Authorship Attribution based on Stylistic Analysis of Particle Distribution

Andreas Linz, Ute Tischer, Stylianos Chronopoulos, Ying-Chi Lin, Huan Meng

March 2015

1 Introduction

Since the motivation of this project is already presented in the presentation, we do not wish to be repetitious here. In this report, we mainly focus on details of our solution process and presenting the analysis results.

In the analyses, the first two numbers of each file name used in sblgnt data source are taken to represent each letter. Pauline letters are grouped into three categories:
Letters confirmed to be written by Paul: 66, 67, 68, 69, 71, 73, 78
Letters confirmed to be NOT written by Paul: 75, 76, 77, 79
Letters not certain if is written by Paul: 70, 72, 74

Petrine letters are noted as 81 and 82.

2 Data Extraction - sblgntparser

To make it easy to do queries on our data set, the SBL Greek New Testament, we needed to parse it and enrich the outcoming data structure with more information, like the position of a word in a sentence.
Luckily, the SBLGNT format stores a lot of information for every word in each of the books. Each book of the Greek New Testament is saved as one or more text files, where each file consists of a number of rows and each row describes a word in the book. An example row looks like this:

010101 N- —-NSF- ?????? ?????? ?????? ??????

The first six digits are three pairs of two numbers each, they store the book, chapter and verse where the word in this row occurs in the book. The next two characters encode the part of speech of the word, in this example the word is a noun. Thereafter are the codes that describe certain features of a word, like person, tense, mood, degree and so on. We had to reverse engineer the meaning of the codes, with the help of this tool, because there was no description available or we havent found it. The last four words in the row are different forms of the word, the text, word, normalized and lemma form. We only used the lemma form for our queries because this is the one that is the same for different shapes of the same word.

While parsing the data set, I enriched each word in the data model with their position in the sentence (remember there is only the verse encoded in the source), the position in the whole text, the distance to the next in-sentence punctuation (g.e. commata) and each word was linked to its neighbours in the sentence. This gave us the possibility to reason about words based on the type and form of their surrounding words.

The data model, in which the parsed text is transformed into, consists of mainly three parts, these are the text, sentence and word object. Each object has different methods that you can use to ask questions on the text, g.e. loop over all words in the text and return only the conjunctions. Our goal was to do an authorship attribution based on particle distribution, so we needed to make queries on the particles in the text. Because the particle type was not encoded directly in the SBLGNT we had to use list of all ancient greek particles to find them in the text.

3 Data Analysis - absolute position analysis

With the help of sblgntparser, the following variables of data are extracted from the sblgnt text for later analyses:

1. sample_no: one number for each data row
2. letter: first two digits of the source file name, for identification
3. author: name of (likely) author

4. authenticity: categories in wikipedia, coding is here.
5. particle
6. part of speech (POS) of the particle
7. position from the start of the sentence
8. position from the end of the sentence
9. left neighbor of the particle
10. POS of left neighbor
11. right neighbor of the particle
12. POS of right neighbor
13. previous in-sentence (PIP) punctuation and the word
14. PIP position (from sentence start)
15. following in-sentence punctuation (FIP) and the word
16. FIP position (from sentence start)

For the parsing, sentences are defined as text (including in-sentence punctuations) separated by period (.), comma (,) and greek question mark (;). In-sentence punctuations are semicolon ( ) and comma (,).

All the particle occurrences in every corpus and their relevant fields mentioned above are extracted using Python. The results are available at: Pauls lettersPeters letters.

3.1 Absolute position analysis

Generally speaking, when a particle is rarely used and has very strict syntax rule so that position variation in a sentence is hardly possible, it is less suitable for stylometric analysis. In the search of suitable particles in the corpus, in addition to extract the overall frequency of each particle in all sample letters, visualisation of the absolute positions of each particle in each letter is also carried out.

3.2 Visualisation of absolute position analysis

Absolute position of a particle is defined as the position of the particle from the nearest previous punctuation which includes sentence separating punctuation and in-sentence punctuation (PIP). For each particle, its frequency (*number of occurrence in the letter*) in each absolute position is extracted using Python. Furthermore, its position distribution is calculated as (*number of occurrence in each position*)/(*total occurrence of the particle in the letter*). The frequency-position plots for each particle in each letter are drawn using R. To distinguish between types of absolute positions, yellow color is used to depict absolute positions from the start of a sentence and green is for absolute position after a PIP. The quellcodes for frequency extraction are at this page.

