

UNIVERSITY OF
BIRMINGHAM

U

When words are
not enough

B

Pernilla Danielsson, School of Humanities
Centre for Corpus Research

Identifying Multi-Word Units

- What are we identifying
- Downward/Upward Method
- Human Identification
- Zipf's Orderliness of Distribution of Words and Units
- Corpus Hub At Birmingham

When words are not enough...

- Words, as units of analysis, have the advantage of being clearly delimited in texts and stable.
- The shortcomings of words become obvious when you try to assign a meaning to the textual representation of a word. Most words represent several meanings. Despite this fact, they rarely cause disambiguating problems to humans.

A different viewpoint

It is not so much that language is fuzzy and ambiguous, but that the units of description are inadequately identified.

Starting from the viewpoint of multi-word units, is to say that we are reconsidering the notion of a word as the basic unit of analysis in language, and instead propose that in an information and meaning carrying system the unit of analysis should be a unit of meaning which may consist of more than one word.

The starting point: Collocations

- J.R Firth was first to identify the importance of words in context

"[...] I propose to bring forward as a technical term, meaning by 'collocation', and to apply the test of 'collocability'." (Firth 1957[1951]:194)

Meaning, however, is tricky. It is pre-theoretical, yet we often want to theorise around it. And we find ourselves stuck in patterns of discussing '*prototypical meanings*'

Collocates as a linguistic tools mainly thanks to John Sinclair

In Corpus Linguistics *collocation* and *collocate* may be used as mere labels indicating empirical facts. As such, counting collocates are simple tools to work with in a stage previous to judgment or concluding statements.

Are all collocates important?

Mike Scott (1998): *“The literature on collocation has never distinguished very satisfactorily between collocates which we think of as associated with a word (letter - stamp) on the one hand, and on the other, the words which do actually co-occur with the word (letter - my, this, a, etc.).*

We could call the first type coherence collocates and the second neighbourhood collocates or horizon collocates. It has been suggested that to detect coherence collocates is very tricky, as once we start looking beyond a horizon of about 4 or 5 words on either side, we get so many words that there is more noise than signal in the system”.

Significant/strong/habitual collocations

- *significant collocation* (McEnery & Wilson 1996: 71; Hunston 2002: 70),
- *strong collocation* (Hunston 2002: 71),
- *habitual collocation* (Fernando 1996: 30-33)
coherence collocation (Scott 1998).

The linguistic tools become computational tools

- The first tools identifying multi-word units in corpora seem to share the following characteristics:
 - Pre-set length (Bigram, Trigram)
 - Adjacent words
 - Statistical Calculations

Automatic Extraction of Meaningful units

Stroke (1558)

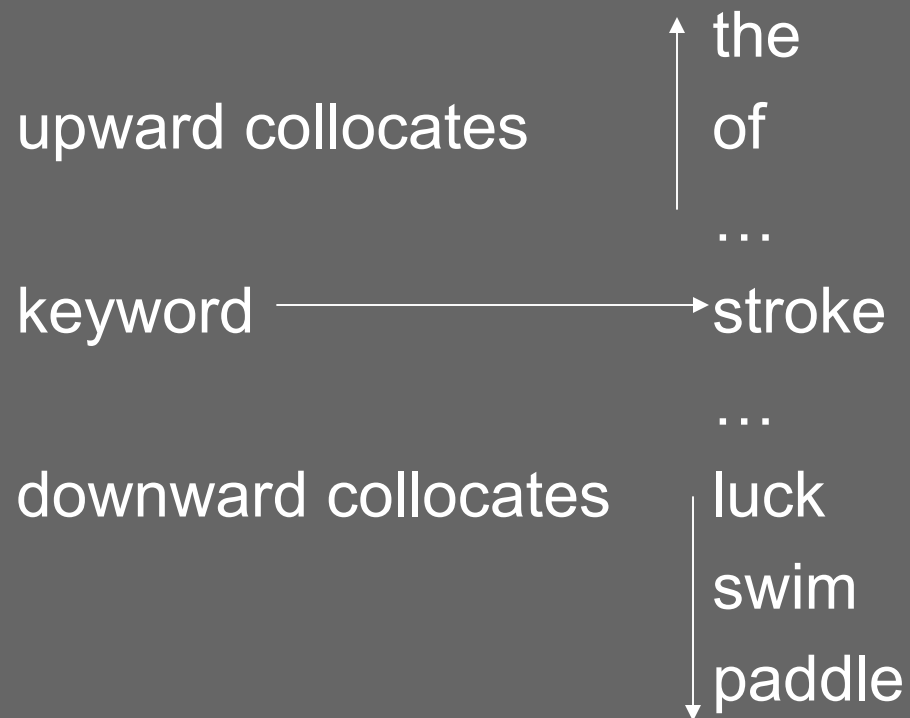
A stroke (407)

