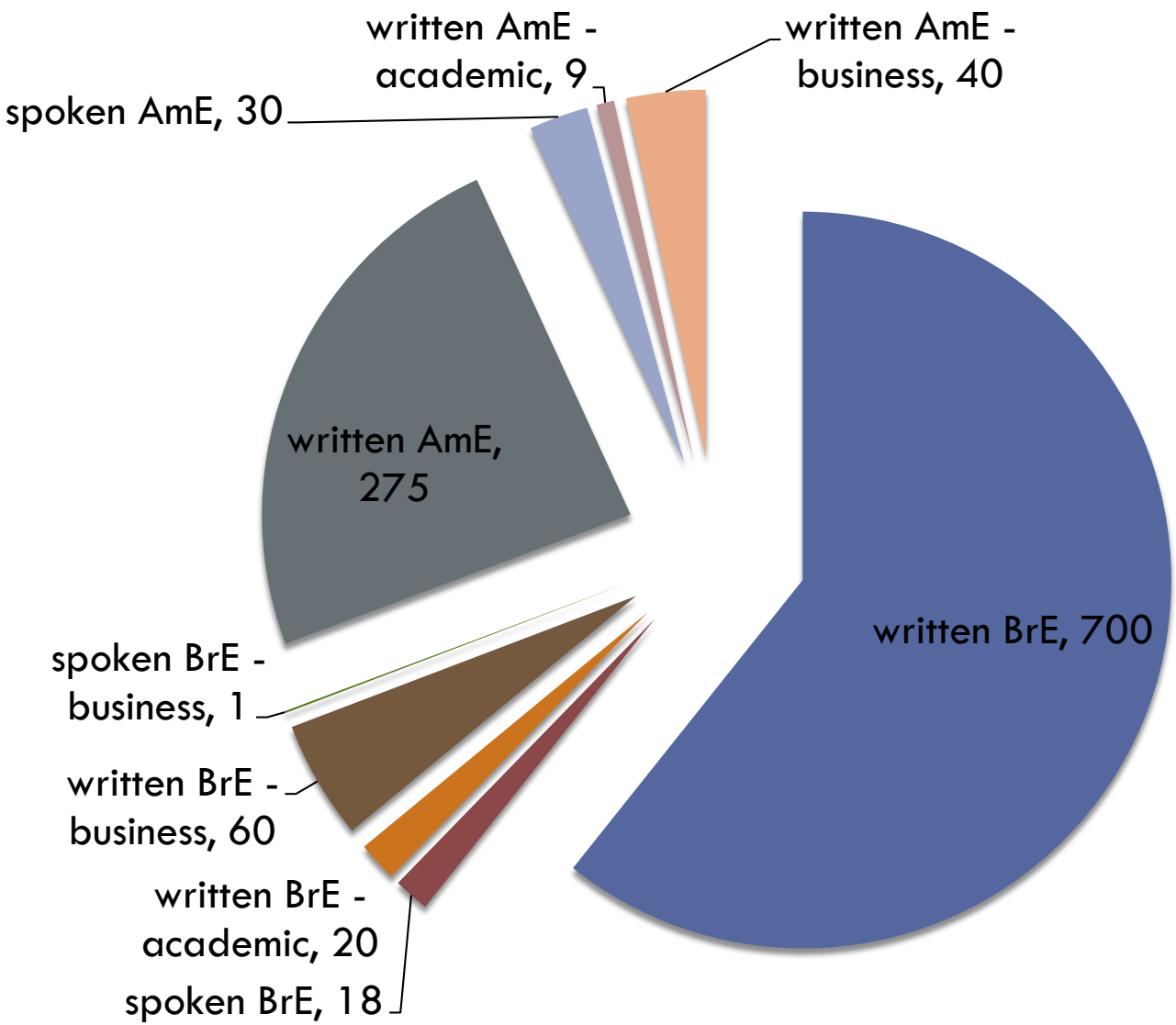
- 1990s-2000s: large corpora: greater depth of analysis, variety of texts, statistical accuracy
  - British National Corpus (100 million words; UK)
  - Bank of English (520 million words; UK, US, Aus, Can)

