Interactive practical workshop 22/01/2016

Automatic identification of formulaic sequences in (fairly) big data: Practical introduction to a procedure

Andreas Buerki
Cardiff University

In this workshop, I will present an automatic procedure for extracting formulaic sequences from corpus data and guide participants through its practical implementation using example data and software tools. By the end of the workshop, participants will be able to use the N-Gram Processor (Buerki 2013) and the software SubString (Buerki 2011) to extract formulaic sequences from corpus data of their own. Participants will also be aware of some of the strengths and weaknesses of the procedure and its theoretical underpinnings. The workshop is divided into three parts.

The first part addresses the question of how (or even whether) extraction procedures relate to theoretical understandings of formulaic sequences. While the procedure presented takes as its starting point a constructionist view of formulaic sequences, which identifies them as units of form and associated meaning that are conventional in a speech community, this understanding is briefly located within a broader context of thinking on the nature of formulaic sequences. Implications for identification procedures, including of views based on psycholinguistic processing, the traditional phraseological criterion triplet of polylexicality, idiomaticity and fixedness or the frequency-only approach that produces lexical bundles will also be discussed.

In part two of the workshop, participants are invited to work through a hands-on example of how formulaic sequences are automatically extracted from corpus materials following the five-stage extraction procedure outlined in Buerki (2012):

- Data preparation (normalisations, formatting)
- N-gram extraction using the N-Gram Processor (including the use of stop-lists)
- Consolidation of different length n-grams to derive a unified list using SubString
- Filtering (application of frequency thresholds and a lexico-structural filter)
- Assessment of accuracy and recall

This includes an introduction to the installation and use of the necessary open-source software tools. A corpus of Wikipedia texts will be provided as example data.

In the final part of the workshop, strengths and limitations of the procedure will be discussed as well as potential alternatives. Strengths include the methodological transparency of the procedure and the ability to process large amounts of corpus data (subject to sufficiently powerful hardware); the limitations consist mainly of the flipside of this, namely that it is less accurate as an automatic procedure when applied to small amounts of data (< 1 million words). In a final discussion section, participants are invited to share their views on any aspect of the workshop topic including how remaining challenges might be overcome.

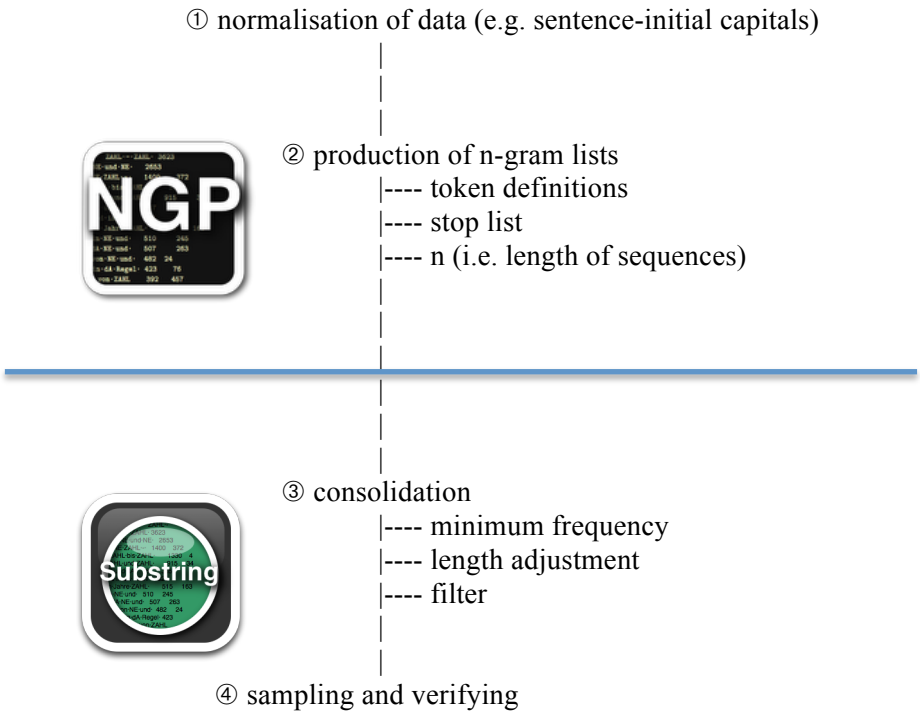
References

Buerki, A. (2013). *N-Gram processor 0.4* [Computer Software]. Available at http://buerki.github.io/ngramprocessor/

Buerki, A. (2011). *SubString* [Computer Software]. Available at http://buerki.github.com/SubString/

Buerki, A. (2012). Korpusgeleitete Extraktion von Mehrwortsequenzen aus (diachronen) Korpora. In N. Filatkina, A. Kleine-Engel, M. Dräger, & H. Burger (Eds.), *Aspekte der historischen Phraseologie und Phraseographie* (pp. 263-92). Heidelberg: Universitätsverlag Winter.

