Automatic N-Gram Analysis on the Basis of Biber et al.'s (1999) Lexical Bundle Categories

Work-in-progress report

Benedikt Heller (KU Leuven)

KU LEUVEN

ICAME 35, Nottingham, 01/05/2014

Overview

1. Introduction

2. Research goals

3. Methodology

4. Preliminary results

Lexical bundles

- *I don't know* (3-Gram), *at the end of* (4-Gram), *you know what I mean* (5-Gram)
- "Lexical bundles are identified empirically, as the combinations of words that in fact recur most commonly […]" (Biber et al. 1999: 992)

Lexical bundles

- Restrictions (cf. Biber et al. 1999: Ch. 13)
  - At least 10 times per 1 million words
  - Spread across five texts

- Some lexical bundles occur more often in certain registers (*conversation*, *academic prose*) (cf. Biber et al. 1999: Ch. 13)

I don't know what (typical in *conversation*)
the nature of the (typical in *academic prose*)

Lexical bundle categories

No.	Category	*conversation*	*academic prose*	Example
1	*personal pronoun + lexical verb phrase*	44	0	*I don't know what*
2	*pronoun/NP + be*	8	2	*it was in the*
3	*active verb*	13	0	*have a look at*
4	*yes-no and wh-question fragment*	12	0	*can I have a*
5	*wh-clause fragment*	4	0	*know what I mean*
6	*NP with post-modifier*	4	30	*the nature of the*
7	*preposition + NP fragment*	3	33	*as a result of*
8	*anticipatory it + VP/adjectiveP*	0	9	*it is possible to*
9	*passive verb + PP fragment*	0	6	*it is based on the*
10	*that-clause fragment*	1	5	*should be noted that*
11	*to-clause fragment*	5	9	*are likely to be*
12	*other expressions*	6	6	
	Total	**100**	**100**	

(Biber et al. 1999: 996)

Research goals

1. Hypothesis testing

   H1: *Using Biber et al's (1999) classification, (corpus)texts can be placed on a meaningful continuum between* conversation *and* academic prose

2. Development of a Perl script that performs the classification of lexical bundles in a (corpus)text automatically

Methodology

- POS tagging of (corpus)texts with CLAWS (cf. Garside & Smith 1997)
- Creation of regular expressions for each of the 12 lexical bundle categories
- Creation of a graphical user interface that is easily usable and displays the rather complex results nicely