- 2000- : huge corpora
  - Web used as data source
  - Corpus collections of publishing houses:

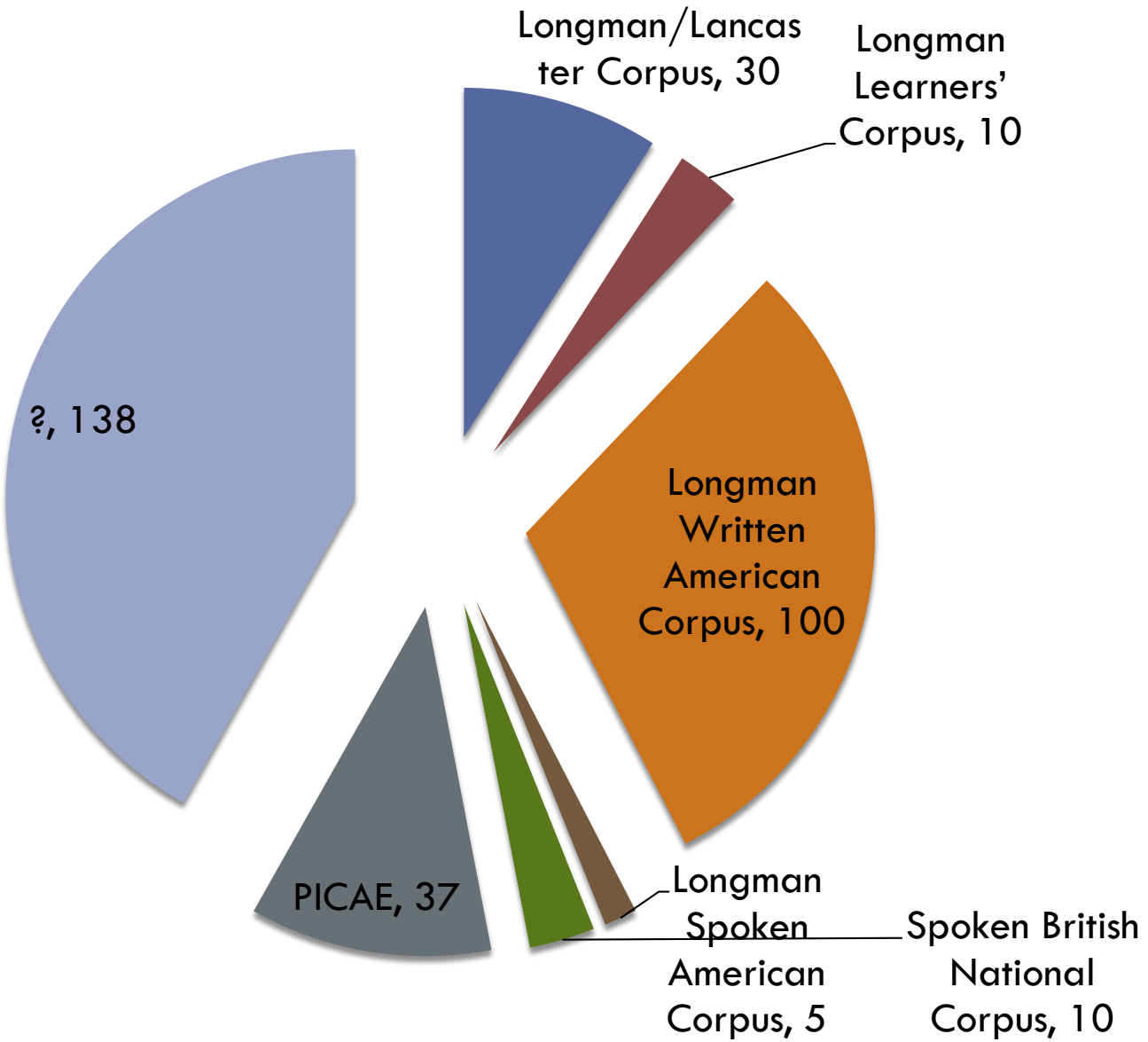
Oxford English Corpus (2 billion words)



Cambridge International Corpus (1,153 billion)

