

Formulaic Language:
What Language Models Must Explain

Norbert Schmitt
University of Nottingham

Formulaic Language is a Major Component of Discourse

- Formulaic language is ubiquitous in both spoken and written discourse
- Formulaic language is the preferred realization of many language functions
- Formulaic language is involved in L1 acquisition
- Much of meaning is encoded in phrasal units, as opposed to individual words
- Formulaic language enables quicker and more efficient language usage (Pawley and Syder)
- Much syntagmatic sequencing is better accounted for by pattern descriptions than grammatical descriptions

Formulaic Language is a Major Component of Discourse

- If formulaic language is an essential component of language, (**the most essential component?**), then models of both language use and language acquisition have to account for it as a core (**not peripheral**) feature
- In particular, they will have to account for at least the following four observations:

Models Must Account For:

1. Formulaic language is ubiquitous in both spoken and written discourse
2. Formulaic language varies widely in its form and function
3. Formulaic language drives much of the sequencing of language
4. Formulaic language facilitates quicker and more efficient language processing than creatively-generated language
 - The processing advantage obtains for native speakers, but not for non-native speakers

1. Formulaic Language is Ubiquitous

- Erman and Warren (2000)
 - 58.6% of spoken English discourse
 - 52.3% of written English discourse
- Foster (2001)
 - 32.3% of unplanned native speech
- Biber's research finds lexical bundles are widespread
- If formulaic language is widespread in discourse, this implies proficient users must know a very large number of formulaic sequences
- Proficient users' mental lexicons may contain as many or more formulaic sequences as individual words

2. Formulaic Language Varies Widely

- Variation within multi-word units (Moon, 1996)
- Variation of form
 - Hang out / air out one's dirty laundry / linen
 - Touch someone with a ten-foot pole / bargepole
- Truncation / variation
 - *Every cloud has a silver lining*
 - The only *silver lining* of the disaster was that it reinforced the need for better disaster response planning.
 - He came away from the experience looking for the *silver lining in the storm cloud*.
 - The unexpected reward arrived like the *proverbial silver lining*.

2. Formulaic Language Varies Widely

- Schmitt and Carter (2004)
- Length
 - *You can lead a horse to water, but you can't make him drink.*
 - *Oh no!*
- Purpose
 - Express a meaning message or idea
 - The early bird get the worm* = do not procrastinate
 - Realize functions
 - [I'm] just looking [thanks]* = declining a shopkeeper's assistance
 - Transact technical information precisely
 - Wind 28 at 7* = wind is from 280 degrees at 7 knots speed

2. Formulaic Language Varies Widely

- Schmitt and Carter (2004)
- Fixedness
 - Fixed: *Ladies and Gentlemen; King and Queen*
 - Slots (with semantic constraints):
[someone / something] *made it plain that* [something as yet unrealized
with authority was intended or desired]
- Degree of Holistic Storage
 - Some formulaic sequences may only be stored holistically
 - Some may have dual storage (holistic + individual words/syntax)
 - In dual cases, holistic route is default?
 - Storage mode will vary from individual to individual

2. Formulaic Language Varies Widely

- Formulaic language is not a homogenous phenomenon
- Language models will have to explain not only the existence of formulaic language, but also the variation in all of the previous parameters

3. Formulaic Language as a Sequencing Principle

- The sequencing in language has traditionally been described by syntax
- But syntax is not meaning-driven
- Communication is meaning-driven
- Formulaticity is meaning-driven (form-meaning composites)
- A formulaic language perspective can often offer a more insightful analysis of sequencing patterns

3. Formulaic Language as a Sequencing Principle

- *border* = boundary or edge
- *border* + *-ed* = past tense verb
- Inflections do not change meaning (grammatical change)
- So *bordered on* should mean:
“existed on the boundary or edge of something”
- Sometimes it does:
The province bordered on a beautiful lake before water loss dried it up.