The main program

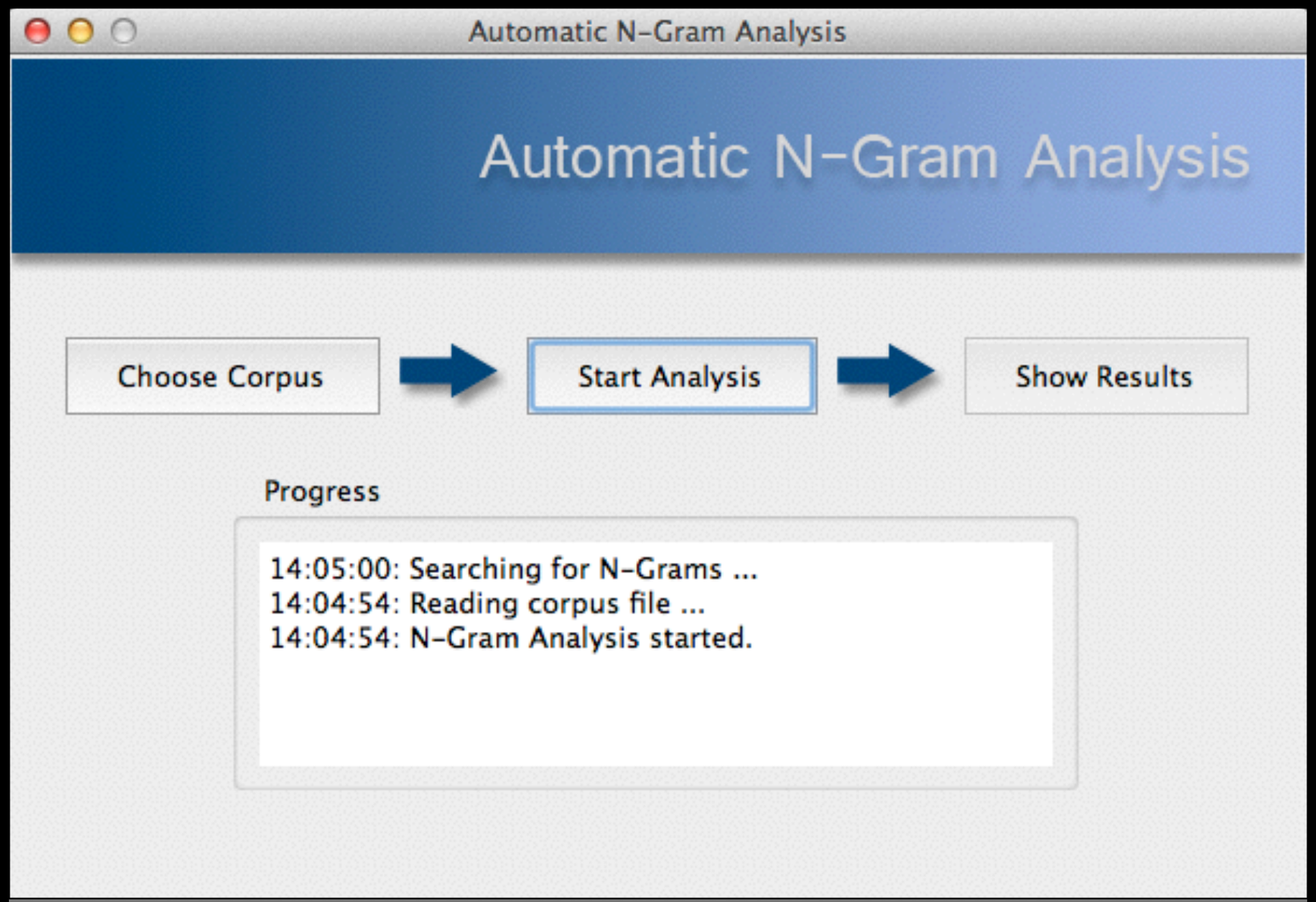


Figure 1: Automatic N-Gram Analysis user interface

Options

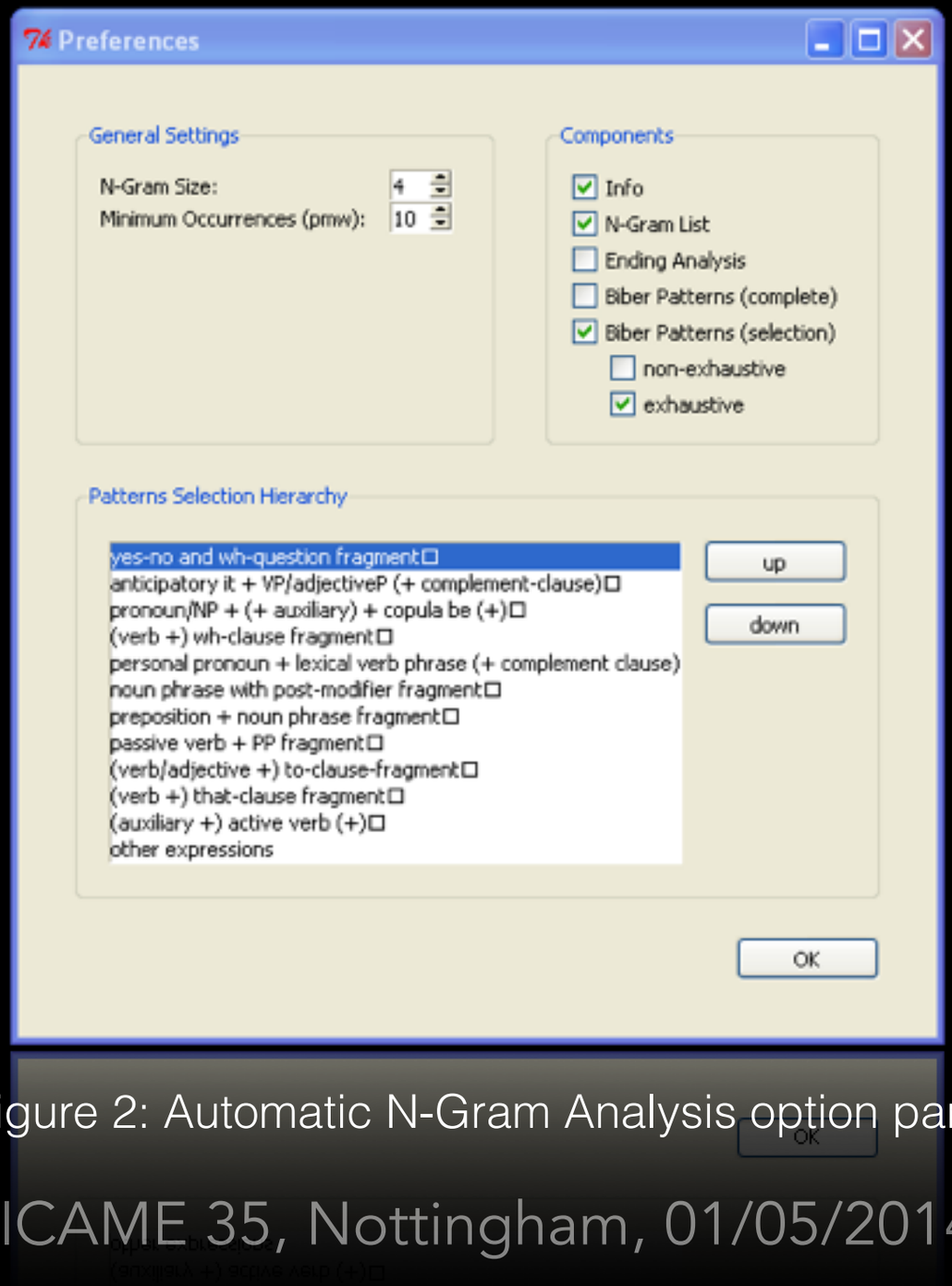


Figure 2: Automatic N-Gram Analysis option panel

Browser-based output

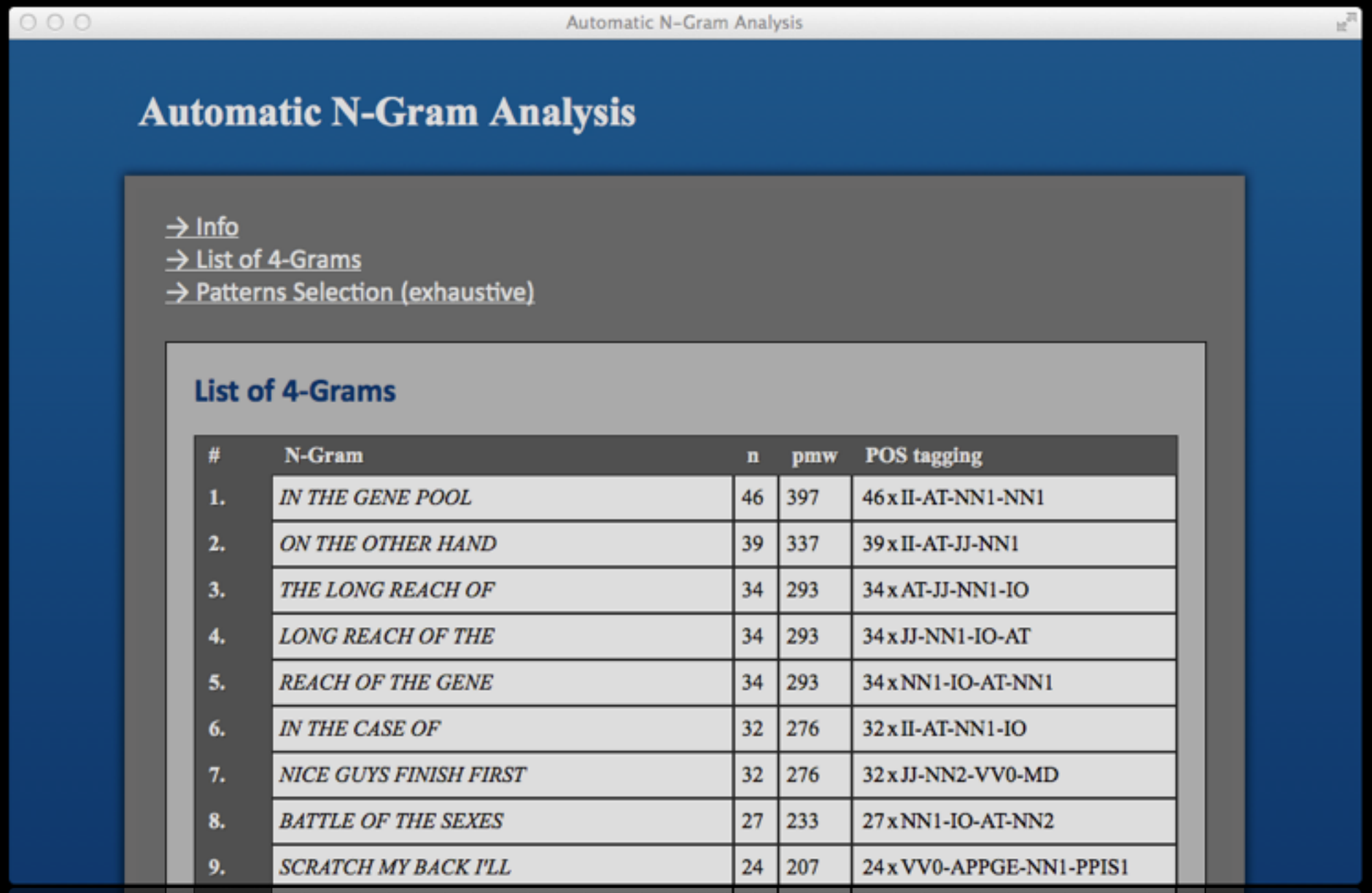


Figure 3: Automatic N-Gram Analysis 4-gram list

Browser-based output

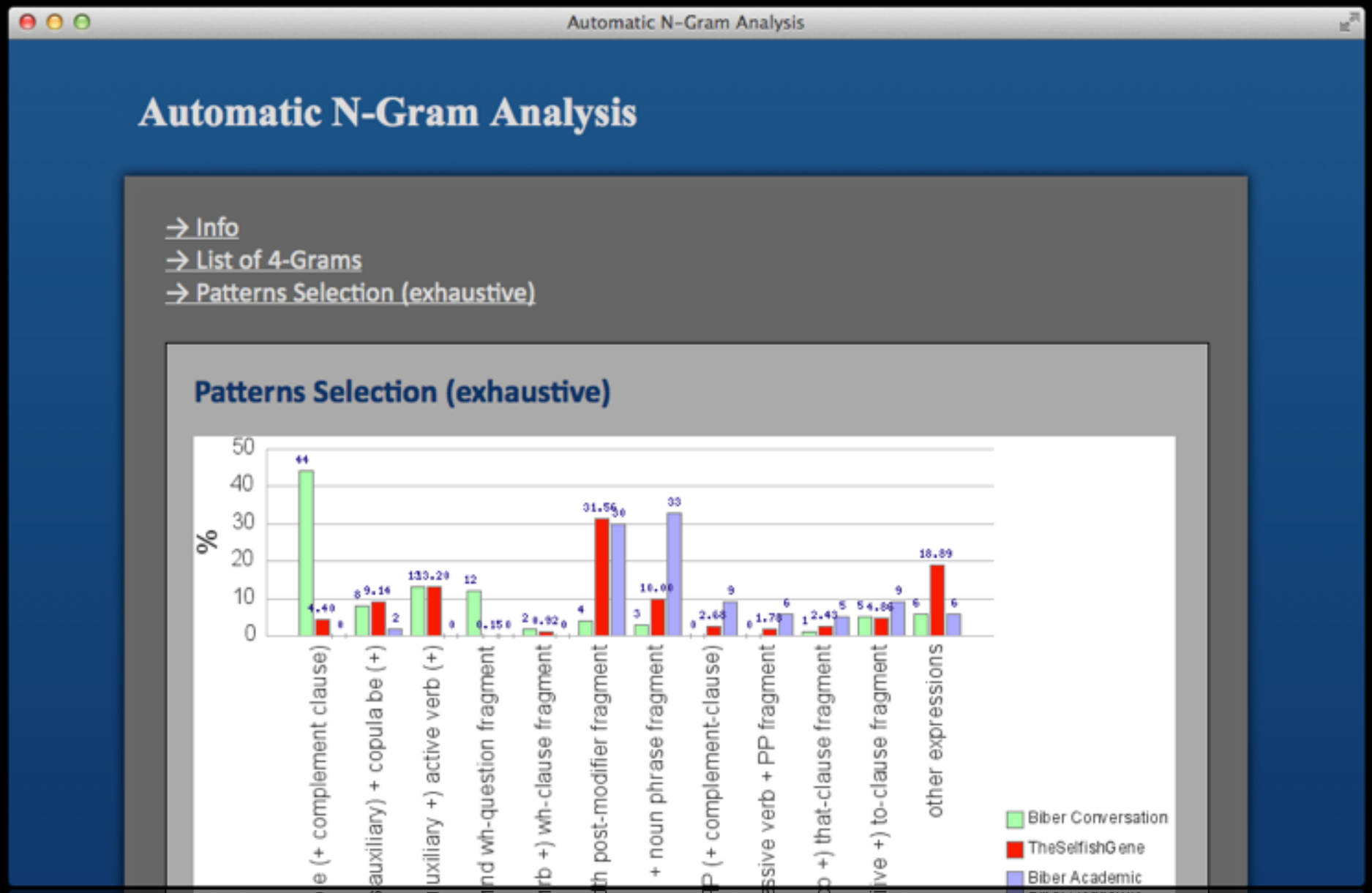


Figure 4: Automatic N-Gram Analysis pattern graph

How the automatic classification works

- Example: *pronoun/noun phrase + copula be*

  1. *pronoun*

  2. *noun phrase*

  3. *be*

- Which *tag* combinations fit into the lexical bundle categories?

Example: pronoun/NP + be – Part 1: pronoun

- Easy to find with the help of the POS tags
- Search for tags *PN* or *PN1*, or *PNQO*, etc.

NP	proper noun, neutral for number (e.g. IBM, Andes)