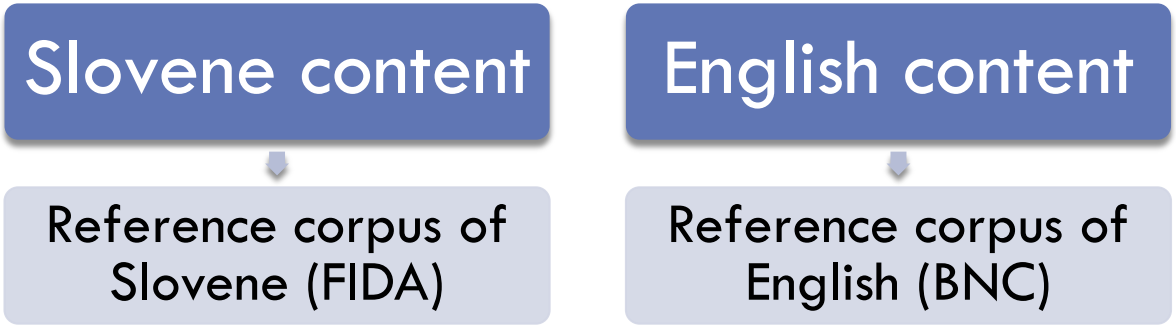


Longman Corpus Network (330 million words)