3. Formulaic Language as a Sequencing Principle

- However, this misses the big picture:

	BNC frequency	X + on	Figurative sense
border	8,011	89 (1%)	
borders	2,539	84 (3%)	
bordering	367	177 (48%)	71%
bordered	356	99 (28%)	75%

3. Formulaic Language as a Sequencing Principle

- Important usage is reporting an undesirable situation:
 - *His passion for self-improvement bordered on the pathological.*
 - *But his approach is unconscionable, bordering on criminal.*
- For further evidence of this usage, here are some other words which occur to the right of *bordered/ing on*:

a slump

a sulk

acute alcoholic poisoning

antagonism

apathy

arrogance

austerity

bad taste

blackmail

carelessness

chaos

conspiracy

contempt

cruelty

cynicism

3. Formulaic Language as a Sequencing Principle

- There is clearly a pattern here, and I would suggest that it is something like this:

[SOMETHING/ SOMEONE] (be) *bordered/bordering* on [AN UNDESIRABLE STATE (OFTEN OF MIND)]

- Contrast this with a grammatical analysis:
noun phrase + BE + *bordered/bordering* + preposition + noun phrase
- This describes the sequence, but gives no hint why it is used. That is, a grammatical description does not really tell us much about the way *bordered/bordering* are used. In contrast, the above pattern-based description tells us much more about why these words are used, and how the sequencing leads to meaning.

3. Formulaic Language as a Sequencing Principle

- While traditional syntactical descriptions are clearly useful in describing sequencing, they typically lack any semantic component which connects the sequencing to meaning and communicative utility. (Although see Larsen-Freeman's idea of grammaring and form/meaning/function)
- Looking at sequencing from a formulaic, pattern-based perspective often gives a better idea of the sequence and why it is used.
- Language models need to consider the role of formulaic language in driving sequencing.

4. Ease of Processing

- Pawley and Syder (1983) *asserted* that formulaic language is easier to process and use than creatively-generated language. This assertion has now been largely substantiated:
- Terminal words in formulaic sequences received fewer fixations and shorter durations than same words in nonformulaic language (NS - yes; NNS – mixed) (Underwood, Schmitt and Galpin, 2004)
- Formulaic sequences (both figurative and literal renderings) were read more quickly than control nonformulaic counterparts (both NS and NNS) (Conklin and Schmitt, 2008)

4. Ease of Processing

- In online grammaticality judgment experiments, both native and nonnative speakers responded to formulaic sequences faster and with fewer errors than they did to nonformulaic sequences (Jiang and Nekrasova, 2007)
- Speakers who need maximum fluency ('smooth talkers' – auctioneers, sports announcers) make heavy use of formulaic language (Kuiper, 2004)
- Two additional studies from the Centre for Research in Applied Linguistics (CRAL) at the University of Nottingham:

4. Ease of Processing

- Siyanova, A. & Conklin, K. (in submission).
Representation and processing of frequent phrases by native speakers and proficient bilinguals.
- Investigated on-line processing of English **binomial expressions** by native and proficient nonnative speakers. Binomials are frequent collocations formed by two words from the same class connected by a conjunction. For example:
 - **king and queen** (N + N)
 - **black and white** (Adj + Adj)
 - **read and write** (V + V)
 - **slowly and carefully** (Adv + Adv)

Binomials

In such expressions, a particular word order is more frequent and considered more acceptable by native speakers. For example:

- ✓ 'king and queen'
- ✓ 'black and white'
- ✓ 'read and write'
- ✓ 'slowly and carefully'

- ✗ 'queen and king'
- ✗ 'white and black'
- ✗ 'write and read'
- ✗ 'carefully and slowly'

What happens if you reverse the word order?

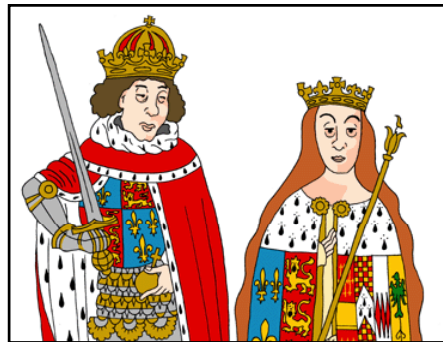
Importantly, the binomial expression and its reversed form are both **grammatically correct** and have the **same meaning**.