In file run.sh contains the workflow of the frequency extraction and plotting, with some batch files in between to connect the subprocesses. For easier comparison, plots are grouped by particles and output as PDF files.

3.3 Selection of particles

Particles with a minimum occurrence in the corpus and with suitable variations in terms of their absolute positions are those we aim for. The following describes the selection by these two criteria:

by frequency For each author, the frequencies of every particle in each letter and their overall occurrence in all letters are shown in Table 1. Particles occur most frequently in Pauls letters correspondent to those in Peters letters and the rankings of occurrence are also identical between both authors. Based on the this frequency results, particles ???, ??, ??? are selected for absolute position analysis.

Particle	Frequency	
	Paul	Peter
???	1406	108
??	696	49
???	546	25
????	327	22
???	123	7
???	71	4
??	44	
???	29	
??	22	
?????????	2	
??????	2	
???	1	
??????	1	
??????	1	

by absolute position variation ??? (engl. and) is the most frequent particle of all. Its absolute position in each letter can be seen in Figure **??**. In most letters of Paul, ??? occurs most frequently at position one, with frequency distribution ranges from 15.6% to 42.7%. Three exceptions with most dominant position at position five are letters 75, 77 and 78. Letters 77 and 78 have the least ??? occurrence (34 and 13 counts, respectively) and letter 78 has moderate frequent of ??? (83 counts). The two letters from Peter show also the same patterns, in letter 79, 39% of the occurrence are at position one and with a small total count of 13, ??? occurs most frequency in position two in letter 78. Since letter 78 is confirmed to be written by Paul and letters 75 and 77 are categorized as uncertain by Paul and additionally, both patterns are observed by two Peters letters, leads to the conclusion that the observed patterns of ??? is not distinct for identification of authorship. ??, the second most frequent particles of all, always dominant at position 2 in all letters. Same pattern is observed with ??? with mostly at least 80% of the occurrence at position two. Little variations on dominant position of ???? is also seen. Therefore, these three particles do not indicate obvious difference between two authors or between different categories of Pauls letters.

The analysis results show that little absolute position variation is observed in the most of the most frequent particles, ??, ??? and ????, in the corpus. Though ??? occurs in more positions in the sentences, its variation in position distribution can be due to small sample size (shorter text??) than authorship difference. Therefore, in order to get more accurate results, longer text and larger corpus is recommended to be used. Further, additional parameters such as xxx

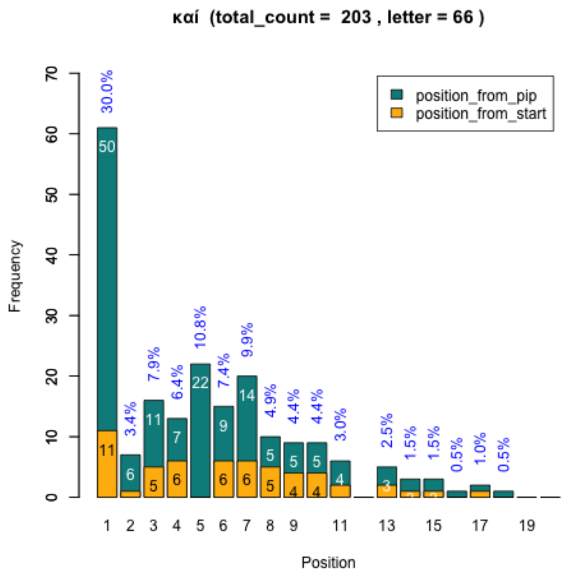
4 References

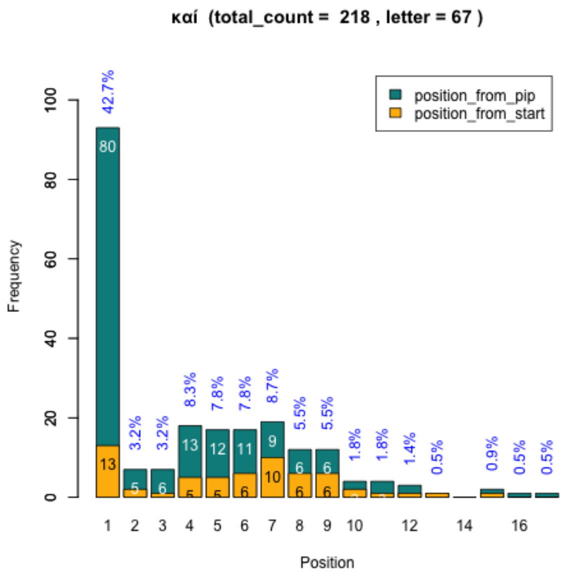
Black, Stephanie (2002), Sentence Conjunctions in the Gospel of Matthew: ???, ??, ????, ???, ???? and Asyndeton in Narrative Discourse. London