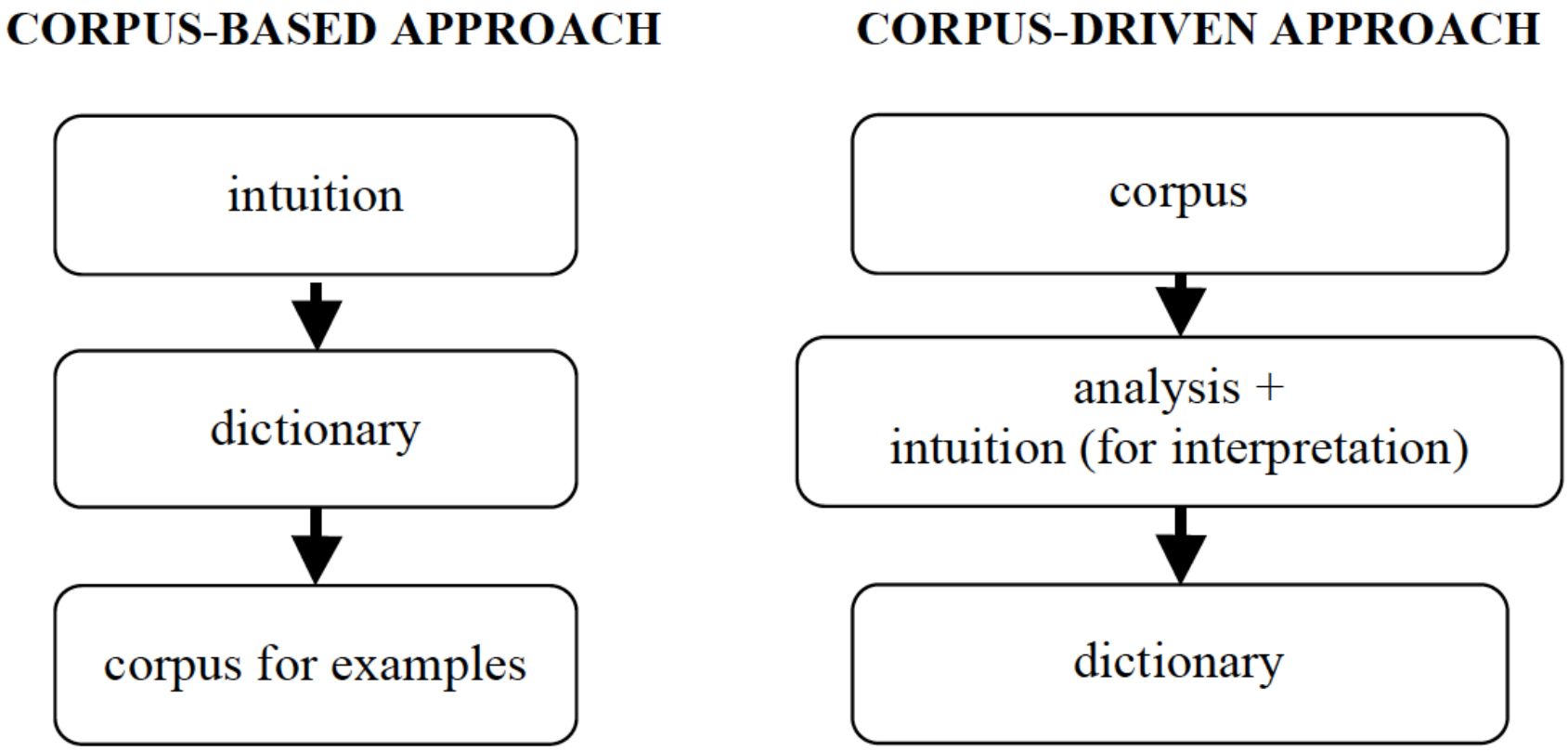
Corpora in lexicography

- Bilingual lexicography:
  - Corpora used less than in monolingual lexicography
  - Oxford Hachette English-French French-English dictionary (1994)
  - Comprehensive English-Slovene dictionary (2005)