At a stroke (90)

The hypothesis

A language consists of a set of units of meaning, where (a) each unit of meaning is made up of one or more word forms, (b) each prototypical unit of meaning has at least one realisation in text which is sufficiently distinctive to differentiate it operationally from all the others, (c) each prototypical unit of meaning may also be abstracted into a descriptive unit, a unit of structure, and, (d) all the information required to identify the units of meaning and make abstractions into units of structure is retrievable from the patterns of occurrence and distribution of word forms in the language text.

Upward and Downward Collocates



Automating Human Discovery Procedures

Downward

Collocates:

half-time

paddle

breast

genius

paralysed

hypertension

sweep

beard

carers

[DC: stroke half-time](UC: the of on) .

[DC: stroke half-time](UC: the on) .

[DC: stroke half-time](UC: the) .

[DC: stroke half-time](UC:) .

Out of signs...

[DC: stroke paddle](UC: the) .

[DC: stroke paddle](UC:) .

Out of signs...

[DC: stroke breast](UC:) .

Out of signs...

[DC: stroke genius](UC: of a) .

[DC: stroke genius](UC: of) .

[DC: stroke genius](UC:) .

Findings

the stroke of half-time (22 occ)
on the stroke of half-time (22 occ)
the paddle stroke (5 occ)
breast stroke (12 occ)
a stroke of genius (7 occ)
paralysed a stroke (5 occ)
the sweep stroke (5 occ)
stroke beard (6 occ)
one stroke adrift (5 occ)
swim stroke (7 occ)
putt stroke (6 occ)
a paralytic stroke (3 occ)

Paradigmatic Variations on: On the stroke of half-time

[(on 22)] the stroke of half-time
on [(the 22)] stroke of half-time
on the stroke [(of 22)] half-time
on the stroke of [(half-time 22) (half 6)
(midnight 5) (the 3) (full 2) (three 2)
(lunch 2) (injury-time 2) (time 1) (each
1) (twelve 1) (two 1) (nine 1) (tea 1)
(noon 1) (stumps 1) (eleven 1)]

Paradigmatic Variations on: A stroke of genius

a stroke of genius 7

[(a 5) (my 1) (every 1) (super 1)]
stroke of genius

a stroke [(of 5)] genius

a stroke of [(luck 25) (good 6) (the 5)
(genius 5) (work 4) (his 2) (about 2)
(absolute 1) (fortune 1) (18 1) (royal
1) (ill-luck 1) (great 1) (some 1)
(coincidence 1) (brilliance 1)
(puddephat's 1)]

Syntagmatic variations

- A [great, remarkable, huge, supreme, fantastic, wonderful] stroke of luck

Remaining problems with MWUs

Are all frequently occurring combinations of words valid MWUs?

Where do the MWUs begin and end?

How do we recognise what is accepted compositionally within a unit?

Jam

- *traffic jam*
- *bread and jam*
- *butter and jam*
- *strawberry jam*
- *spread jam*
- *apricot jam*
- *a jam sandwich*
- *raspberry jam*
- *toast and strawberry jam*
- *toast and jam*
- *jam and cream*
- *jam or marmalade*
- *jam tarts*
- *blackcurrant jam*
- *pot of jam*
- *cake with jam*
- *jam sandwiches*
- *empty jam jars*
- *tomato jam*
- *jam session*

Prototypical meaning: Whatever you get if your do a Google Image search



Sounds like a wasp/hornet in a jam jar

Sounds like

a good idea

a lot of work

*a load of [rubbish, rhubarb,
gobbledegook]*

a recipe for [disaster, boredom]

a [dream, nightmare]

a contradiction in terms

Chunks and prefabs

Looking systematically at a lot of linguistic evidence points more and more to the conclusion that people's communicative performance, in a higher degree than had ever been thought, leans on the fact of having stored prefabricated sequences of words — “chunks” of language, informally —, and that, therefore, the use of “rules” of the “system” is, to a certain extent, not very operative.