NP1	singular proper noun (e.g. London, Jane, Frederick)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	singular weekday noun (e.g. Sunday)
NPD2	plural weekday noun (e.g. Sundays)
NPM1	singular month noun (e.g. October)
NPM2	plural month noun (e.g. Octobers)
PN	indefinite pronoun, neutral for number (none)
PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	objective wh-pronoun (whom)
PNQS	subjective wh-pronoun (who)
PNQV	wh-ever pronoun (whoever)
PNX1	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHO2	3rd person plural objective personal pronoun (them)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPHS2	3rd person plural subjective personal pronoun (they)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	2nd person personal pronoun (you)
RA	adverb, after nominal head (e.g. else, galore)
REX	adverb introducing appositional constructions (namely, e.g.)
RG	degree adverb (very, so, too)
RGQ	wh- degree adverb (how)
RGQV	wh-ever degree adverb (however)
RGR	comparative degree adverb (more, less)

Figure 5: List of C7 tags

Example: pronoun/NP + be – Part 2: noun phrase

- Simplified structure of the English noun phrase:

(determiners)	(pre-modifiers)	noun	(post-modifiers)
a	*new*	*edition*	*of the book*
some	*large*	*sheets*	*of paper*
the	*old*	*man*	*who lives near us*

(Greenbaum & Nelson 2009: 66)

Example: pronoun/NP + be – Part 3: copula be

- Again, easy to find (*be*, *am*, *are*, *is*, *being*, *was*, *were*, and *been*)

Automatic classification

- Search for the pattern *pronoun/noun phrase + copula be*:

```
$pattern =~ /($POS{"pronoun"}|
$POS{"noun_phrase"})-(\w+-)*$POS{"be"}/
```