  - parallel corpora used more and more

Corpus-based vs. corpus-driven



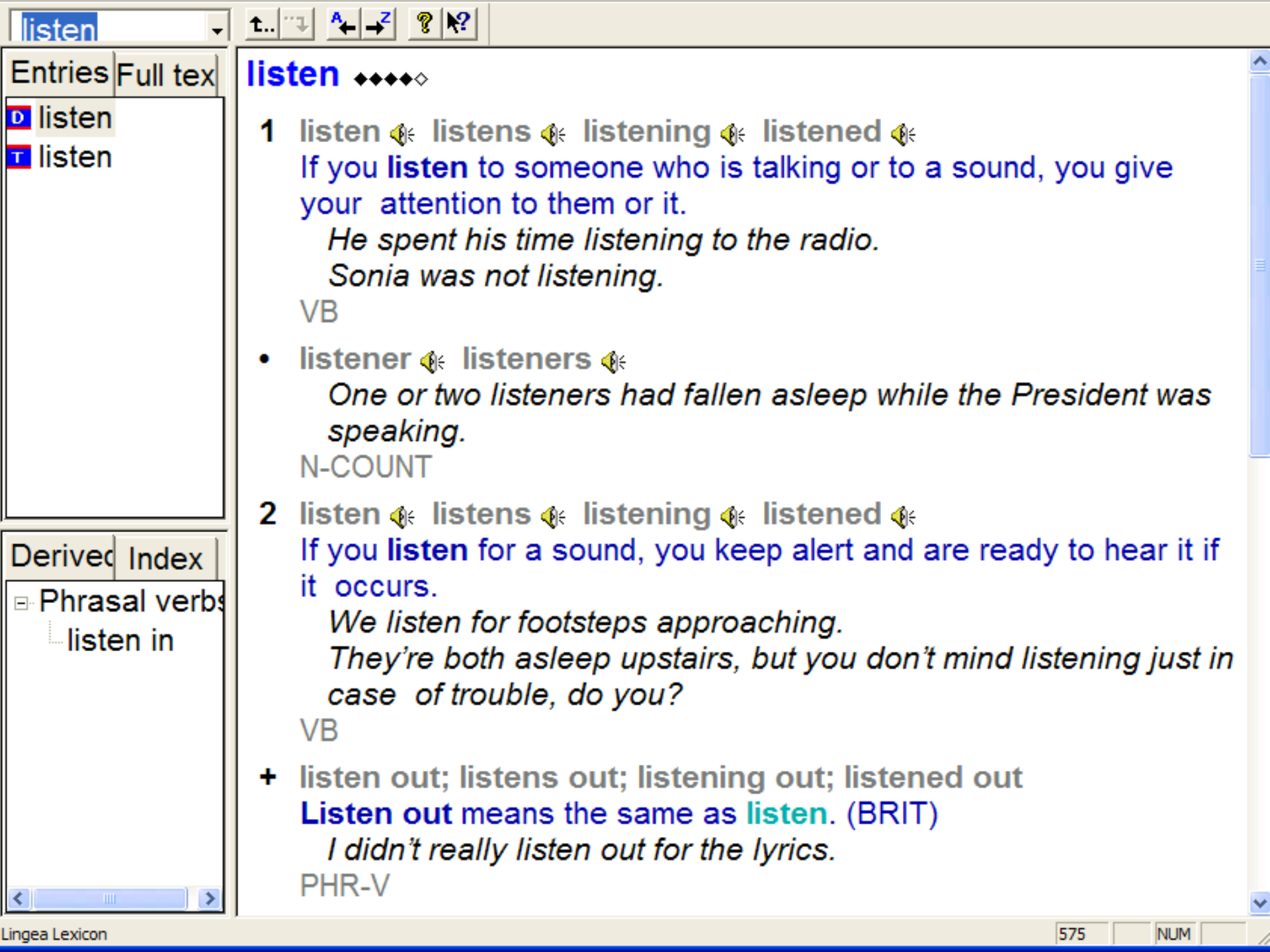
Corpus-based vs. corpus-driven

- "The results of [corpus] analysis are incorporated into specially designed **usage notes** and **study pages** in Cambridge dictionaries… In addition, **dictionary examples** illustrating word use can be taken from the corpus, making them sound natural and realistic." (Cambridge University Press)

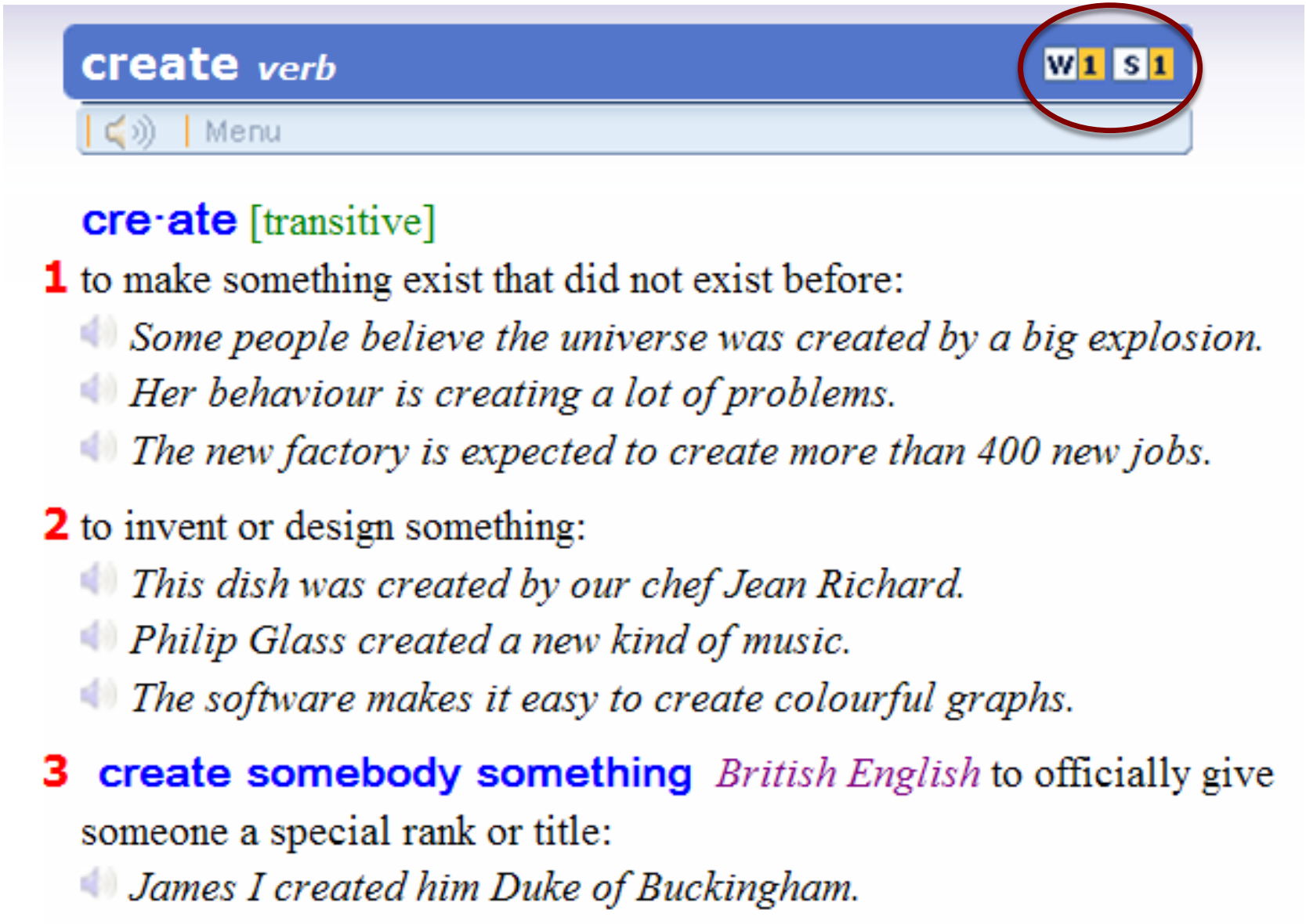
- "A corpus-driven approach involves a bottom-up methodology, beginning by selecting unedited examples from the corpus, identifying their shared and individual features, and only then grouping them for the purpose of lexicographic presentation." (Krishnamurthy, 2008: 231)