What we are looking for depends on
(at least) two factors:

- Experience/corpus
- Application

Erman, B. & Warren, B. (2000)

The idiom principle and the open choice principle. Text 20(1), pp 29-62

- 'A prefab is a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization'

Sinclair, J. 1991.

Corpus, Concordance, Collocation. Oxford:
Oxford University Press.

- The **open-choice principle** sees language as a result of many complex choices. '*At each point where a unit is completed (a word, phrase, clause), a large range of choice opens up and the only restraint is grammaticalness.*' (Sinclair 1991:109)
- You may think of this as taking the next word in a sentence and see if you could have chosen any other word within the same (or similar) word class.

The idiom-principle

- **The idiom principle** treats language as a combination of specific words into large prefabricated chunks, which makes it more difficult to define. *'At its simplest, the principle of idiom can be seen in the apparently simultaneous choice of two words, for example, of course. This phrase operates effectively as one word [...]'.*

Not really two competing models

- Sinclair goes on to state that '*The open choice analysis could be imagined as an analytical process which goes on in principle all the time, but whose results are only intermittently called for.*'(Sinclair 1991:114)

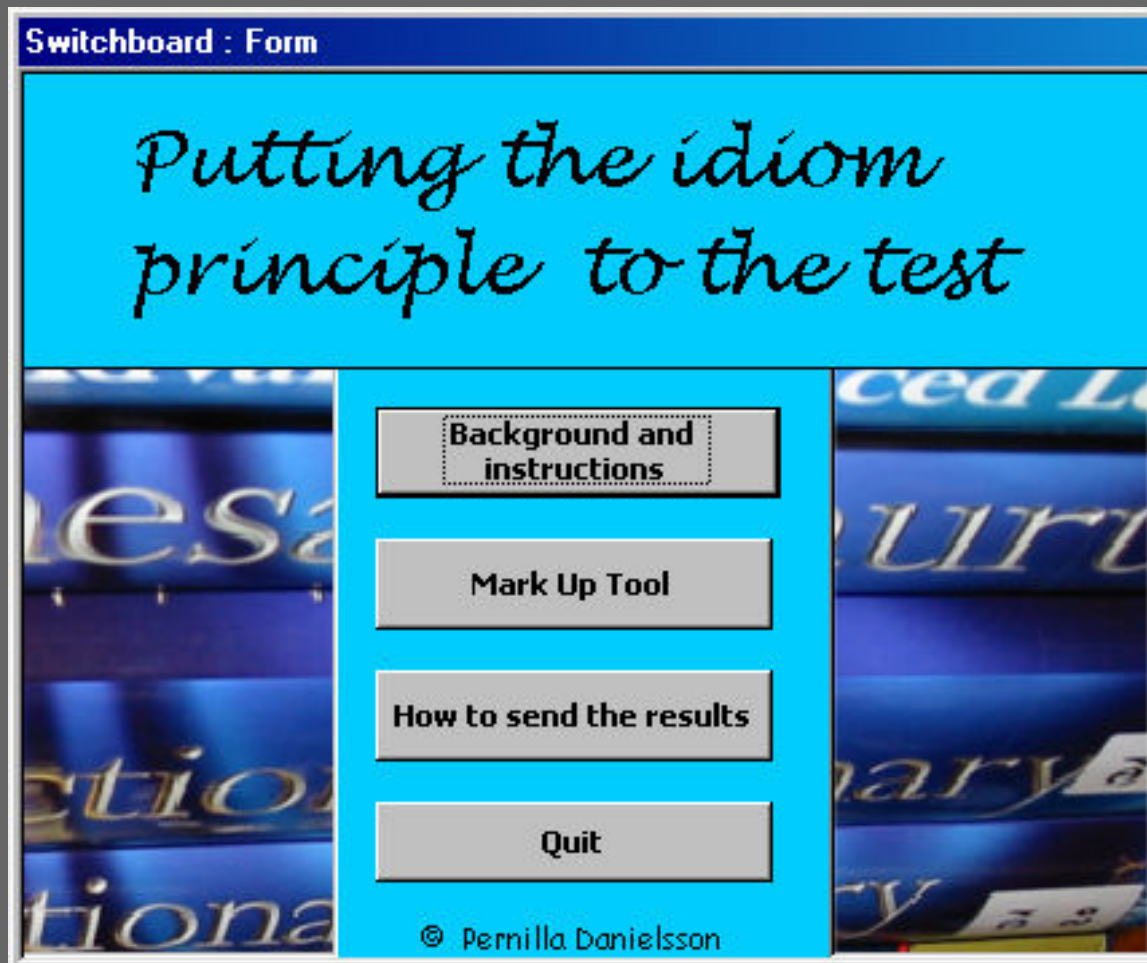
Quantifying prefabs

- Erman and Warren (2000) tried to quantify the proportions between the open-choice and the idiom principle in authentic texts. As much as 55% was said to belong to the idiom principle, whereas only 45% agreed to the open-choice model.
- This turns linguistic theories on their head, as no theory can yet fully incorporate larger chunks into the system.

The problem:

- 'Prefabs, it must be remembered, are not like phonemes and morphemes, or noun phrases and verb phrases. They are probabilistic, some more than others.'
- Of course, this goes straight back to the linguistic/computational tools about collocates: it is based on frequency of the distribution of words

“Le tool”



Postgraduate mark-up

THE world's biggest squad ran out of inspiration,
organisation and even players last night, when
they were effectively reduced to ten men