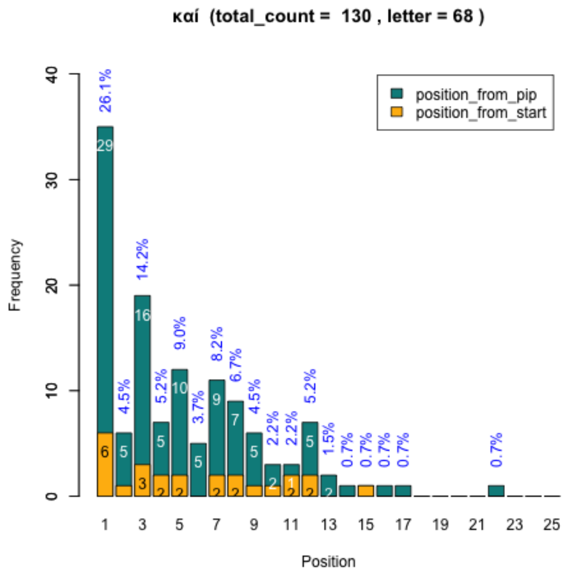
Denniston, J. D. (1954), The Greek Particles. Oxford.

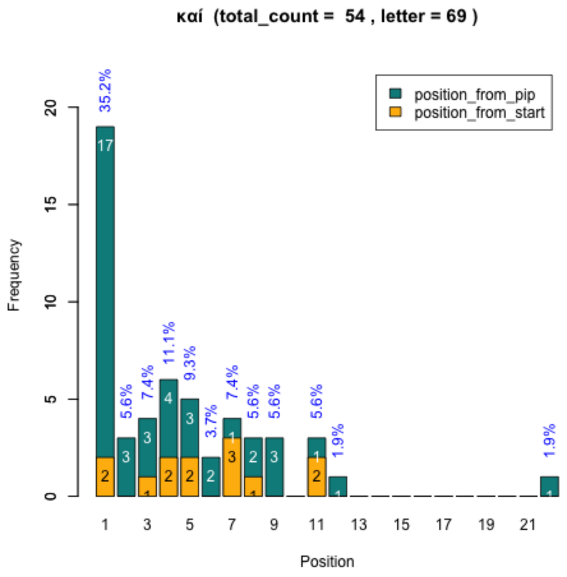
Frischer, Bernard (1991): Shifting paradigms: new approaches to Horaces Ars poetica. American classical studies. Scholars Press, 1991. isbn: 9781555406196. url: http: //books.google.de/books?id=T1F0AAAAIAAJ.

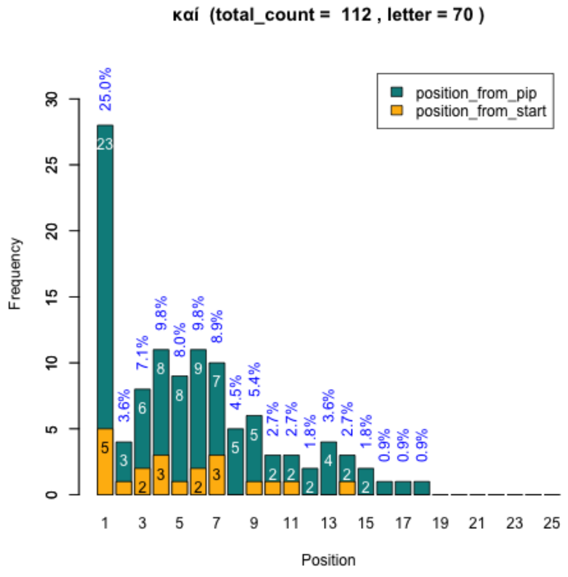
Thrall, Margaret Eleanor (1962) Greek Particles in the New Testament (Linguistic and Exegetical Studies 3). Leiden.