  - OR relationship between *pronoun* and *noun phrase*
  - after that, an optional number (also 0) of other words
  - finally, a form of *to be*

```
"pronoun"       =>
     qr/(EX          # Existential there   (cf. Biber et al. 1999: 1005)
        |PN          # indefinite pronoun, neutral for number (none)
        |PN1         # indefinite pronoun, singular (e.g. anyone, everything,
                     # nobody, one)
        |PNQO        # objective wh-pronoun (whom)
        |PNQS        # subjective wh-pronoun (who)
        |PNQV        # wh-ever pronoun (whoever)
        |PNX1        # reflexive indefinite pronoun (oneself)
        |PPGE        # nominal possessive personal pronoun (e.g. mine, yours)
        |PPH1        # 3rd person sing. neuter personal pronoun (it)
        |PPHO1  # 3rd person sing. objective personal pronoun (him, her)
        |PPHO2  # 3rd person plural objective personal pronoun (them)
        |PPHS1  # 3rd person sing. subjective personal pronoun (he, she)
        |PPHS2  # 3rd person plural subjective personal pronoun (they)
        |PPIO1  # 1st person sing. objective personal pronoun (me)
        |PPIO2  # 1st person plural objective personal pronoun (us)
        |PPIS1  # 1st person sing. subjective personal pronoun (I)
        |PPIS2  # 1st person plural subjective personal pronoun (we)
        |PPX1        # singular reflexive personal pronoun (e.g. yourself,
                     # itself)
        |PPX2        # plural reflexive personal pronoun (e.g. yourselves,
                     # themselves)
        |PPY)        # 2nd person personal pronoun (you)
   /x,
```

```
"noun_phrase"    =>
        qr/      # determiner
                 (
                         $determiner
                         (CC-$determiner)*
                 )*
                 # pre-modifiers
                 (
                         $pre_modifier
                         (CC-$pre_modifier)*
                 )*
                 # noun
                 (
                         (ND1     # singular noun of direction (e.g. north, south
                         |NN      # common noun, neutral for number (e.g. sheep,
                                  # headquarters)
                         |NN1     # singular common noun
                         |NN2     # plural common noun (
                         |NNA     # following noun of ti
                         |NNB     # preceding noun of ti
                         |NNL1    # singular locative no
                         |NNL2    # plural locative noun
                         |NNO     # numeral noun, neutra
                                  # hundred)
                         |NNO2    # numeral noun, plural
                         |NNT1    # temporal noun, singu
                         |NNT2    # temporal noun, plura
                         |NNU     # unit of measurement,
                                  # cc)
                         |NNU1    # singular unit of mea
                         |NNU2    # plural unit of measu
                         |NP      # proper noun, neutral
                         |NP1     # singular proper noun
                         |NP2     # plural proper noun (
                         |NPD1    # singular weekday nou
                         |NPD2    # plural weekday noun
                         |NPM1    # singular month noun
                         |NPM2)   # plural month noun (e
                 )+
        /x,
```

```
my $determiner =
        qr/
        # predeterminer
                (
                        (DB      # before determiner or pre-determiner capable of
                                 # pronominal function (all, half)
                        |DB2)-   # plural before-determiner ( both)
                )*
                # central determiner
                (
                  (APPGE # possessive pronoun, pre-nominal (e.g. my, your, our)
                  |AT    # article (e.g. the, no)
                  |AT1   # singular article (e.g. a, an, every)
                  |DD1   # singular determiner (e.g. this, that, another)
                  |DD2)- # plural determiner ( these,those)
                )*
                # post-determiner
                (
                  (DA              # after-determiner or post-determiner capable of
                                   # pronominal function (e.g. such, former, same)
                  |DA1             # singular after-determiner (e.g. little, much)
                  |DA2             # plural after-determiner (e.g. few, several, many)
                  minerminer (e.g. more, less,

                  iner (e.g. most, least,
```

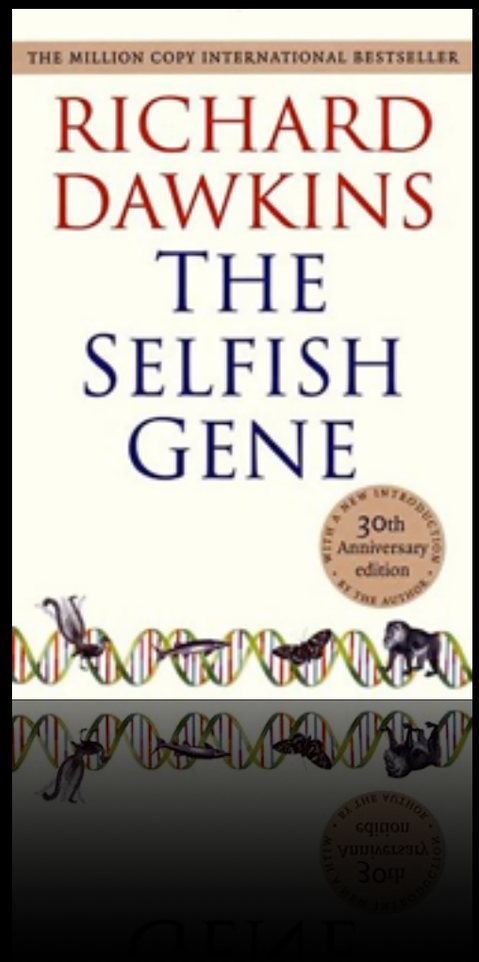
```
my $pre_modifier =
        qr/     # adverb
                (
                        (RGR     # comparative degree adverb (more, less)
                        |RGT     # superlative degree adverb (most, least)
                        |RR)-    # general adverb
                )*

                # adjective
                (
                        (JJ      # general adjective
                        |JJR     # general comparative adjective (e.g. older, better,
                                 # stronger)
                        |JJT     # general superlative adjective (e.g. oldest, best,
                                 # strongest)
                        |MC      # cardinal number,neutral for number (two, three..)
                        |MC1     # singular cardinal number (one)
                        |MD)-    # ordinal number (e.g. first, second, next, last)
                )*
 /x;
```

```
"be"      =>
    qr/(VB0   # be, base form
        |VBDR # were
        |VBDZ # was
        |VBG  # being
        |VBI  # be, infinitive (To be or not...)
        |VBM  # am
        |VBN  # been
        |VBR  # are
        |VBZ) # is
    /x,
```