Mental Representation of Binomials

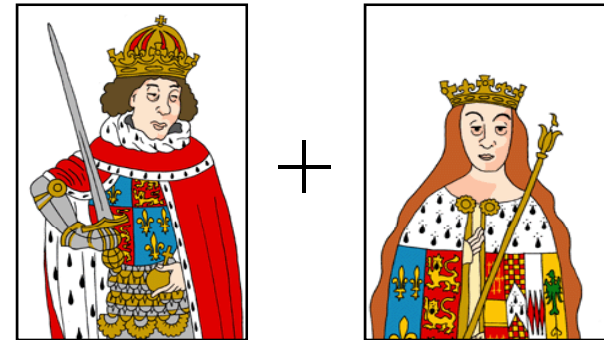
There are two ways in which binomials may be represented in the mental lexicon:

- holistically, that is, as a single unit
- as individual words

Representation



king and queen



king and queen

Research Question

Do native and proficient nonnative speakers process binomials **holistically** or in a **word-by-word** manner?

- If their processing is holistic, a processing advantage should be found for binomials (e.g. *king and queen*) but not their reversed forms (e.g. *queen and king*).
- However, if binomials are processed as individual words, no difference will be found.
- Experiment 1: reaction time
- Experiment 2: eye-tracking

Experiment 1

Participants:

- 40 native English speakers
- 40 proficient nonnative English speakers

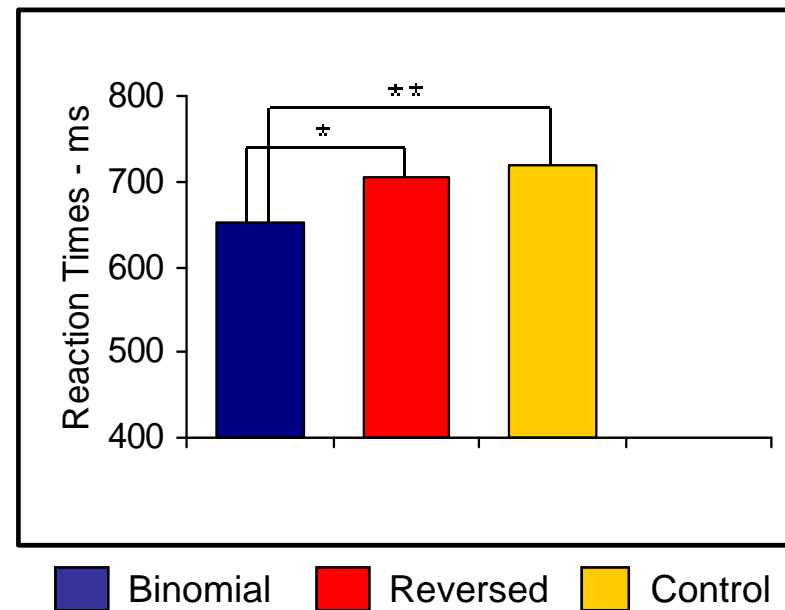
Procedure:

- Sentences were presented using phrase-by-phrase moving window paradigm. For example:

Despite the crisis the king and queen are still popular among the people.

- Reaction times were measured for the middle chunk only
- 25% of trials were followed by a comprehension question

