Collocations in corpora and in speakers' minds

Ewa Dąbrowska

northumbria
UNIVERSITY

What is the mental status of collocations?

- □ Epiphenomenal? (cf. Bley-Vroman 2002)
  - e.g. *dark night*
- □ BUT
  - many collocations are semi-idiomatic
  - very difficult for L2 learners

Corpus-based measures of association strength

- Raw frequency, MI, z, t, DP, conditional probability…
- Psychological reality?
  - weak correlations
  - inconsistent results

Need an appropriate measuring instrument

This paper

- □ The instrument: Words that go together well
- □ Validation study
- □ Some preliminary research

Words that go together well

"Choose the phrase that sounds the most natural or familiar"

Two examples:

- □ delicate tea
- □ feeble tea
- □ frail tea
- □ powerless tea
- ☒ weak tea

- ☒ deliver a speech
- □ hold a speech
- □ perform a speech
- □ present a speech
- □ utter a speech

Developing the test

- Initial list extracted from a dictionary of collocations (Douglas-Kozłowska and Dzierżanowska 2004)
- Their collocational status confirmed using data from the British National Corpus (overall frequency of at least 5 in the BNC **and** MI of at least 4)
- Collocations involving abstract nouns
  - idiosyncratic (avoids the *dark night* problem; difficult to construct good foils for concrete nouns)
  - fairly regular

Foils

- MI of less than 2 **and** not listed in the dictionary of collocations; the majority were also unattested in the corpus
- Synonyms of the target or of other collocates of the target; semantically and pragmatically plausible

Examples of test items

- blatant lie
- clear lie
- conspicuous lie
- distinct lie
- recognizable lie

- boost production
- double production
- enlarge production
- extend production
- redouble production

Examples of test items

- ☒ blatant lie
- □ clear lie
- □ conspicuous lie
- □ distinct lie
- □ recognizable lie

- ☒ boost production
- □ double production
- □ enlarge production
- □ extend production
- □ redouble production

The final test

- 38 items (half verb-noun, half adjective noun)
- Range of difficulty
  - frequency: mean 87, median 42, range 6-619
  - t score: mean 7.9, median 6.5, range 2.4 – 24.6
  - MI: mean 7.8, median 7.7, range 4.4-15.6
- Frequency and MI not correlated (r=0.05)

Validation study

- 62 adult native speakers of English
  - varying ages (18-60)
  - varying educational backgrounds (from no formal qualifications to doctorate)
- Part of a larger study:
  - Three linguistic tests (grammar, vocabulary, collocations)
  - Three non-linguistic tests: print exposure (Author Recognition Test), nonverbal IQ (Shipley 2 Block Design), and metalinguistic abilities (Pimsleur Language Analysis)
  - Also information about education level and reading habits

Reliability

- Test-retest:0.80
- Split half: 0.79
- Cronbach’s alpha: 72

Validity

- □ Convergent validity
  - Colloc x ART: r=0.54, p<.001
  - Colloc x Hours reading: r=0.27, p = 0.035
  - Colloc x Education: r=0.40, p=.001
  - Colloc x Age r=0.25, p= 0.048 (0.37 for under 35's)
- □ Divergent validity
  - Colloc x Blocks: r=0.21, p = 0.90