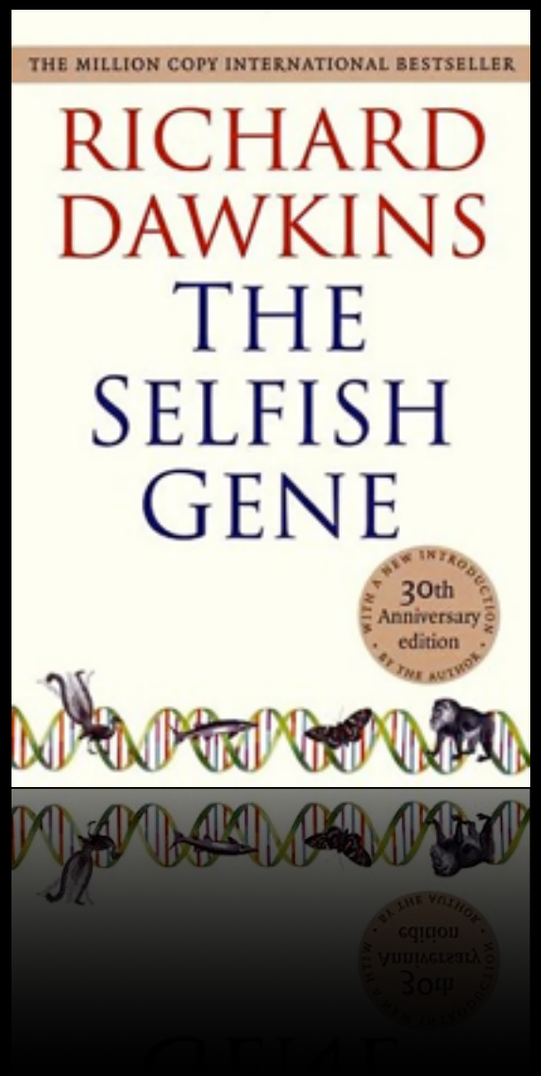
Example: *The Selfish Gene*

- Popular scientific book about evolutionary biology/psychology
- Published in 2006 [1976]



Example: *The Selfish Gene*

An uneasy tension disturbs the heart of the selfish gene theory. It is the tension between gene and individual body as fundamental agent of life. On the one hand we have the [...]



Example: *The Selfish Gene*

- The ten most frequent 4-Grams

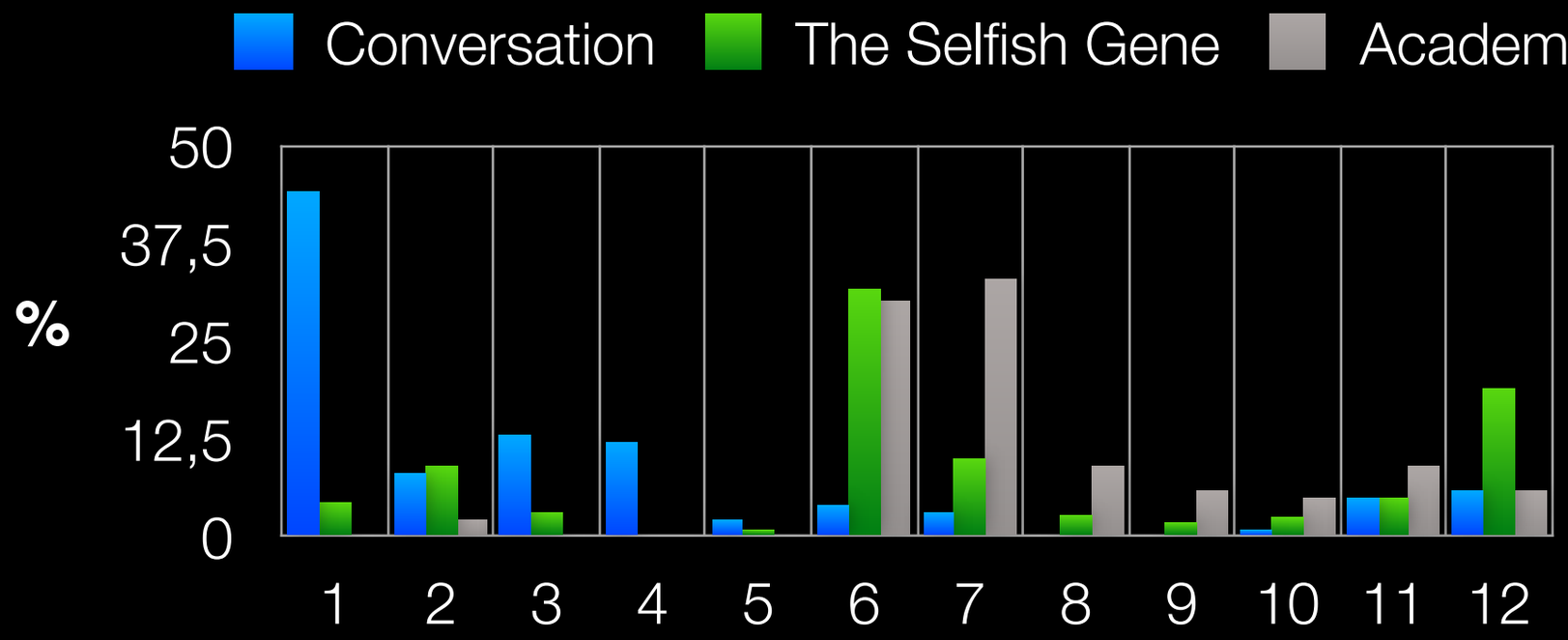
#	N-Gram	n	pmw
1	*IN THE GENE POOL*	46	397
2	*ON THE OTHER HAND*	39	337
3	*THE LONG REACH OF*	34	293
4	*LONG REACH OF THE*	34	293
5	*REACH OF THE GENE*	34	293
6	*IN THE CASE OF*	32	276
7	*NICE GUYS FINISH FIRST*	32	276
8	*BATTLE OF THE SEXES*	27	233
9	*SCRATCH MY BACK I'LL*	24	207
10	*I'LL RIDE ON YOURS*	23	198