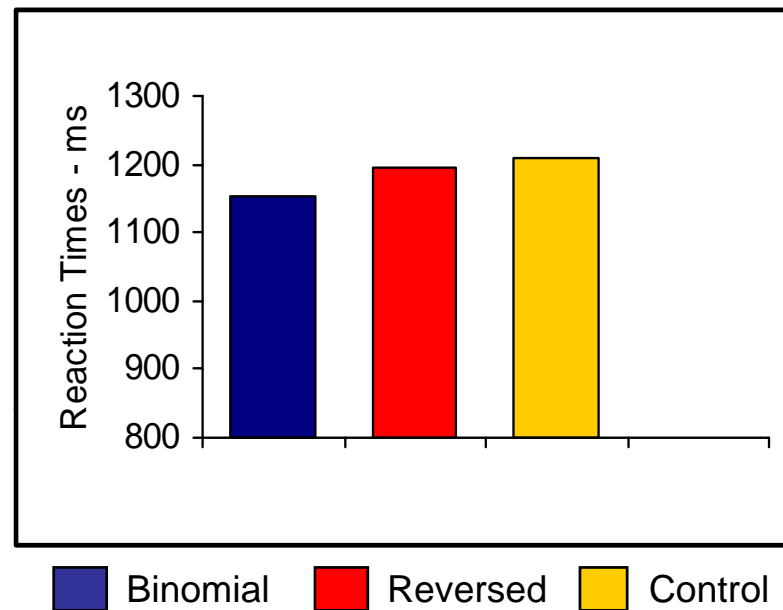
Experiment 1 - results



Native speakers showed:

- shorter response times to binomials than to reversed forms
- shorter response times to binomials than to controls
- no significant differences in response times to reversed forms and controls

Experiment 1 - results



Nonnative speakers showed:

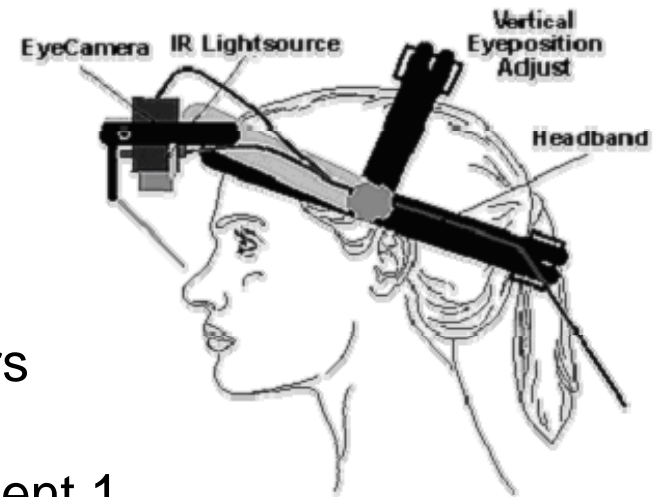
- no differences in response times to binomials and reversed forms
- no differences in response times to binomials and controls
- no differences in response times to reversed forms and controls

Experiment 2

Participants:

- 28 native English speakers
- 28 proficient nonnative English speakers

The same stimuli were used as in Experiment 1.

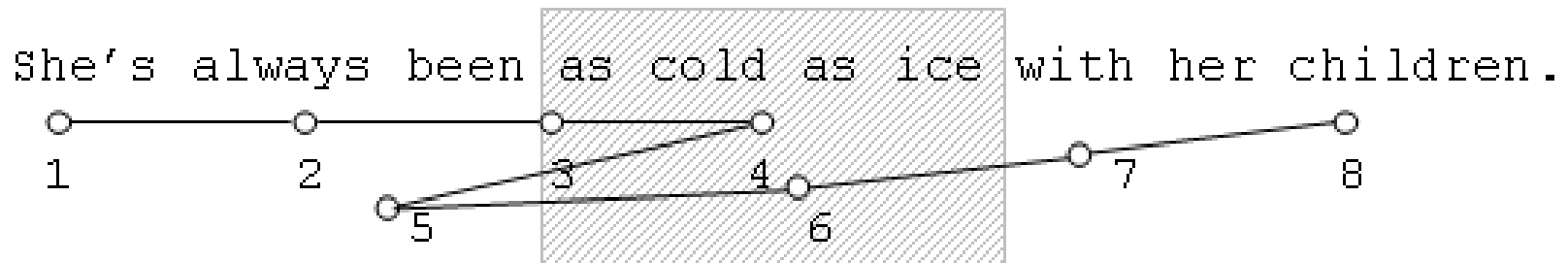


Procedure:

- whole sentences were presented
- 25% trials were followed by a comprehension question