Corpus information in dictionary entries

- Headword list
- examples
- Labels
- Examples
- Phrases
- Collocations
- definitions like COBUILD
- Usage notes


listen
Entries
Full tex
listen
listen
Derived
Index
Phrasal verbs
listen in
listen ◆◆◆◆◇
1 listen listens listening listened
If you listen to someone who is talking or to a sound, you give your attention to them or it.
He spent his time listening to the radio.
Sonia was not listening.
VB
• listener listeners
One or two listeners had fallen asleep while the President was speaking.
N-COUNT
2 listen listens listening listened
If you listen for a sound, you keep alert and are ready to hear it if it occurs.
We listen for footsteps approaching.
They're both asleep upstairs, but you don't mind listening just in case of trouble, do you?
VB
+ listen out; listens out; listening out; listened out
Listen out means the same as listen. (BRIT)
I didn't really listen out for the lyrics.
PHR-V
Lingea Lexicon
575
NUM

Longman Dictionary of Contemporary English



create *verb* W1 S1

Menu

cre·ate [transitive]

1 to make something exist that did not exist before:

- *Some people believe the universe was created by a big explosion.*
- *Her behaviour is creating a lot of problems.*
- *The new factory is expected to create more than 400 new jobs.*

2 to invent or design something:

- *This dish was created by our chef Jean Richard.*
- *Philip Glass created a new kind of music.*
- *The software makes it easy to create colourful graphs.*

3 **create somebody something** *British English* to officially give someone a special rank or title:

- *James I created him Duke of Buckingham.*

Longman Exams Dictionary

mere[1] *adjective* W 3

/mɪə $ mɪr/ *superlative* **merest** [only before noun, no comparative]

1 used to emphasize how small or unimportant something or someone is:

- *She lost the election by a mere 20 votes.*
- *He's a mere child.*
- *It can't be a* ***mere coincidence*** *that they left at the same time.*
- *Many of the soldiers who went to war were mere boys.*

2 used to emphasize that something which is small or not extreme has a big effect or is important:

- *The merest little noise makes him nervous.*
- *The mere thought of food made her feel sick.*
- ***The mere fact*** *that the talks are continuing is a positive sign.*