4-Grams typical of *academic prose*

4-Grams typical of *conversation*

Example: *The Selfish Gene*

- 4-Gram distribution in *The Selfish Gene*



Example: *The Selfish Gene*

Category	1	2	3	4	5	6	7	8	9	10	11	12
Biber *conversation*	44	8	13	12	4	4	3	0	0	1	5	6
Biber *academic*	0	2	0	0	0	30	33	9	6	5	9	6
The Selfish Gene	18.86	6.7	6.95	2.48	0.5	25.56	8,68	3,47	0	0,99	6,56	19,35

- For now: Distances between is estimated by average deviation from reference values
- Search for most appropriate distance measure from the many measures available (cf. Cha 2007)

Example: *The Selfish Gene*

- Deviation from *conversation*: **15 %**
- Deviation from *academic prose*: **9,24 %**

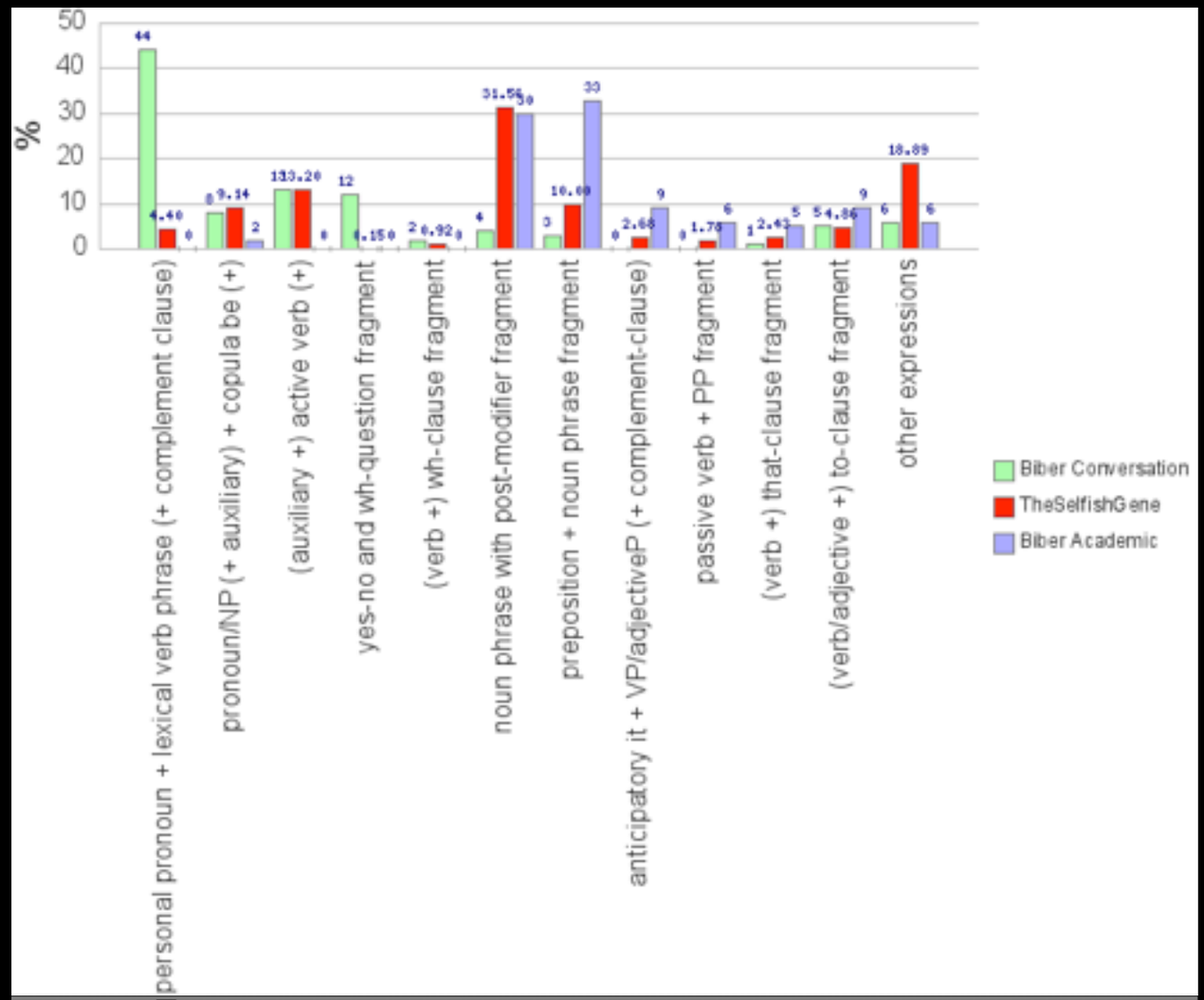


Figure 6: Distribution of 4-grams in *The Selfish Gene*

Example: *The Selfish Gene*

