Zipf's Law in Aphasia Across Languages: A Comparison of English, Hungarian and Greek

Article *in* Journal of Quantitative Linguistics · January 2017

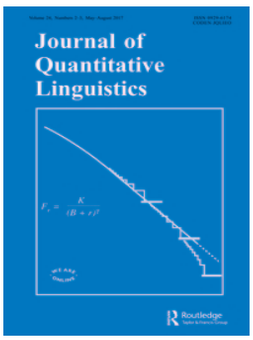
DOI: 10.1080/09296174.2016.1263786

3 authors:

Kyriaki Neophytou
Johns Hopkins University
9 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Marjolein van Egmond
Amsterdam University Medical Center
126 PUBLICATIONS **6,630** CITATIONS
SEE PROFILE

S. Avrutin
Utrecht University
127 PUBLICATIONS **1,403** CITATIONS
SEE PROFILE



Routledge
Taylor & Francis Group

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: http://www.tandfonline.com/loi/njql20

Zipf's Law in Aphasia Across Languages: A Comparison of English, Hungarian and Greek

Kyriaki Neophytou, Marjolein van Egmond & Sergey Avrutin

To cite this article: Kyriaki Neophytou, Marjolein van Egmond & Sergey Avrutin (2017) Zipf's Law in Aphasia Across Languages: A Comparison of English, Hungarian and Greek, Journal of Quantitative Linguistics, 24:2-3, 178-196, DOI: 10.1080/09296174.2016.1263786

To link to this article: https://doi.org/10.1080/09296174.2016.1263786

JOURNAL OF QUANTITATIVE LINGUISTICS, 2017
VOL. 24, NOS. 2–3, 178–196
http://dx.doi.org/10.1080/09296174.2016.1263786

Routledge
Taylor & Francis Group

Zipf's Law in Aphasia Across Languages: A Comparison of English, Hungarian and Greek

Kyriaki Neophytou[a†], Marjolein van Egmond[b] and Sergey Avrutin[b]

[a]Neuroscience of Language Lab, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates; [b]Department of Language, Literature and Communication, Utrecht Institute of Linguistics, Utrecht University, Utrecht, The Netherlands

ABSTRACT

We investigated Zipf's law in fluent and non-fluent aphasics' spontaneous speech in English, Hungarian, and Greek. A previous study showed that the word frequency distribution in Dutch non-fluent aphasic speech conforms to Zipf's law, although with a different slope. In this project we investigated to what extent these results can be generalized to other languages and to fluent aphasic speech. The results suggest that both the fluent and the non-fluent aphasic speech of English, Hungarian and Greek conform to Zipf's law, and that differences in slope can be related to a language's morphological properties and a group's particular language impairments.

1. Introduction

Zipf's law is a power law that characterizes the word frequency distribution in a corpus of natural language (Zipf, 1949). More specifically, it describes the rank-frequency distribution of words in a text. In the formula used to express this distribution (Formula 1), f_r is the frequency of the rth word, given that the words of the text are ordered by decreasing frequency. Therefore, in Zipf's law, the frequency of any word in a text is inversely proportional to its rank. C is a constant, dependent on the text size (i.e. number of tokens). For large corpora, it is usually found that $\alpha \approx 1$.

Formula 1: Zipf's law

$$f_r \approx \frac{C}{r^{\alpha}}$$

Word frequency distributions systematically deviate from Zipf's law for the first few ranks, as Zipf's law overestimates their frequency. Mandelbrot (1953)

CONTACT Kyriaki Neophytou Kyriaki.neophytou@nyu.edu
†This research was conducted at the Utrecht Institute of Linguistics.

therefore added an extra parameter to Zipf's law to accommodate for this discrepancy, namely β (Formula 2, henceforth ZML). By adding this parameter, Mandelbrot achieved a better fit of real linguistic data compared to the traditional Zipf's law. This extra parameter has a heavy influence on the value of f_r for low rank numbers, those with the highest frequency (which were overestimated by the traditional Zipf's law), but only a small influence on the higher ranks (for discussion see Mandelbrot, 1953; Vogt, 2004). Usually, it is found that the highest frequency words are function words, as they are, among other things, responsible for the formation of grammatical structure (e.g. Caplan, 1987). Since β helps to calculate a more accurate α by accounting better for the high frequency of the function words, it is plausible that the β-values are closer linked to grammatical structure rather than the structure of the lexicon per se.