Relationship between grammar, vocabulary and collocations

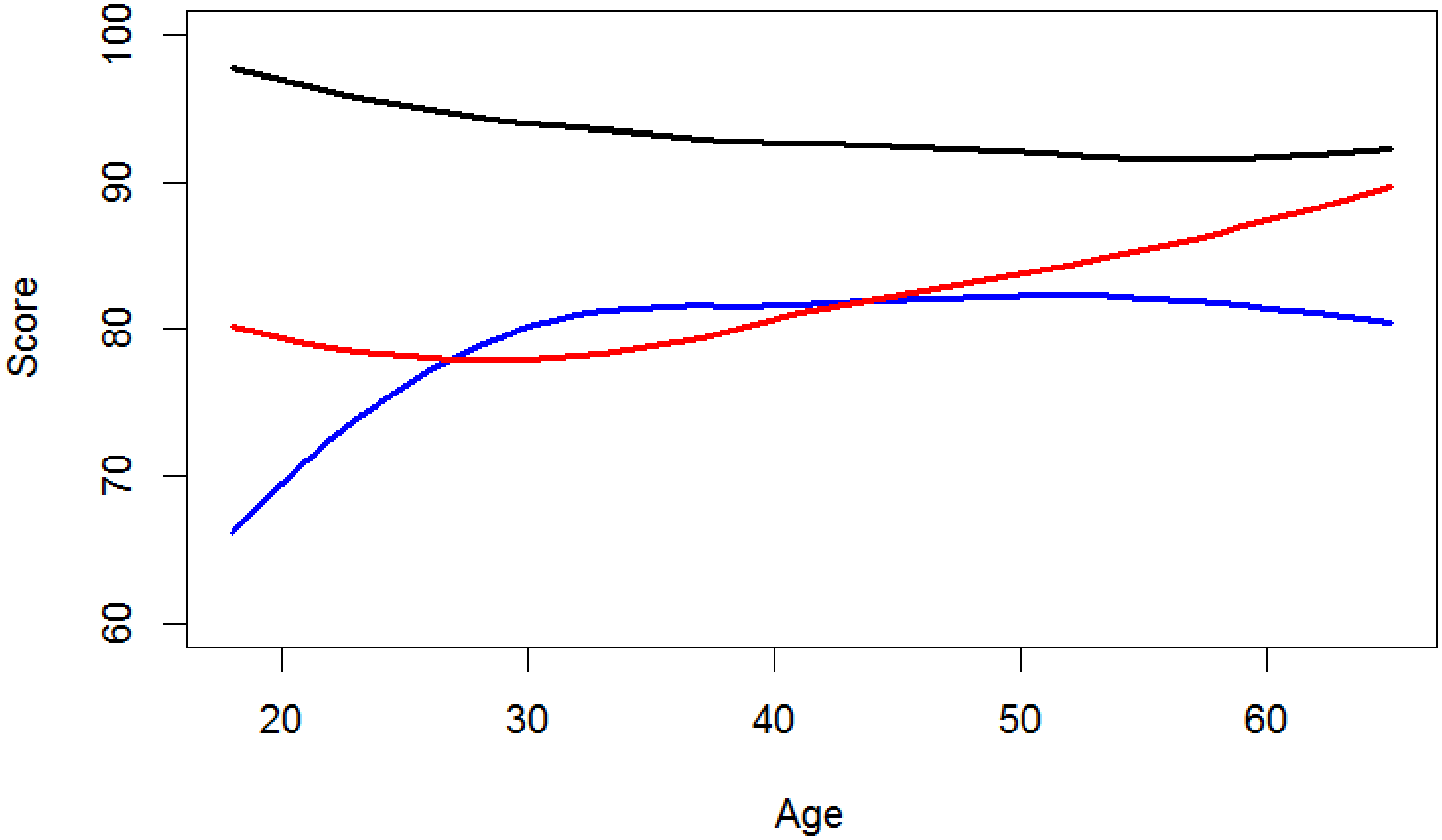
- Usage-based models: all three should be correlated
- Modular models do not predict a correlation (but don’t necessarily rule it out)
- Declarative-Procedural model: link between grammar and collocations (both involve procedural memory), no link between these two and vocabulary (declarative memory)
- Distributional learning of vocabulary: predicts correlation between collocations and vocabulary

Relationship between grammar, vocabulary and collocations

- □ Colloc x Vocab: r=0.70*** (0.40)
- □ Grammar x Vocab: r=0.46*** (0.22)
- □ Colloc x Grammar: r=0.38** (0.13)

- ✓ Usage-based theories
- ✗ Modular theories
- ✗ Declarative/Procedural model
- ✓ Distributional learning of vocabulary

Relationship between age, grammar, vocabulary and collocations



Relationship with corpus measures of collocation strength

- □ Colloc x Frequency: r=.10
- □ Colloc x z score: r=0.04
- □ Colloc x t score: r=0.10
- □ Colloc x MI: r=-0.01

Conclusions

- "Words that go together well" is a valid and reliable test of individual speakers' collocational knowledge
  - correlates with measures of linguistic experience
  - doesn't correlate with non-verbal IQ
- It does not correlate with any of the corpus-based measures of association

More conclusions

- As predicted by usage-based theories (and contra modular theories), there is a relationship between speakers' knowledge of grammar, vocabulary and collocations.
- Particularly strong relationship (0.7) between collocations and vocabulary size – in line with the hypothesis that the acquisition of non-basic vocabulary depends strongly on distributional learning mechanisms.
- Linguistic knowledge continues to develop in adulthood; the relationship between the three components changes in the course of development.

THANK
YOU