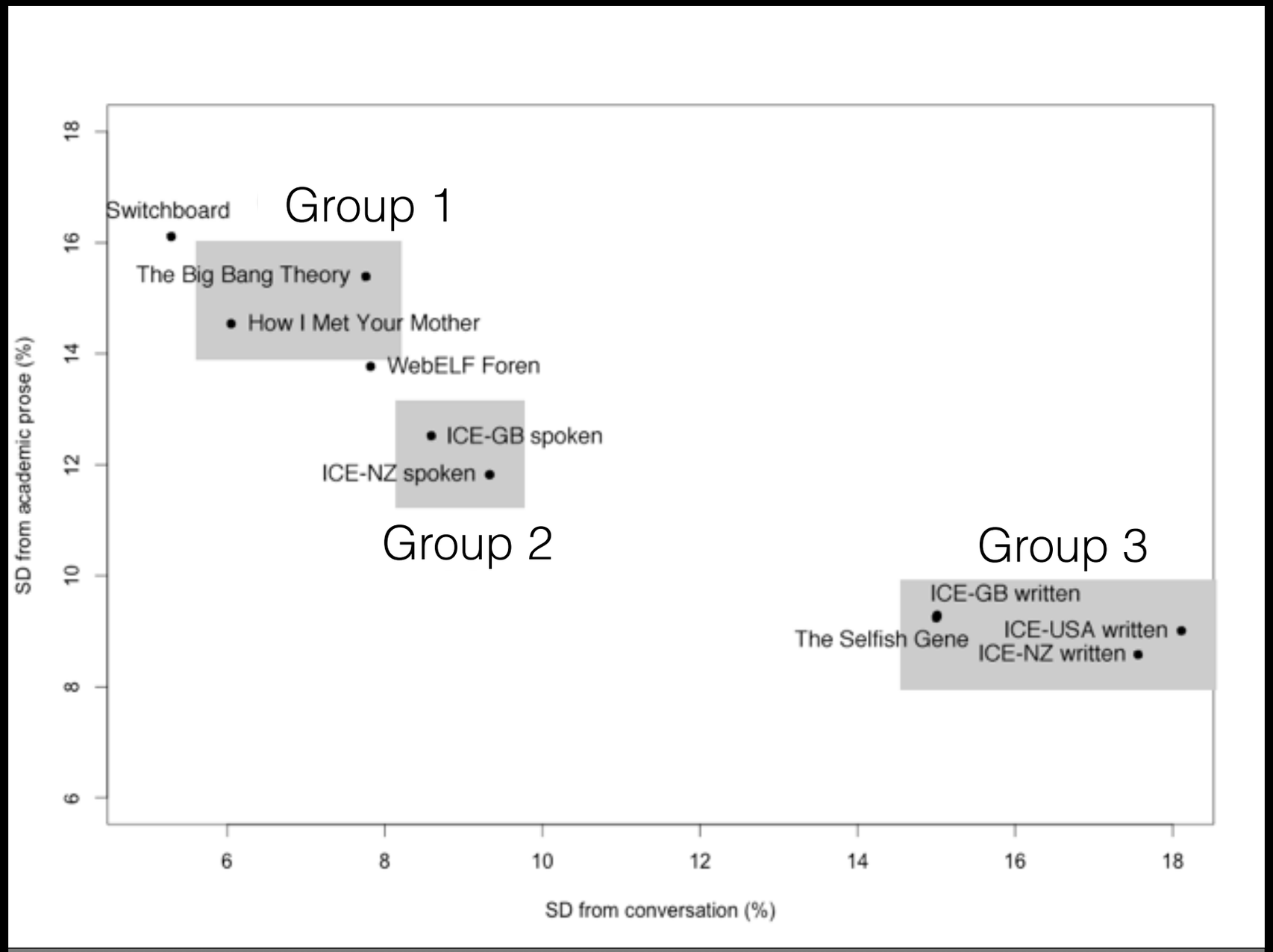


Figure 7: Deviation from the registers *conversation* and *academic prose*

References

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. and R. Quirk (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Cha, S. (2007): "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions". In: *International Journal of Mathematical Models and Methods in Applied Sciences*. Issue 4, Vol. 1. 300–307.

Garside, R., and N. Smith (1997). "A hybrid grammatical tagger: CLAWS4". In: Garside, R., Leech, G., and A. McEnery (eds.). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman. 102–121.

Greenbaum, S. & G. Nelson (2009). *An Introduction to English Grammar*. Third Edition. Harlow: Pearson.

Thank you :-)

Contact:
Benedikt.Heller@kuleuven.be

QLVL:
http://wwwling.arts.kuleuven.be/qlvl