Formula 2: Zipf–Mandelbrot law (ZML)

$$f_r \approx \frac{C}{(r+\beta)^{\alpha}}$$

ZML has been shown to apply to many different texts from various languages, including Greek (Hatzigeorgiu, Mikros, & Carayannis, 2001), Hungarian, German (Németh & Zainkó, 2002), Spanish, Irish and Latin (Ha, Stewart, Hanna, & Smith, 2006). It is almost exclusively studied in written texts, and much less frequently in spontaneous, spoken language (some exceptions being Ridley, 1982; Ridley & Gonzales, 1994). Zipf's law has been minimally investigated in impaired speech as well. Piotrovskii, Pashkovskii, and Piotrovskii (1994) and Piotrowski and Spivak (2007) studied word frequency distributions in Russian, in speech from people with schizophrenia, Down syndrome children and women 1–2 weeks before and 3–4 days after giving birth, a period during which the authors consider women to be under 'birth stress'. However, the sample sizes per text were highly diverse in these studies, rendering it hard to compare the observed values. Howes and Geschwind were the first who studied Zipf's law in aphasic speech in English (Howes, 1964; Howes & Geschwind, 1964), followed by van Egmond, van Ewijk, and Avrutin (2015)[1] in Dutch. In this paper, we focus on spontaneous speech from people with aphasia, a language disorder most commonly acquired after a stroke or a trauma. Aphasic speech sounds distinctively different compared to healthy speech, but still little is known about the underlying properties of it.

Aphasia can be divided into several sub-types, depending on the exact impairments. The broadest distinction made is the one between *fluent* and *non-fluent* aphasia. Non-fluent aphasic speech, like the speech from Broca's aphasic people, is usually described as effortful and telegraphic, with multiple omissions and/or substitutions, mostly of functional categories (Goodglass, Fodor, & Schulhoff, 1967). As a result, non-fluent aphasic speech sounds markedly different from healthy speech, but it is still meaningful. On the contrary, fluent aphasic speech, like the speech from Wernicke's aphasic people, usually

sounds effortless and continuous. Nevertheless, it is hard to follow: paraphasias and neologisms are included while content words are omitted or substituted (Bastiaanse, 2011; Wernicke, 1874), thus creating incomprehensible phrases and sentences.

The difficulties of non-fluent aphasic people in speech production are thought to be due to the reduction of processing resources, not to the loss of syntactic abilities (Avrutin, 2006). Previous studies show that the syntactic processes are intact in people with non-fluent aphasia (e.g. Zurif, Swinney, Prather, Solomon, & Bushell, 1993). The problem is that they are slower. The brain injuries in these patients affect the amount of available processing resources. In order to compensate for the reduction of those resources, more processing time is required, thus causing the various distortions in speech.

The language problems of fluent aphasics are also thought to relate to the processing system and not to the lexicon itself, but they surface very differently. The content words that fluent aphasics have problems with are mostly finite verbs (Bastiaanse, 2011). Specifically, the variability in finite verbs is reduced compared to the variability in infinitive verbs. This shows that there is no lexical retrieval problem in non-fluent aphasics, but either a grammatical complexity problem or an integration problem. Since, similarly to non-fluent aphasia, the problem is not in the lexicon itself, the processing system is likely to be responsible for the linguistic difficulties of fluent aphasic people as well.

The only study on Zipf's law (using the formula that more accurately accounts for the high frequency words) in aphasic speech so far is by van Egmond et al. (2015) on a group of Dutch non-fluent aphasic patients. The results show that the speech from non-fluent aphasic people, although sounding markedly different from healthy speech, still conforms to Zipf's law but with a different slope. Given equal size samples, this difference in slope reflects the fact that speech from aphasic speakers includes less different word types (i.e. less variation of linguistic items) and that these types are used more frequently compared to normal speech. van Egmond et al. (2015) argue that these findings provide an insight into the origins of Zipf's law: they suggest that it is the organization of the mental lexicon, which is intact in this group of people, that renders language to conform to Zipf's law.

The study by van Egmond et al. (2015) provided evidence that Zipf's law applies to the speech of non-fluent aphasic patients in Dutch. Therefore, the aim of the current study is twofold. First, we aim to see to what extent the non-fluent aphasic speech findings can be generalized to other languages. Second, this study seeks to explore Zipf's law in fluent aphasic speech for the first time. We investigate whether fluent and non-fluent aphasic speech share the same properties or not. To this end, we examine both the fit and the slope of the rank-frequency distribution.

We first ran a within-language analysis to test whether the word frequency distribution in the speech of aphasic patients and healthy people has significant

differences in a language other than Dutch; namely, English. Following the findings by van Egmond et al. (2015), it was expected that the fit to ZML should not differ among the different groups. The slopes of the distribution, however, were expected to vary. Given equal sample sizes, non-fluent aphasic patients were expected to produce fewer different word types than healthy controls, but to use the produced items more frequently. Therefore, we predicted that non-fluent aphasic speech would have a steeper slope compared to healthy speech. Fluent aphasic patients, on the other hand, were expected to produce more different word types than healthy controls, due to the fluent nature of their speech and the production of neologisms and paraphrases. Therefore, we predicted that fluent aphasic speech would have a less steep slope compared to healthy speech.