by injury and Claudio Ranieri's compulsion to
make substitutions.

the	has	be	NODE	to	a	to
and	to	been	NODE	the	the	of
<p>	have	a	NODE	by	to	the
of	will	was	NODE	<p>	of	and
s	had	is	NODE	from	and	a
to	<p>	and	NODE	and	pound	<p>
will	and	have	NODE	in	<p>	by
</h>	is	has	NODE	its	for	in
have	a	were	NODE	their	in	for
that	be	are	NODE	rate	s	men
a	of	already	NODE	or	two	per
in	should	greatly	NODE	his	number	s
has	the	the	NODE	but	that	2
which	can	had	NODE	price	dollar	on
for	at	being	NODE	it	one	is
can	would	of	NODE	fat	an	it
is	that	now	NODE	rates	by	with
it	could	or	NODE	prices	10	more
would	are	much	NODE	as	but	from
had	we	significan	NODE	if	on	are
was	it	with	NODE	cost	or	years
they	s	ve	NODE	that	deficit	but
with	was	<p>	NODE	on	arrears	1

"to". Tot freq:1375856. Freq as coll:560. t-sc:20.8797. MI:3.0875. `?' for



	has	been	NODE	to	a	men
	had	be	NODE		the	and
	have	was	NODE		pound	the
	to	were	NODE		two	by
y	is	is	NODE		10	of
	will	now	NODE		rubble	years
	the	are	NODE		dollar	for
l	a	being	NODE		just	more
	would	match	NODE		tears	overs
h	and	and	NODE		three	<p>
	were	<p>	NODE		one	a
s	should	re	NODE		12	single
t	can	costs	NODE		an	as
	are	later	NODE		four	2
e	they	either	NODE		16	minimum
se	was	men	NODE		ten	in
ch	he	years	NODE		six	or
	this	even	NODE		nothing	few
e	it	then	NODE		14	to
	could	team	NODE		five	than
	we	lead	NODE		<KPD>	when
	with	often	NODE		less	per
il	s	thus	NODE		eight	5

e". Tot freq:2872094. Freq as coll:26. t-sc:-0.3946. MI:-0.1075. `?' for he

Telnet - titania.cobuild.collins.co.uk

connect Edit Terminal Help

e	they	were	NODE	to	10	men
ter	and		NODE		a	of
st	areas		NODE		nine	and
>	crowd		NODE		one	by
ch	we		NODE		14	the
d	indies		NODE		piles	to
	two		NODE		the	for
notes	somerset		NODE		being	20
ny	these		NODE		12	in
	support		NODE		four	at
	other		NODE		just	s
s	course		NODE		daily	gas
	who		NODE		wearing	as
en	side		NODE		giving	when
cal	prices		NODE		13	more
ese	nsw		NODE		six	fit
om	fishermen		NODE		catching	davi
t	charges		NODE		42	hand
re	london		NODE		94	whet
nancial	buildings		NODE		rubble	5
th	ways		NODE		fielding	60
ze	officials		NODE		120	comm
wn	leaders		NODE		concrete	she

he". Tot freq:2872094. Freq as coll:4. t-sc:0.7745. MI:0.7067. `?' f

In Swedish, we don't.....