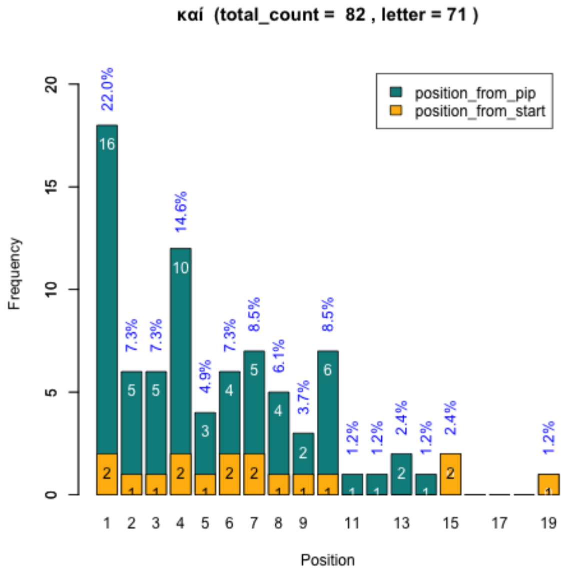

καί (total_count = 203 , letter = 66)
Frequency
Position
position_from_pip
position_from_start
30.0%
50
11
3.4%
6
7.9%
11
5
6.4%
7
6
10.8%
22
7.4%
9
6
9.9%
14
6
4.9%
5
5
4.4%
5
4
4.4%
5
4
3.0%
4
2.5%
3
1.5%
1.5%
0.5%
1.0%
0.5%


καί (total_count = 218 , letter = 67)
Frequency
Position
position_from_pip
position_from_start
42.7%
80
13
3.2%
5
3.2%
6
8.3%
13
5
7.8%
12
5
7.8%
11
6
8.7%
9
10
5.5%
6
6
5.5%
6
6
1.8%
1.8%
1.4%
0.5%
0.9%
0.5%
0.5%

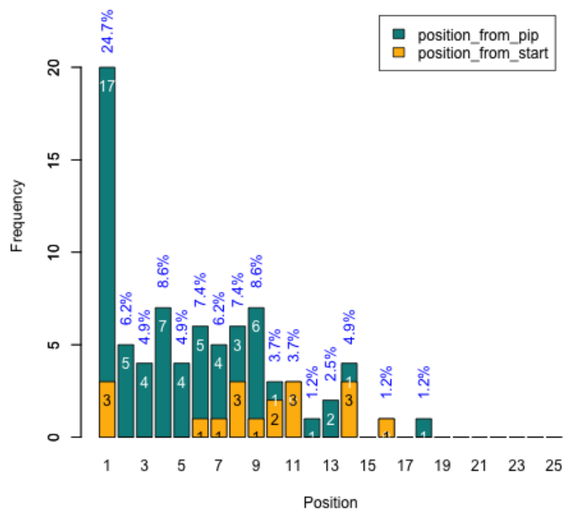

καί (total_count = 130 , letter = 68)
Frequency
Position
position_from_pip
position_from_start
26.1%
29
6
4.5%
5
14.2%
16
3
5.2%
5
2
9.0%
10
2
3.7%
5
8.2%
9
2
6.7%
7
2
4.5%
5
2.2%
2
2.2%
1
2
5.2%
5
2
1.5%
0.7%
0.7%
0.7%
0.7%
0.7%


καί (total_count = 54 , letter = 69)
Frequency
Position
position_from_pip
position_from_start
35.2%
17
2
5.6%
3
7.4%
3
11.1%
4
2
9.3%
3
2
3.7%
2
7.4%
1
3
5.6%
2
5.6%
3
5.6%
1
2
1.9%
1
1.9%
1