Secondly, we investigated the properties of ZML in aphasic speech, both fluent and non-fluent, across different languages; more specifically, Greek, English and Hungarian. In this between-languages analysis, similarly to the within-languages analysis, we did not expect to find any differences in the fit to ZML, but we did expect to find differences in slope between the different languages. Languages differ on various levels, ranging from alphabet to phonology and morphology. Even in the same language, an older and a newer version of it may differ to a great extent. A previous study by Bentz, Kiela, Hill, and Buttery (2014) showed a difference in the parameters of ZML between Old English and Modern English, a difference that lies in the morphological variability of these two historical versions of English. Specifically, a language with poor morphology, like Modern English, displays a steeper slope than a language with rich morphology, like Old English. In the current study we investigated whether fluent and non-fluent aphasic speech also reflects this difference in morphological variability among the different languages we tested. Given that ZML is intact in aphasic speech, we expect to find that speech from the morphologically simple English has a steeper slope compared to speech from the morphologically complex Hungarian and Greek.

2. Methods

2.1. Data

All spontaneous speech transcripts were taken from AphasiaBank (MacWhinney, Fromm, Forbes, & Holland, 2011), an online database with speech from aphasic patients in different languages. We selected all texts that fitted two criteria: the speech had to be spontaneous and non-directed (i.e. an unstructured interview) and the patient's aphasia type should be clearly categorized into fluent or non-fluent. This resulted in English transcripts from fluent aphasic, non-fluent aphasic and healthy speakers; and Greek and Hungarian transcripts from fluent and non-fluent aphasic speakers (English: Bates, Friederici, Wulfeck, & Juarez, 1988; Hungarian: MacWhinney & Osmán-Sági, 1991; Greek: Goutsos, Potagas, Kasselimis, Varkanitsa, & Evdokimidis, 2011). No healthy speech samples were

included in the corpus for Greek, whereas the healthy Hungarian speech samples that were included in the corpus were collected through directed speech tasks, and therefore did not meet our data inclusion criteria. Detailed information for the participants can be found in Appendix 1.

All conversations were in CHAT-format (MacWhinney, 2000). In the analysis for each conversation, the speech from the investigator was ignored. Thus, similarly to the van Egmond et al. (2015) study, the analysed speech was as it naturally occurred but it does not form continuous streams of speech.

2.2. Analysis

As the parameters of ZML are highly dependent on text size, we selected the same number of tokens from each participant. In order to have equal size samples, the first 200 words were analysed. This number was the balance between both including as many speech samples as possible and including as many words as possible, because many participants' recordings were relatively short.

For the within-language analysis, the comparisons were made between fluent, non-fluent and healthy speech for English only. For the between-languages analysis, the comparisons were made between fluent aphasic speech and non-fluent aphasic speech for all three languages.

First, rank and frequency were logarithmically transformed. Then, the R^2, the alpha- and the beta-values were calculated through maximum likelihood estimation. The R^2-values were used to define whether ZML applies. Although it is still a matter of debate of when it can be claimed that ZML does not apply, the closer the R^2values are to 1, the better the fit to ZML is. These values were also statistically compared to see if any of the groups displayed a significantly lower fit than the others. The alpha values were statistically compared between groups. The beta-values, which, as discussed, are mainly based on the first few ranks, were calculated but not further analysed. In these samples, the number of ranks is so small that statistical analyses are unreliable.

For the within-language analysis we conducted a simple ANOVA analysis, since there was only a single independent variable: group. For the between-languages analysis we conducted a factorial ANOVA because there were two independent variables: group and language.

3. Results

Mean outcomes per group and the minimum and maximum value per group are given in Table 1. Detailed results per speaker are provided in Appendix 2.

A box plot depicting the outcomes for α for the three groups in English is given in Figure 1. A box plot depicting the outcomes for α for the two groups in English, Greek and Hungarian is given in Figure 2.

Table 1. Mean values, SD, minimum and maximum values per group, per outcome value.