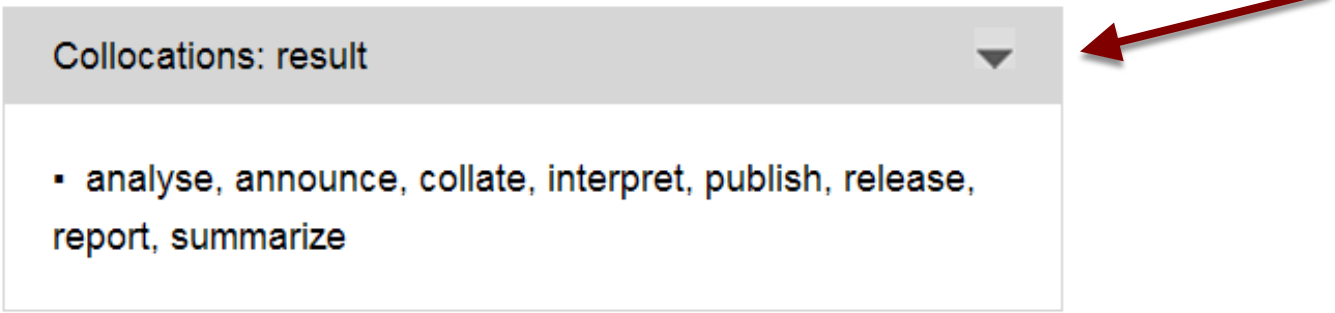
result n. (Macmillan English Dictionary)

3 [COUNTABLE] [OFTEN PLURAL] **a piece of information that is obtained by examining, studying, or calculating something**

Our results show that an effective vaccine is feasible.

result of: *The results of the survey will be published shortly.*

Thesaurus entry for this meaning of result



Get it right: maybe

Don't confuse the adverb **maybe** (one word), which means 'perhaps', with **may be** (two words), which means 'could be':

✗ *In an earthquake your house **maybe** badly damaged.*

✓ *In an earthquake your house **may be** badly damaged.*

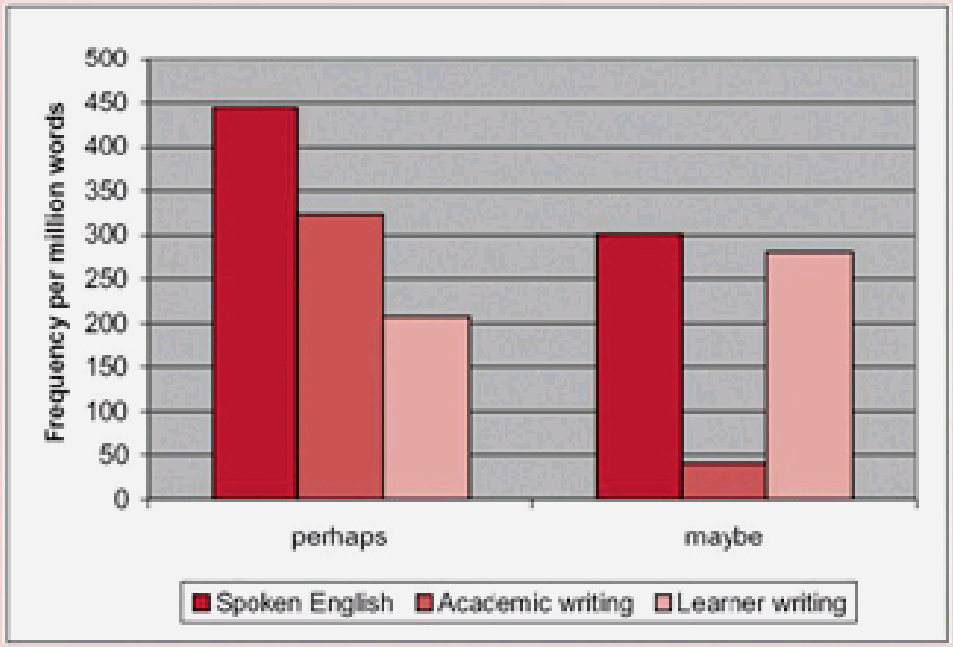
✗ *It **maybe** an unfulfilled dream.*

✓ *It **may be** an unfulfilled dream.*

Maybe and **perhaps** have the same meaning, but **maybe** is used mainly in spoken English and informal writing. In more formal English, **perhaps** is far more common:

*Now, **maybe** I haven't explained myself very well.*

*There are, **perhaps**, three principles which must be followed.*



Trends in modern lexicography

- Large corpora → more data to analyse
- More data better for computers (to exclude noise)
- Automating as much as possible
  - Automatic data collection and annotation (WebBootCat, Baroni et al., 2006)
  - Identifying salient data and presenting them to the lexicographer
- Lexicographer validates the data, makes the final selection
- Technology: from supportive to proactive role (Rundell, 2011)