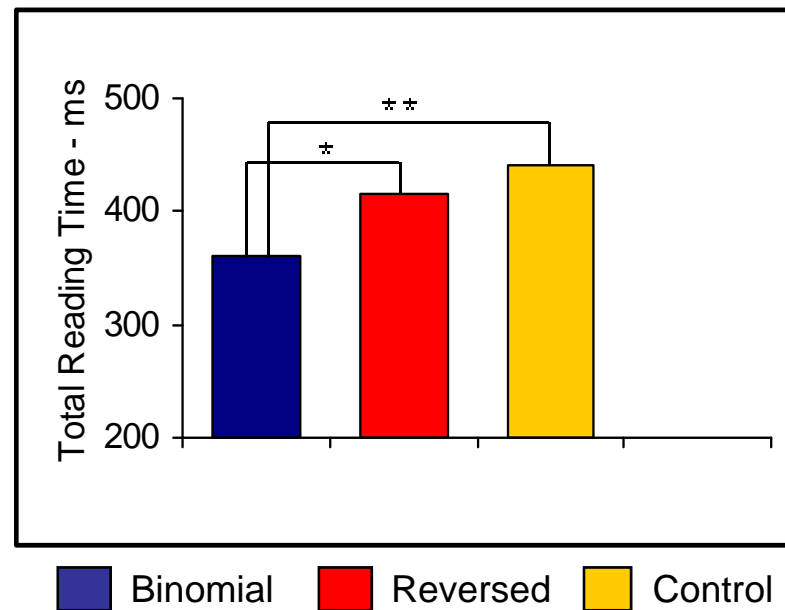
Experiment 2

We analysed **four** different eye-tracking measures:



1. Total Reading Time = 3 + 4 + 6
2. First Pass Reading Time = 3 + 4
3. Fixation Count = 3 + 4 + 6
4. Regression Path Duration = 3 + 4 + 5 + 6

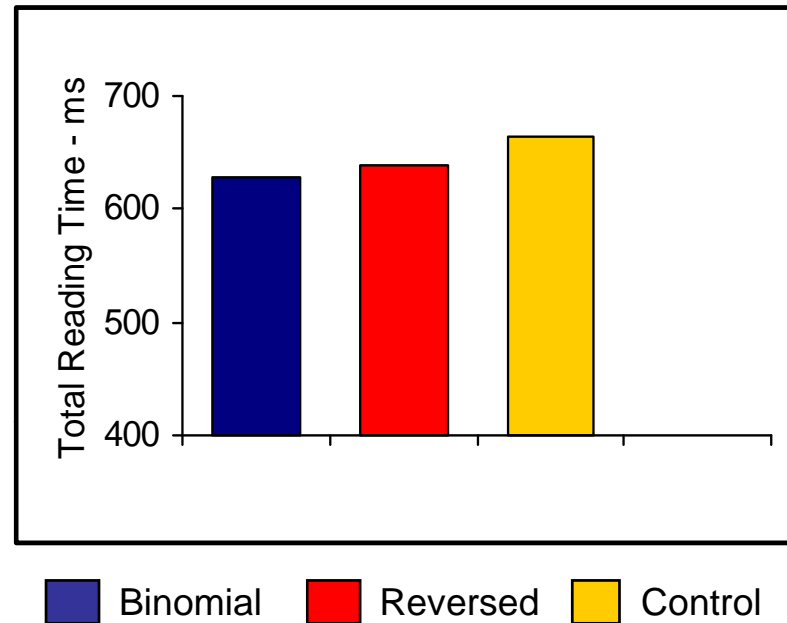
Experiment 2 - results



Native speakers showed:

- shorter response times to binomials than to reversed forms
- shorter response times to binomials than to controls
- no significant differences in response times to reversed forms and controls

Experiment 2 - results



Nonnative speakers showed:

- no differences in response times to binomials and reversed forms
- no differences in response times to binomials and controls
- no differences in response times to reversed forms and controls

Conclusions

Native speakers:

- The processing advantage displayed for binomials (e.g. 'king and queen') over their reversed forms (e.g. 'queen and king') suggests that not only the individual words but binomial phrases themselves are **represented in the lexicon** of native speakers. This means that for them, frequent phrases are a unit of lexical access and storage.