			R^2				a			
Language	Group	*N*	Mean	SD	Min	Max	Mean	SD	Min	Max
English	Control	26	0.972	0.019	0.897	0.989	0.685	0.976	0.57	0.976
	Non-fluent	26	0.968	0.019	0.927	0.99	0.928	1.993	0.593	1.993
	Fluent	24	0.971	0.016	0.919	0.988	0.801	1.162	0.582	1.162
Greek	Non-fluent	5	0.957	0.019	0.943	0.99	0.713	0.826	0.608	0.826
	Fluent	12	0.96	0.016	0.933	0.978	0.662	0.808	0.563	0.808
Hungarian	Non-fluent	5	0.963	0.01	0.955	0.977	0.689	0.798	0.63	0.798
	Fluent	5	0.964	0.015	0.942	0.978	0.725	0.84	0.563	0.84

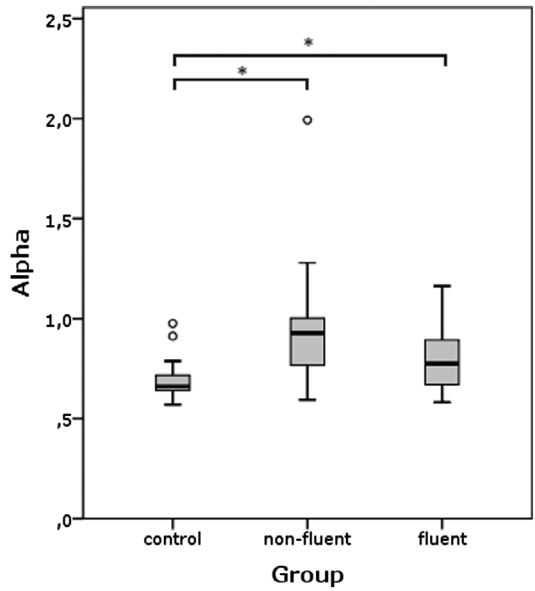


Figure 1. Boxplot: *a* for English control speakers and fluent and non-fluent English aphasic speakers. (Stars indicate the significant differences between groups.)

3.1. Within-language Analysis

The within-language analysis concerned the text samples from English fluent and non-fluent aphasic speakers and control speakers.

R^2

From Levene's test, we were able to conclude that there was homogeneity of variance, $F(2, 73) = 0.44$, $p = 0.43$. Using simple ANOVAs, no difference between groups was found. Thus, R^2 was equally high for all groups. The lowest R^2 value that was found for any of the texts is 0.897, which was for one of the control speakers.

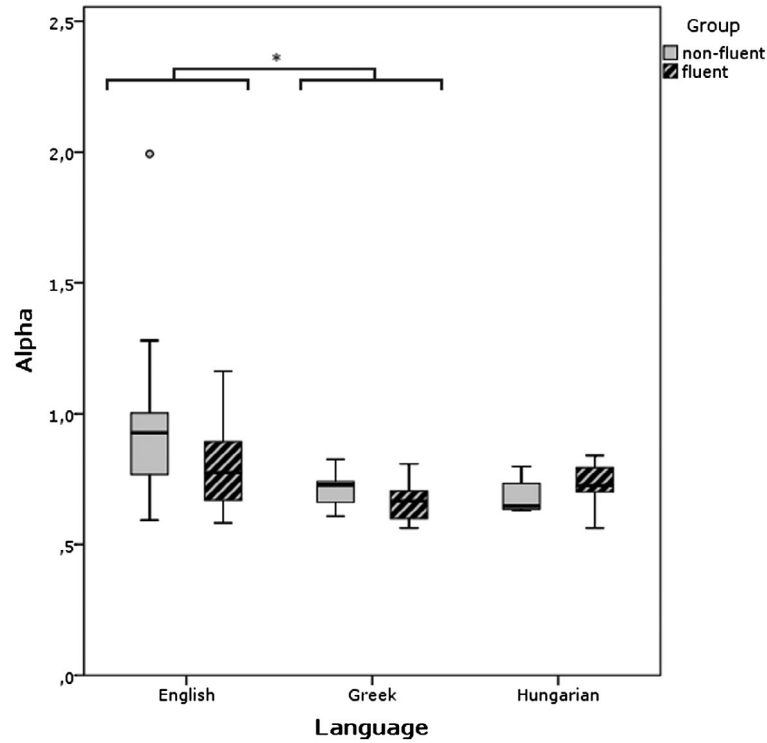


Figure 2. Boxplot: *α* for English, Greek and Hungarian fluent and non-fluent aphasic speakers. (Stars indicate the significant differences between groups.)

Alpha

Levene's test showed that there was no homogeneity of variance, $F(2, 73) = 4.96$, $p = 0.04$. Reciprocal transformation of the data solved this, $F(2, 73) = 2.06$, $p = 0.14$.

Simple ANOVA analysis revealed that there existed a significant difference between groups, $F(2, 73) = 13.06$, $p < 0.001$. *Post hoc* tests using the Bonferroni correction showed that both the non-fluent group ($p < 0.01$) and the fluent group ($p < 0.05$) displayed a significantly higher alpha compared to the control group. The fluent group and non-fluent group did not differ significantly from each other.