- Looking in a Swedish corpus tells us that we don't talk about “*reducerat till tio man*”, we talk about played or played with only
- *Spelade med bara tio man till slutsingnalen*

Units in Translation

- According to Dorothy Kenny (2004) experienced translators identify larger units of meaning than beginners and amateur translators

Right than wrong

The Italian has got

more decisions right than wrong

this season

he does, he's **more** often right than wrong." <xr> 8047 </xr> <h1> S
badly. We're doing **more** right than wrong." But Villa's final pass
bad job. We're doing **more** right than wrong." Villa had emerged from
he had been **more** often right than wrong. But he was soon to find
results were **more** often right than wrong, but who can tell? He of
<p> The general was **more** right than wrong for the singer had alrea
more of the assignment right than wrong. It offers good design,
DIY that's easier to get right than wrong. <p> Sagbag No sulky tee

Usually a core of two words

- Accused of
 - Are accused of
 - Are often accused of
 - Are often accused of being
 - Are [often] accused of being NASTY
-
- Tot up
 - Tot up the cost of
 - Tot up the cost of any

Our figures

First test, 16 participants.

In total 2281 words marked up, out of which 1153 were mark-ed as part of a prefab

51,5% prefabs

But the full range was between **14%-89%**

Second test, 6 participants.

Out of 804 words, 536 were marked as part of prefabs

66,6% prefabs

The full range this time was between **46%-82%**

What did we agree on?

- We were most likely to agree on as prefabs:
 - 2 word combinations (exceptions: ran out of, the cost of) (laws of probability?)
 - Lexical prefabs
 - With clear semantic status
- Words everyone agreed on **not** to be prefabs:
 - surprisingly few:
 - As, his, they, had, their
 - Almost all content words were marked by someone as part of a larger unit

Subjective choices or objective units

- There may be a learning curve involved in identifying them
- Chunking may be more basic than 'unitizing'

George Kingsley Zipf

Zipf introduced the Orderliness of Distribution of Words

$$a b^2 = k$$

$$a (b*b)=k$$

Where a is the number of words at a given frequency and b is the number of occurrences (the actual frequency)

The Orderliness of Distribution of Words

betraktade 48

vilka 47

vapen 47

tala 47

polis 47

fyra 47

borgström 47

tyst 46

a = 6 occurrences

b = 47

$6(47 \cdot 47) = 13254$

And further down the same list....

både	40
ännu	39
tjänst	39
tio	39
sam	39
peter	39
känna	39
ihop	39
förstås	39
dels	39
såna	38



a = 9 occurrences

b = 39

$9 (39 \cdot 39) = 13689$

What does this means?

- Zipf discovered a mathematical regularity amongst the orderliness of distribution of words. And, he found that the characteristics of a frequency list applies almost regardless of languages, genre, text type, sample size etc.

What does the Zipf curve tell us?

- A) a relation between the frequency of occurrences of an event and the number of different events occurring within that list
- B) a relation between a frequency and its rank in the frequency list

Now to the interesting part

Zipf is famous for having a less accepted analysis of his own findings, but he could also be famous for what he couldn't show. Namely he couldn't prove this regularity to hold all the way up to the top of the frequency list. In fact the top 2% does not agree to this regularity. WHY?

The top 2 %

The word *the* occurs 4290 times in a text. The size of K is here approximately 4200, giving this formulae:

$$a * (4290*4290=18404100) = 4200$$

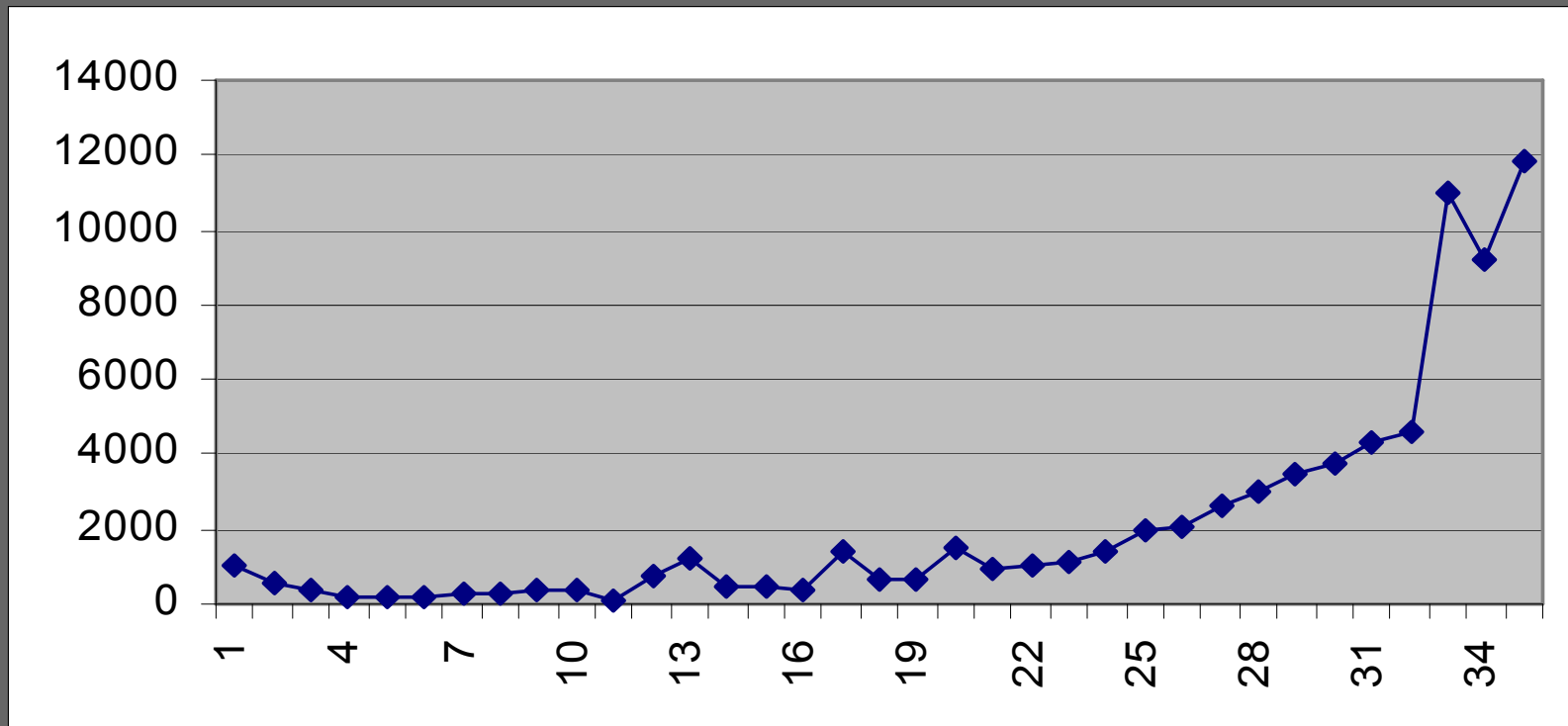
(a = the number of words at this frequency)

The quantity, *a*, then, which represents the number of words occurring 4290 times would be $4200/(4290)^2$; that is, *the*, in English would represent .000025 of a word – **a very absurd statement no matter how a word is defined.**