Conclusions

Nonnative speakers:

- Because no such advantage was found for nonnative speakers, it appears that they do not process these structures (e.g. 'king and queen') as a unitary whole, but rather in a **word-by-word** manner.
- We can thus argue that binomials are **not represented** in their lexicon.
- This suggests that even highly proficient nonnative speakers may not be sensitive to collocation patterns and phrasal frequencies

4. Ease of Processing

Siyanova, A., Conklin, K., & Schmitt, N. (in submission).

Processing of idioms by native speakers and proficient L2 learners: An eye-tracking study.

We investigated **native** and **nonnative** processing of an idiom's:

figurative meaning (e.g. *ring a bell* – remind)

literal meaning (e.g. *ring a bell* – produce sound with a small metal object),

novel phrase (e.g. *ring the bell*).



Research Questions

We aimed to provide a more detailed account of idiom processing in native and nonnative speakers of English when a **story context strongly biases** towards the figurative or literal idiom interpretation.

1. How do native and proficient nonnative speakers process **idioms vs. novel phrases**?
2. How do native and proficient nonnative speakers process idioms' **figurative vs. literal meanings**?

Design - Materials

- 21 idioms in highly **biasing story contexts**
- Examples of sentences:

Figurative:

“.... He told me how she used to look when we were at school, her name and even her nickname, but it didn't ring a bell at all! ...”

Literal:

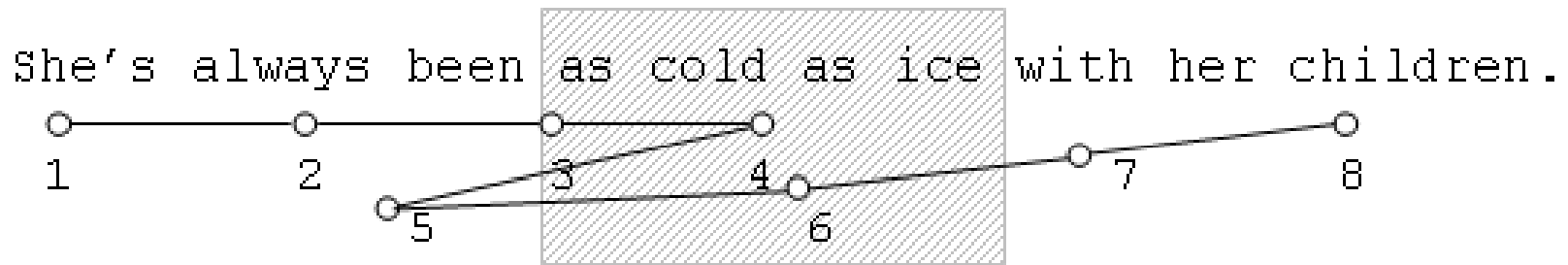
“.... We would be given a task that we had to do within a certain period of time. As soon as the time was up, the invigilator would ring a bell to let everyone know that we had to stop writing and put our pens down. ...”

Novel:

“... there's hardly anyone waiting for appointment. Every time I need to speak to a receptionist I need to ring the bell to be noticed, otherwise I may end up waiting for half an hour ...”