3.2. Between-languages Analysis

The between-languages analysis concerned the text samples from English, Hungarian and Greek fluent and non-fluent aphasic speakers. It should be noted though that, as the Greek non-fluent aphasic group and the Hungarian

fluent and non-fluent aphasic groups were rather small, statistical analyses are only a tentative indication of the strength of the effect.

R^2

From Levene's test, we were able to conclude that there was homogeneity of variance, $F(5, 71) = 0.46$, $p = 0.81$. Using factorial ANOVAs, no difference between groups was found. Thus, R^2 was equally high for all groups. The lowest R^2 value that was found for any of the texts was 0.897, which was for one of the English control speakers.

Alpha

Levene's test showed that there was homogeneity of variance, $F(5, 71) = 1.58$, $p = 0.18$.

Factorial ANOVA analysis revealed that there was a significant effect of language, $F(5, 71) = 6.19$, $p = 0.003$. There was no significant main effect of group and no interaction effect between language and group. *Post hoc* tests using the Bonferroni correction showed that Greek and English were significantly different from each other ($p = 0.003$), with English displaying higher alpha values. The difference between English and Hungarian approached significance ($p = 0.06$), with English tending to have higher alpha values. No significant difference was found between Hungarian and Greek.

4. Discussion

4.1. Within-language Analysis

The high R^2 values and the fact that no significant differences between the various groups were found suggest that speech from all groups conforms to ZML. Although the speech from fluent aphasic people and non-fluent aphasic people sounds markedly different from the speech of healthy speakers, this difference is not in any way reflected in the fit of their frequency distributions to ZML. These results confirm the earlier findings by van Egmond et al. (2015) for Dutch non-fluent aphasic speech. More importantly, they provide evidence that this finding applies cross-linguistically and to fluent aphasic speech.

The statistical analysis of the alpha values shows that both fluent and non-fluent aphasic speech display a steeper slope compared to healthy control speech. These results suggest that although the speech from all three groups conforms to ZML, the linguistic impairments of the two aphasic groups still affect the parameters of their word frequency distribution. The steeper slope of the non-fluent aphasic speech is in line with the van Egmond et al. (2015) findings for Dutch. These results are also in accordance with the results by Piotrovskii et al. (1994) and Piotrowski

and Spivak (2007) who studied Zipf's law in other impaired populations (discussed above) and reported steeper slopes than the average natural language texts.

The results for the fluent aphasic speakers were not as expected. The healthy control speech was expected to have a steeper slope compared to the fluent aphasic speech, but the results show a significant effect in the opposite direction. This means that both groups of aphasic speakers perform the same way. As this is the first study investigating ZML in fluent aphasic speech, further research is required to find out why this might be so. ZML only concerns quantitative data. However, a qualitative investigation into the difference between fluent and non-fluent aphasic speech in light of these findings may be needed to explain them. Such line of research should try to identify differences in the nature (i.e. neologisms and paraphasias versus existing words) and the linguistic properties (i.e. the linguistic category) of the items the two aphasic groups are producing.

4.2. Between-languages Analysis

Adding to the within-languages analysis for English, the between-languages analysis revealed that the speech from Greek and Hungarian aphasic speakers also conforms to ZML. All R^2 values were very high, thus providing evidence that the fit to Zipf's law is a cross-linguistic phenomenon, independent of the linguistic impairments manifested in fluent and non-fluent aphasia.

The comparisons of the alpha values revealed a significant effect of language. Specifically, English has a steeper slope than Greek, and a trend for a steeper slope compared to Hungarian, while Greek and Hungarian show no significant difference between them.

Bentz et al. (2014), in a comparison between Old English and Modern English, showed that a language with poor morphology displays a steeper slope (i.e. higher alpha values) than a language with rich morphology. Thus, the morphologically poor Modern English has a steeper slope than the richer Old English. We suggest then that the current results could reflect this difference in the morphology of the three languages we studied. English is a morphologically poor language, especially when compared to Greek (and Hungarian). In Greek, nouns and adjectives are marked for number, gender and case, while verbs are marked for tense, person and number. However, in English, nouns and adjectives are only marked for number, and verbs are (only to some extent) marked for tense and person. This difference in grammatical encoding strategies is mirrored in the length difference of the tails of the frequency distributions, with English having a significantly steeper slope than Greek. We speculate that the reason why we did not find a statistically significant difference between English and Hungarian is because the Hungarian sample was too small (only 10 participants), and not because of morphology. Greek and Hungarian, both morphologically rich languages, show no significant difference in alpha values between them.