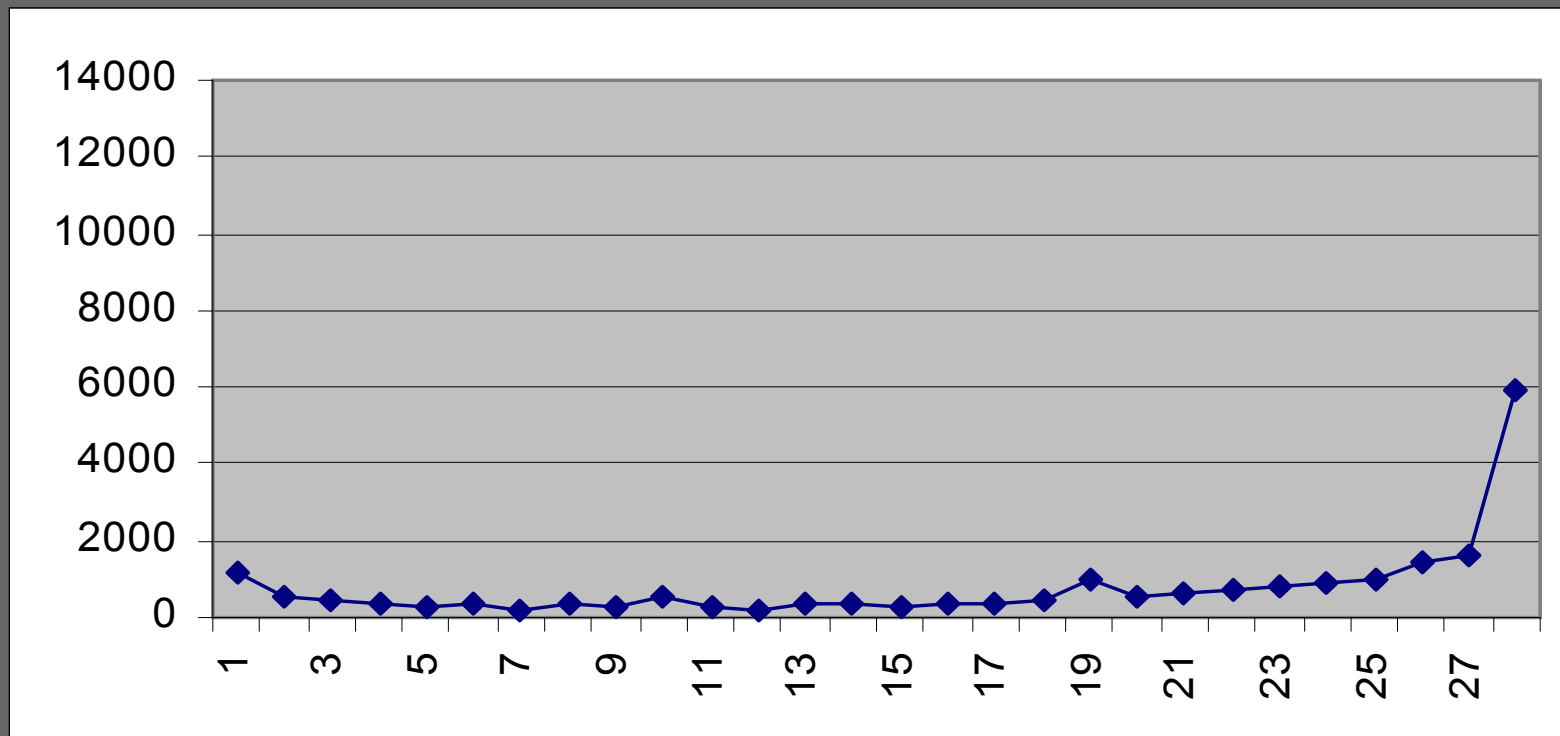
Unitized text

När han <kom tillbaka> <gick [han] till sängs> och läste men låg <långa stunder> och <tittade på> henne. Hon <sov djupt> när <telefonen ringde>. När han väckte henne och gav henne luren hade hon <svårt att förstå> <vad det <rörde sig om>>. <<Klockan [var nästan> ett] på natten>. Han <gick ut> och <satte på tevattnen> och <rostade bröd>. Hon brukade <må illa> <när [hon] vaknade> och var tom i magen. När han kom in med teet pratade hon fortfarande. Hon sa egentligen inte mycket. Lugnade nån. <Lyckades [kanske för] till slut> <la [hon] på luren> och <kröp ihop> igen.

Before...



After



So...

- If you intend to use a statistical measurement that assumes normal distribution, you are bound to run into problems with the top 2% of a frequency list
- The result shown here suggests that if our units of analysis were MWUs rather than words, Zipf's ODWs may prove to hold the whole frequency list through

CHAB –

Corpus Hub At Birmingham

Danielsson, P & Sawyers, A (forthcoming) Corpus Hub at Birmingham: Concordance software for the analytic linguist