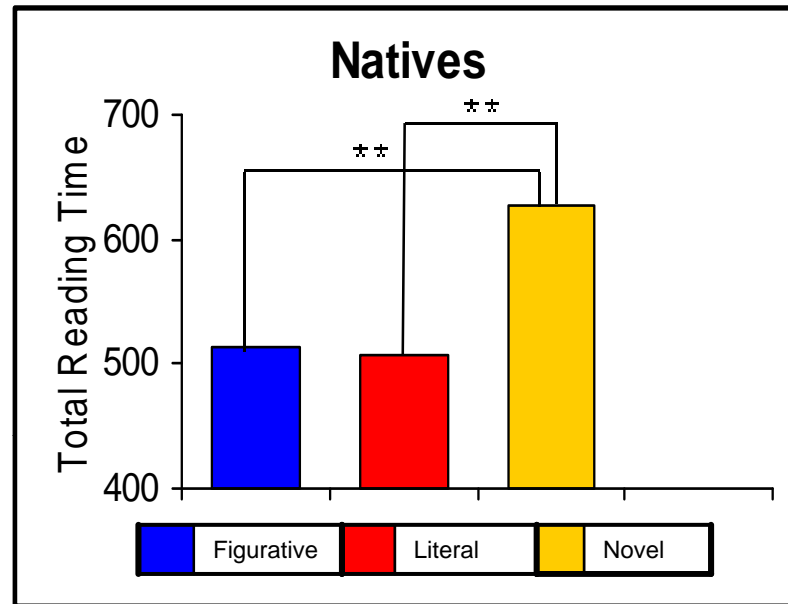
Experiment

We analysed **five** different eye-tracking measures:



1. Total Reading Time = 3 + 4 + 6
2. First Pass Reading Time = 3 + 4
3. Fixation Count = 3 + 4 + 6
4. Regression Path Duration = 3 + 4 + 5 + 6
5. Rereading = 5 + 6

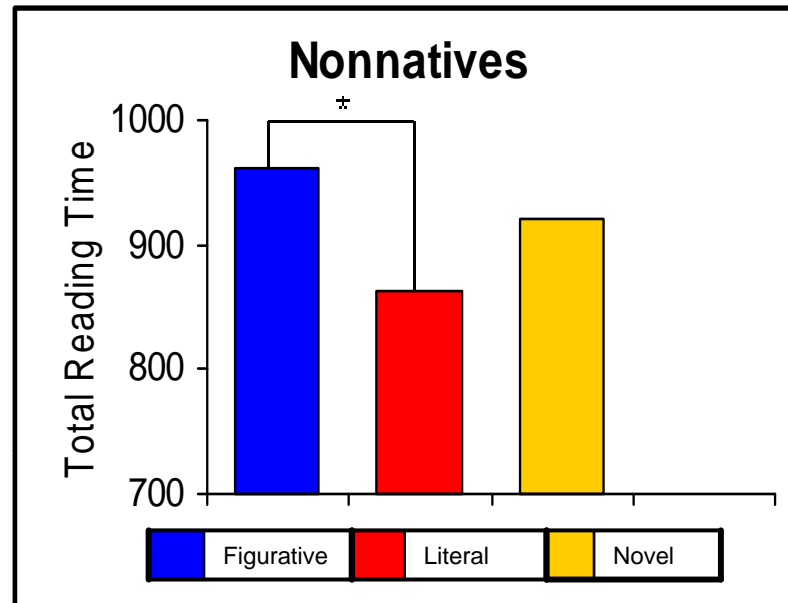
Results



Native speakers showed:

- shorter reading times to figurative uses than to novel phrases
- shorter reading times to literal uses than to novel phrases
- no differences in reading times to figurative and literal uses

Results



Nonnative group showed:

- no differences in reading times to figurative uses and novel phrases
- no differences in reading times to literal uses and novel phrases
- longer reading times to figurative uses than to literal ones

Conclusions

Native speakers

- Process idioms faster than novel language. Formulaic language is important. Native speakers rely on it all the time in both comprehension and production. It speeds up and facilitates both.
- We also found that idioms' figurative and literal interpretations were processed with the same speed.
 - This implies that the preceding disambiguating discourse plays an important role when processing phrases that have multiple interpretations.

Conclusions

Nonnative speakers

- Slow reading times for figurative uses suggest that nonnative speakers do not use **formulaic language** to the same extent as native speakers. In fact, idioms slow down comprehension due to their non-compositionality.
- Nonnatives **cannot use context effectively** to activate the appropriate (i.e. figurative) meaning, even in the presence of strong contextual cues.

4. Ease of Processing

- Natives benefit from formulaic language in terms of processing ease
- Nonnatives do not seem to share this advantage to any great degree, as they seem to have a predilection to process in a much more word-by-word basis
- Language models need to account for:
 - natives' tendency to process holistically
 - the disparity between native and nonnative (less proficient?) processing of formulaic language