Contrary to our predictions, but similar to the findings for the within-language analysis, the comparison of the alpha values between fluent and non-fluent aphasic speech revealed no significant difference between the two. We speculate that the small speech sample sizes (i.e. 200 words) might not include as many paraphasias and neologisms for the fluent aphasic group and as many repetitions for the non-fluent aphasic group as needed to yield the predicted difference. Therefore, as discussed before, further investigation is needed to look into the nature and the linguistic properties of the items the two groups produce to identify the exact differences and similarities between the two.

4.3. Further Implications for Aphasia

Previous studies have shown that natural language texts conform to the specific word-frequency distribution defined by Zipf's law. This law has been observed to exist in many languages with varying morphological properties. Based on these findings, we expect that the text produced by any unimpaired language system will follow a Zipfian word-frequency distribution. If this system is not functioning properly, though, we do not really know what the properties of the output of such system will be. One would expect it to be disorganized, lacking basic structure and disobeying fundamental linguistic rules.

However, it all depends on the nature of the impairment and the aspects of the system that it is touching upon. If the impairment is such that it affects the core aspects of the language system, like the structure of the lexicon and the selection and short-term storage of the lexical items, then we would expect a distorted Zipfian distribution in the output. Nonetheless, the current findings suggest that this is not the case either for fluent or non-fluent aphasic speech. Although the former includes paraphasias and neologisms and the latter is effortful and telegraphic, both of them conform to Zipf's law. These findings suggest that something deep, underlying the observed impairments in aphasia, is still intact. If the whole system was disrupted, the speech of these people would not match the word-frequency distribution of healthy, control speech. The next steps should then be to identify what exactly is affected in aphasia and what language properties are still untouched.

5. Conclusion

In this paper we investigated the frequency distribution properties of aphasic speech, both fluent and non-fluent, in English, Greek and Hungarian. Van Egmond et al. (2015) reported that Dutch non-fluent aphasic speech conforms to Zipf's law. That was the first study to provide this kind of evidence for any type of aphasic speech. The current study shows that it is not only non-fluent but also fluent aphasic speech that conforms to ZML, and that this applies across languages with varying morphological complexities.

The variations in slope that we report in this study between the non-fluent aphasic and healthy speech, with non-fluent aphasic speech having a steeper slope than the healthy speech, are in line with the particular language deficits in this group. However, the finding that the fluent aphasic speech also had a steeper slope than the healthy speech was unexpected. A qualitative investigation into the difference between fluent and non-fluent aphasic speech in light of these findings is needed to explain them.

The variations in slope between English and Greek appear to reflect the language-specific morphological properties of the two languages. This is supported by the trending difference between English and Hungarian, which is expected to reach statistical significance with bigger sample sizes. At the same time, we would also like to see whether the frequency distributions of languages of increasing morphological complexity show scalar differences in the parameters of ZML, an interesting topic for further research.

To conclude, this is the first study to provide cross-linguistic evidence that aphasic speech, both fluent and non-fluent, although sounding markedly different from healthy speech, conforms to Zipf-Mandelbrot's law. We take these findings as an extra step in revealing what exactly is impaired in aphasia, independently of its type. If the language system of an aphasic person is completely disrupted it would not adhere to fundamental aspects of natural speech, such as the Zipfian distribution of word-frequency. As we have shown in the present study, this is not the case.

Based on these findings, we can follow two different lines of research, both equally important. On the one hand, we need to work further into understanding what makes the speech of the aphasic population sound so distinctively different from healthy speech. Since their speech conforms to Zipf's law, we expect that some other language properties are responsible for the observed output. On the other hand, these findings raise more questions regarding the lexicon and its properties on a more general, cognitive level. The lexical system cannot be solely defined by its frequency distribution properties. There are other aspects of this system that are yet to be defined and connected to each other for a complete picture of lexical access and language production. These two lines of research complement each other, as each step forward in one will help to better understand the other.

Note

1. The frequency distribution of aphasic speech was first investigated by Howes and Geschwind (Howes, 1964; Howes & Geschwind, 1964). Unfortunately, they did not use Zipf's law, but a cumulative version concerning the percentage of words that occur with frequencies up to and including each frequency value. This formulation is not very sensitive: disruptions in the higher frequency classes are easily concealed if the lower frequency classes do follow a Zipfian distribution. van Egmond et al. (2015) were the first to study Zipf's law in aphasic speech as such.

Acknowledgements

This research was conducted at the Utrecht Institute of Linguistics, as part of the dissertation of Marjolein van Egmond.

Disclosure Statement

The authors declare that there are no conflicts of interest.