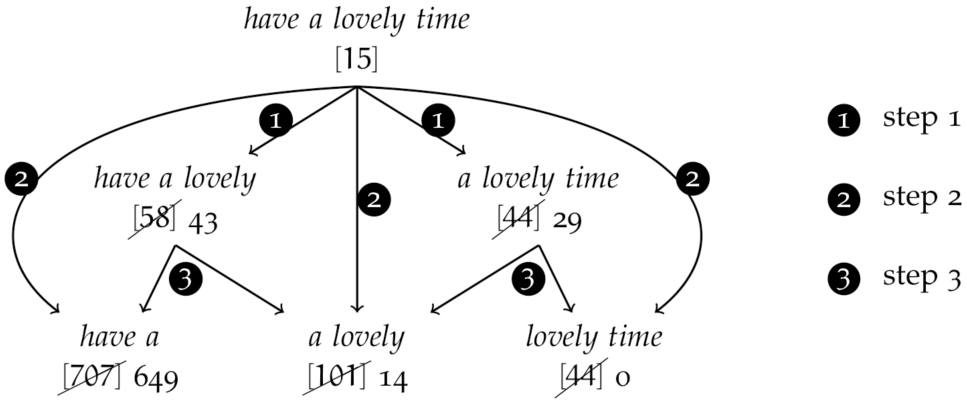
Stages of Extraction



Additive stoplist: top 200 most frequent words of English based on the Leipzig word lists (http://wortschatz.uni-leipzig.de/Papers/top1000en.txt)

I	city	governm	like	our	some	us
a	come	ent	lot	out	state	very
about	company	group	made	over	still	want
according	could	had	make	part	such	was
after	county	has	man	people	take	way
against	day	have	many	per	team	we
all	did	he	market	percent	than	week
also	do	help	may	play	that	well
although	don	her	me	points	the	were
an	down	here	million	police	their	what
and	during	high	more	public	them	when
another	each	him	most	put	then	where
any	end	his	much	re	there	which
are	even	home	my	report	these	while
around	family	how	need	right	they	who
as	few	i	new	run	think	will
at	first	if	next	said	third	win
back	five	in	night	same	this	with
be	for	including	no	say	those	won
because	former	into	not	says	three	work
been	found	is	now	school	through	world
before	four	it	of	season	time	would
being	from	its	off	second	to	year
between	game	just	officials	see	today	years
both	get	know	on	set	told	you
business	go	last	one	she	too	you
but	going	lead	only	should	took	
by	good	least	or	since	two	
can	got	left	other	so	under	
					up	

Frequency consolidation and substring reduction



(1)a

have a lovely time	15
have a lovely	58
a lovely time	44
have a	707
a lovely	101
lovely time	44

(1)b

have a lovely time	15
have a lovely	43
a lovely time	29
have a	649
a lovely	14
lovely time	0

References:

Altenberg, B. and M. Eeg-Olofsson. (1990). 'Phraseology in spoken Eng- lish: presentation of a project'. In J. Aarts and W. Meijs (eds). *Theory and practice in corpus linguistics.* Amsterdam: Rodopi. 1-26.

O'Donnell, M.B. (2011). 'The adjusted frequency list: A method to produce cluster-sensitive frequency lists' *ICAME Journal*, 35(April). <http://icame.uib.no/ij35/Matthew Brook ODonnell.pdf> [accessed 21 June 2013]

Length adjustment

dA zum Beispiel	[65]
einA zum Beispiel	[17]
es zum Beispiel	[17]
man zum Beispiel	[20]
so zum Beispiel	[54]
wie zum Beispiel	[68]
zum Beispiel	[451]
zum Beispiel NE	[70]
zum Beispiel NUM	[31]
zum Beispiel auf	[16]
zum Beispiel bei	[20]
zum Beispiel dass	[16]
zum Beispiel durch	[20]
zum Beispiel für	[26]
zum Beispiel mit	[37]

Lexico-structural filter

(Substring v. 0.9.8)

```
\/·       |\
^%·'·|\
^%·NE·
          |\
^&·NE·|\
^&·amp·
          |\
^'·[^·]·  |\
^(NE·)+
          |\
^(NUM·)+
          |\
^\(·\)|·\(·\)|\
^-·|\
^-·|^HYPH·|^—
·|^–·|\
^/·(NUM·)+/·|\
^/·/·|·/·/·|\
^/·NE|\
^NE·|\
^NUM·\(·|\
^NUM·\)·|\
^[^·]·'·  |\
^\+·NE|\
^\+·\+·NE|\
^_|\
^·|\
·'·       |^'·|\
·-·       |\
·-·       |·HYPH·
          |·—·
          |·–·
          |\
·/·       |^/·|\
·_|\
·–·       |\
^
          [[:digit:]]
* |\
^and·|^\+·|\
^but·|\
^for·the·[^·]*·
          |\
^from·the·|\
·he·      |\
^people·|\
·she·     |\
^than·|\
^that·|\
^their·|\
^them·|\
·they·    |\
^the·[^·]*·of·
          |\
^to·[^·]*·the·
          |\
^we·|\
^what·|\
·when·    |\
·where·   |\
^which·[^·]*·
          |\
^[^·]*·which·
          |\
^who·|\
^you·     |\
·and·     |\
·a·       |\
·by·      |\
·for      |\
·from·    |\
·her·     |\
·he·      |\
·his·     |\
·in·      |\
·its·     |\
·it·      |\
·my·      |\
·our·     |\
·their·   |\
·the·     |\
·to·      |\
·your·    |\
·been·    |\
·be·      |\
·are·[^·]*·
          |\
·had·     |\
·has·     |\
·have·    |\
·is·      |\
·was·     |\
^had·
```