The screenshot shows a web browser window with the address bar displaying "http://127.0.0.1/chab/public_html/" and a search query "race condition". The main content area is titled "CORPUS HUB AT BIRMINGHAM" and displays a concordance for the word "brink". The text is highlighted in yellow, and a "More" link is visible next to each occurrence. On the left side, there is a navigation menu with "Find" and "Analysis" options. Below "Analysis", there are three radio buttons: "Text select tool", "Line select tool" (which is selected), and "Cluster select tool". A "Line select tool" panel is open, showing a list of words: "catastrophe", "river", and "catastrophe". The "river" word is selected, and there are buttons for "Add", "Remove", "Hide these lines", and "Show only these lines".

www.corpus.bham.ac.uk/CHAB

(available from June)

Storm

later described the incident as "a storm in a teacup", was admonished. A student from a Portland high school died in a storm in 1986. The three climbers from the Saturday afternoon was a real storm in a tea cup. But by the time change and adjust. During a strong storm in January 1987, the Nauset Beach and the Gold Coast Meter Maids. The storm in a DD-cup rivalry has even

<p> As with Operation Desert Storm in 1991, Operation Allied Force to tender. This might all seem a storm in a teacup - after all Dunkley you could dismiss the incident as a storm in a teacup: naive student shoes were lost overboard during a storm in May 1990, 800 kilometres

latest row is just a silly season storm in a cappuccino. The ITC will p As far as I am concerned this is a storm in a beer mug. In an increasingly their train - take conservatism by storm in 1994, rather than 1984, or

The Kansas-Nebraska Act Raises a Storm In January 1854, Senator Stephen MPU-401 took the computer world by storm in 1984 as the first standard f Tech. <p> Hal Zeron: I remember one storm in the '40s. I believe it was a power plant. A major geomagnetic storm in 1989 caused a blackout in Ca overtake. <p> He added: "That was a storm in a bloody tea cup. OK, so they

UNIVERSITY OF
BIRMINGHAM

CHAB clustering:

CUP=[teacup, tea cup, DD-cup, cappuccino, beer mug, bloody tea cup]

later described the incident as `a storm in a **teacup**", was admonished. From a Portland high school died in a storm in 1986. The three climbers for the Saturday afternoon was a real storm in a **tea cup**. But by the time change and adjust. During a strong storm in January 1987, the Nauset Beachy Gold Coast Meter Maids. The storm in a **DD-cup** rivalry has even

<p> As with Operation Desert Storm in 1991, Operation Allied Force to tender. This might all seem a storm in a **teacup** - after all Dunkley you could dismiss the incident as a storm in a **teacup**: naive student shoes were lost overboard during a storm in May 1990, 800 kilometres latest row is just a silly season storm in a **cappuccino**. The ITC will p As far as I am concerned this is a storm in a **beer mug**. In an increasingly their train - take conservatism by storm in 1994, rather than 1984, or

The Kansas-Nebraska Act Raises a Storm In January 1854, Senator Stephen MPU-401 took the computer world by storm in 1984 as the first standard f Tech. <p> Hal Zeron: I remember one storm in the `40s. I believe it was a power plant. A major geomagnetic storm in 1989 caused a blackout in Ca overtake. <p> He added: `That was a storm in a **bloody tea cup**. OK, so they

CHAB unitizing: STORM='storm in a CUP'

later described the incident as `a storm in a teacup", was admonished. the Saturday afternoon was a real storm in a tea cup. But by the time aunchy Gold Coast Meter Maids. The storm in a DD-cup rivalry has even to tender. This might all seem a storm in a teacup - after all Dunkley ou could dismiss the incident as a storm in a teacup: naive student latest row is just a silly season storm in a cappuccino. The ITC will p As far as I am concerned this is a storm in a beer mug. In an increasingl overtake. <p> He added: `That was a storm in a bloody tea cup. OK, so they

Summing Up

- Recognising that words may not serve as the best units of analysis in language is a good starting point
- We still haven't answered what constitutes a multi-word unit is, nor where it begins and ends. This seem to be a problem that is shared between humans an computers.
- New research has shown that rather than validating our MWUs around the idea of 'prototypical meaning', we may choose to look at the units as prefabricated units, hence probabilistic.
- Erman and Warren's claim of above 50% of words in a text belonging to the idiom-principle seem to be confirmed or even strengthened in the study done at Birmingham
- There is still a fair amount of discrepancy between each annotator's units and a learning curve seems to be involved: Experience?
- Nonetheless, the multi-word units are concrete. If we apply MWUs to text as a basis for the calculation of Zipf's curve then we see indications that they may hold the key also to regularities of distribution