References

Avrutin, S. (2006). Weak syntax. In Y.Grodzinsky & K.Amunds (Eds.), *Broca's region* (pp. 49–62). Oxford: Oxford University Press.

Bastiaanse, R. (2011). The retrieval and inflection of verbs in the spontaneous speech of fluent aphasic speakers. *Journal of Neurolinguistics, 24*, 163–172.

Bates, E. A., Friederici, A. D., Wulfeck, B. B., & Juarez, L. A. (1988). On the preservation of word order in aphasia: Cross-linguistic evidence. *Brain and Language, 33*, 323–364.

Bentz, C., Kiela, D., Hill, F., & Buttery, P. (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory, 10*, 175–211.

Caplan, D. (1987). *Neurolinguistics and linguistic aphasiology: An introduction*. Cambridge: Cambridge University Press.

Goodglass, H., Fodor, I. G., & Schulhoff, C. (1967). Prosodic factors in grammar-evidence from aphasia. *Journal of Speech, Language, and Hearing Research, 10*, 5–20.

Goutsos, D., Potagas, C., Kasselimis, D., Varkanitsa, M., & Evdokimidis, I. (Eds.). (2011). *Studying paraphasias in the Corpus of Greek Aphasic Speech*. Athens: Synapses.

Ha, L. Q., Stewart, D. W., Hanna, P. J., & Smith, F. J. (2006). Zipf and type-token rules for the English, Spanish, Irish and Latin languages. *Web Journal of Formal, Computational and Cognitive Linguistics, 1*(8), 1–12.

Hatzigeorgiu, N., Mikros, G., & Carayannis, G. (2001). Word length, word frequencies and zipf's law in the Greek language. *Journal of Quantitative Linguistics, 8*(3), 175–185.

Howes, D. (1964). Application of the word-frequency concept to aphasia. In A. V. S.d. Reuck & M.O'Connor (Eds.), *Disorders of language. Ciba foundation symposium* (pp. 47–78). London: J. & A. Churchill,.

Howes, D., & Geschwind, N. (1964). Quantitative studies of aphasic language. *Research Publications–Association for Research in Nervous and Mental Disease, 42*, 229–244.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*, 1286–1307.

MacWhinney, B., & Osmán-Sági, J. (1991). Inflectional marking in Hungarian aphasics. *Brain and Language, 41*, 165–183.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication theory, 84*, 486–502.

Németh, G., & Zainkó, C. (2002). Multilingual statistical text analysis, Zipf's law and Hungarian speech generation. *Acta Linguistica Hungarica, 49*, 385–405.

Piotrovskii, R. G., Pashkovskii, V. E., & Piotrovskii, V. R. (1994). Psychiatric linguistics and automatic text processing. *Automatic documentation and mathematical linguistics translations of selected articles from nauchno-tekhnicheskaia informatsiia, 28*, 28–28.

Piotrowski, R. G., & Spivak, D. L. (2007). Linguistic disorders and pathologies: Synergetic aspects. In P. Grzybek & R. Köhler (Eds.), *Exact methods in the study of language and text. To honor Gabriel Altmann* (pp. 545–554). Berlin: Gruyter.

Ridley, D. R. (1982). Zipf's Law in transcribed speech. *Psychological Research, 44*, 97–103.

Ridley, D. R., & Gonzales, E. A. (1994). Zipf's law extended to small samples of adult speech. *Perceptual and Motor Skills, 79*, 153–154.

van Egmond, M., van Ewijk, L., & Avrutin, S. (2015). Zipf's law in non-fluent aphasia. *Journal of Quantitative Linguistics, 22*, 233–249.

Vogt, P. (2004). Minimum cost and the emergence of the Zipf-Mandelbrot law. Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems. Cambridge, MA: MIT Press.

Wernicke, C. (1874). *The aphasia symtom complex: A psychological study on an anatomical basis*. Reprinted in G. Eggert (1977). *Wernicke's works on aphasia: A sourcebook and review*. Berlin: Mouton de Gruyter.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press.

Zurif, E., Swinney, D., Prather, P., Solomon, J., & Bushell, C. (1993). An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. *Brain and Language, 45*, 448–464.