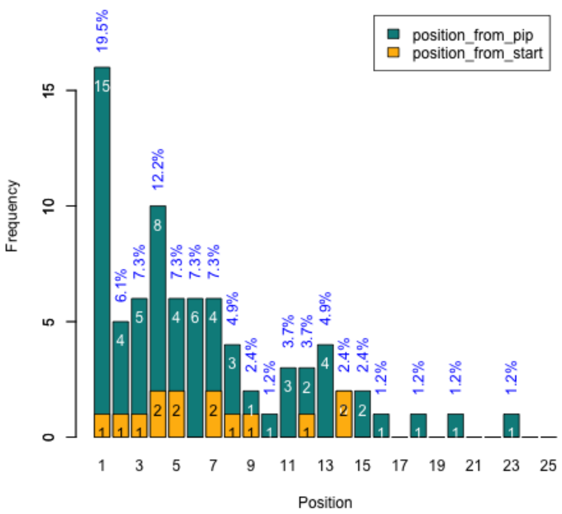

καί (total_count = 112 , letter = 70)
Frequency
Position
position_from_pip
position_from_start
25.0%
23
5
3.6%
3
7.1%
6
2
9.8%
8
3
8.0%
8
9.8%
9
2
8.9%
7
3
4.5%
5
5.4%
5
2.7%
2
2.7%
2
1.8%
2
3.6%
4
2.7%
2
1.8%
2
0.9%
0.9%
0.9%


καί (total_count = 82 , letter = 71)
Frequency
Position
position_from_pip
position_from_start
22.0%
16
2
7.3%
5
1
7.3%
5
1
14.6%
10
2
4.9%
3
1
7.3%
4
2
8.5%
5
2
6.1%
4
1
3.7%
2
1
8.5%
6
1
1.2%
1
1.2%
1
2.4%
2
1.2%
1
2.4%
2
1.2%
1

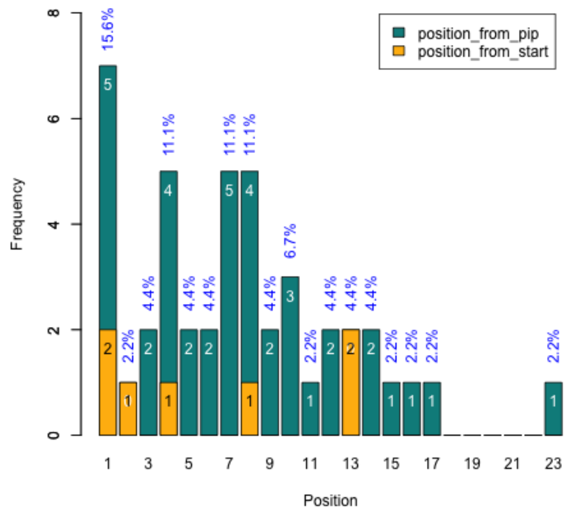
καί (total_count = 79 , letter = 72)



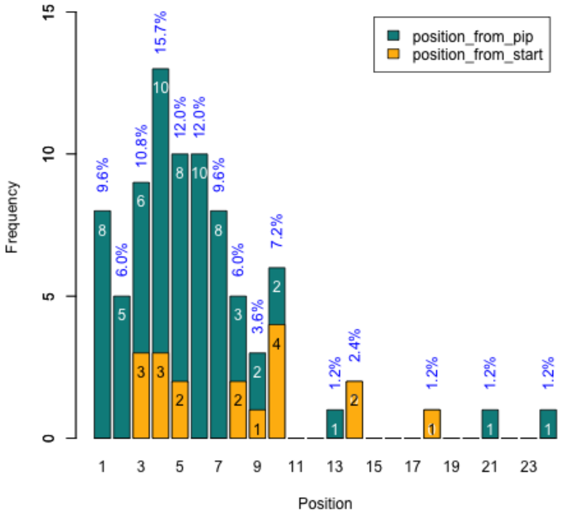
καί (total_count = 80 , letter = 73)



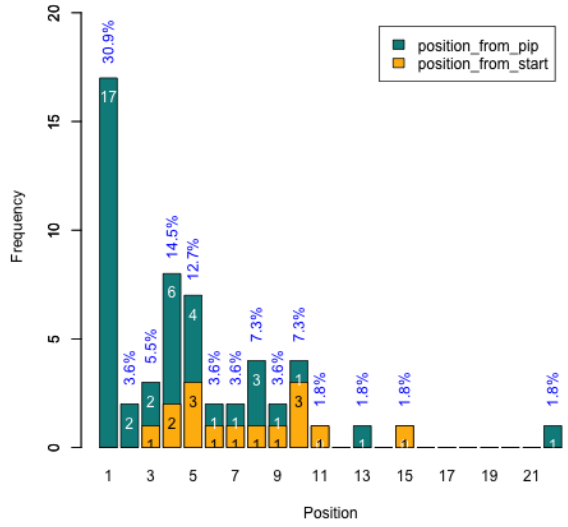
καί (total_count = 45 , letter = 74)



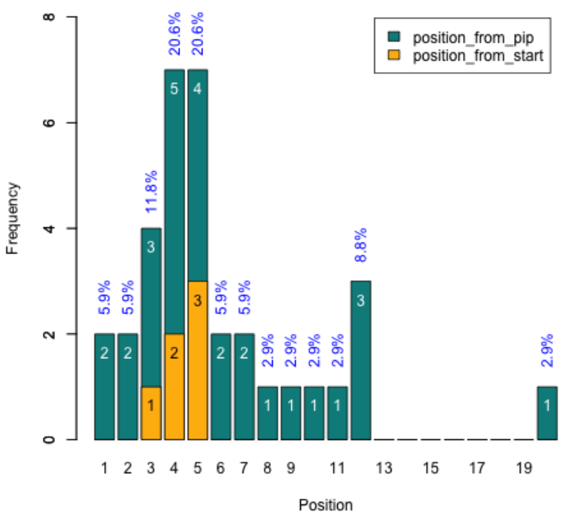
καί (total_count = 83 , letter = 75)

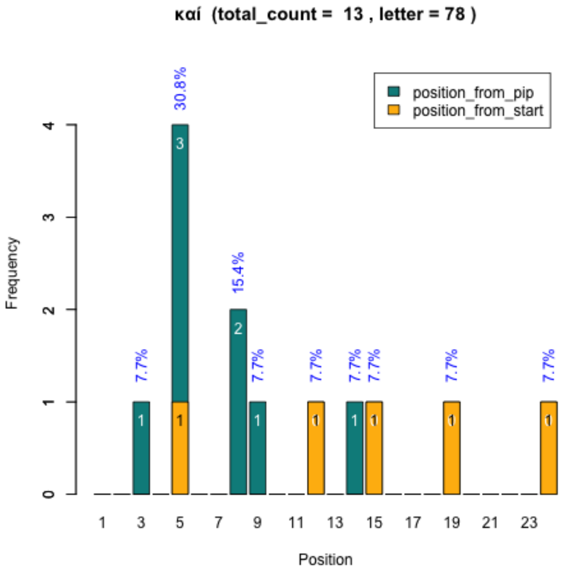


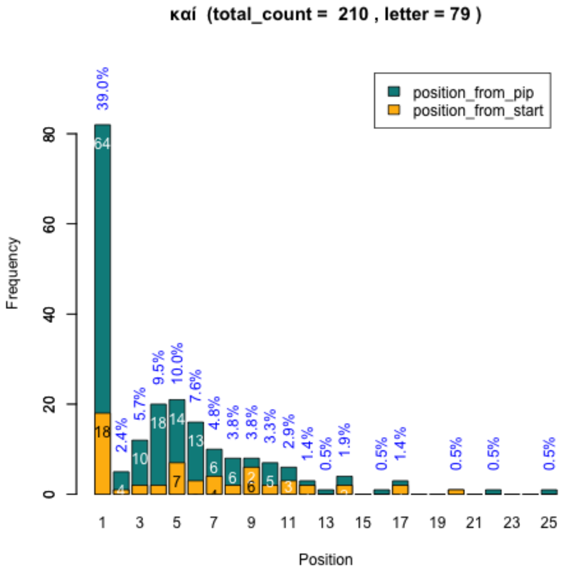
καί (total_count = 55 , letter = 76)



καί (total_count = 34 , letter = 77)




καί (total_count = 13 , letter = 78)
position_from_pip
position_from_start
Frequency
0
1
2
3
4
Position
1
3
5
7
9
11
13
15
17
19
21
23
7.7%
30.8%
15.4%
7.7%
7.7%
7.7%
7.7%
7.7%
7.7%
1
3
1
2
1
1
1
1
1
1


καί (total_count = 210 , letter = 79)
position_from_pip
position_from_start
Frequency
0
20
40
60
80
Position
1
3
5
7
9
11
13
15
17
19
21
23
25
39.0%
2.4%
5.7%
9.5%
10.0%
7.6%
4.8%
3.8%
3.8%
3.3%
2.9%
1.4%
0.5%
1.9%
0.5%
1.4%
0.5%
0.5%
0.5%
64
18
4
10
18
14
7
13
6
6
2
6
5
3