Appendix 1. Detailed information per speaker

	English					Greek					Hungarian				
	Speaker ID	*Age*	*Sex*	*Type of Aphasia*	*Time after injury*	*Speaker ID*	*Age*	*Sex*	*Type of Aphasia**	*Time after injury*	*Speaker ID*	*Speaker ID*	*Sex*	*Type of Aphasia*	*Time after injury*
Fluent Aphasics	ACWT11a	48;4	male	Wernicke	N/A	02_combined	71	female	Fluent	17 months	c01_combined	c01_combined	male	Conduction	9.5 months
	adler23a	81;3.	male	Wernicke	N/A	04_combined	78	female	Fluent	3.5 years	c02_combined	c02_combined	female	Conduction	4 months
	BU10a	75;9.	male	Wernicke	N/A	15_combined	72	male	Fluent	8 days	c05_combined	c05_combined	male	Conduction	3 months
	elman12a	57;4.	male	Wernicke	N/A	19_combined	64	female	Fluent	9 days	w02_combined	w02_combined	female	Wernicke's	7 months
	kansas14a	77;4.	female	Wernicke	N/A	20_combined	74	female	Fluent	4 months	w04_combined	w04_combined	female	Wernicke's	4 months
	kansas23a	75;6.	female	Wernicke	N/A	30_combined	60	female	Fluent	20 days	.	.	.	.	.
	kurland01c	58;9.	male	Wernicke	N/A	32_combined	73	male	Fluent	20 days	.	.	.	.	.
	kurland01d	59;0.	male	Wernicke	N/A	33_combined	34	female	Fluent	8 days	.	.	.	.	.
	scale11b	91;8.	female	Wernicke	N/A	34_combined	68	male	Fluent	4.5 years	.	.	.	.	.
	scale24a	61;9.	male	Wernicke	N/A	37_combined	58	female	Fluent	2.5 months	.	.	.	.	.
	tucson03a	46;9.	female	Wernicke	N/A	38_combined	72	female	Fluent	3.5 years	.	.	.	.	.
	tucson13a	68;2.	male	Wernicke	N/A	39_combined	55	female	Fluent	20 days	.	.	.	.	.
	whiteside10a	66;2.	male	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	williamson23a	60;9.	male	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	garrett01a	76;8.	female	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	kansas05a	69;10.	female	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	thompson03a	67;6.	male	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	thompson05a	63;10.	female	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	tucson15a	74;1.	male	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	elman14a	76;3.	female	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	whiteside14a	40;6.	female	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	ACWT10a	48;4.	male	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	adler06a	70;7.	male	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.
	kansas12a	N/A	male	Wernicke	N/A	.	.	.	.	.	.	.	.	.	.

Non-Fluent Aphasics	adler13a	52;4.	male	Broca	N/A	05_combined	58	male	Non-fluent	4 weeks	b05_combined	b05_combined	male	Broca's	7.5 moths
	adler16a	63;6.	male	Broca	N/A	09_combined	56	male	Non-fluent	3 months	b07_combined	b07_combined	female	Broca's	21 months
	adler25a	66;2.	male	Broca	N/A	11_combined	50	male	Non-fluent	20 months	a06_combined	a06_combined	male	Anomia	10 weeks
	BU07a	52;4.	male	Broca	N/A	35_combined	86	female	Non-fluent	2 days	a07_combined	a07_combined	male	Anomia	10 months
	BU08a	*64;6.*	*male*	Broca	N/A	36_combined	63	male	Non-fluent	6 days	a11_combined	a11_combined	male	Anomia	6 months
	elman03a	*64;6.*	*male*	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	elman06a	76;10.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	elman11a	52;1.	male	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	fridriksson03a	46;3.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	fridriksson10a	64;9.	male	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	fridriksson12a	47;10.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	kempler03a	64;6.	male	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	kempler04a	60;3.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	kurland10b	78;4.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	scale01a	78;3.	male	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	scale15b	59;4.	male	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	scale25a	52;7.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	scale26a	58;9.	male	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	tap13a	49;3.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	tap19a	N/A	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	tcu02a	42;7.	male	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	tcu03a	41;9.	male	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	tcu07a	49;2.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	tcu08a	57;2.	male	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	whiteside15a	53;10.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
	wright206a	39;0.	female	Broca	N/A	.	.	.	.	.	.	.	.	.	.
Healthy Controls	capilouto02a	85;2.	male	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto03a	75;0.	female	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto04a	80;6.	female	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto05a	72;3.	male	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto06a	82;4.	female	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto07a	72;0.	female	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto08a	74;0.	male	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto09a	82;7.	male	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto10a	72;11.	male	.	.	.	.	.	.	.	.	.	.	.	.

(*Continued*)

Appendix 1. (*Continued*)

English					Greek					Hungarian				
Speaker ID	*Age*	*Sex*	*Type of Aphasia*	*Time after injury*	*Speaker ID*	*Age*	*Sex*	*Type of Aphasia**	*Time after injury*	*Speaker ID*	*Speaker ID*	*Sex*	*Type of Aphasia*	*Time after injury*
capilouto12a	54;11.	female	.	.	.	.	.	.	.	.	.	.	.	.
capilouto13a	71;5.	female	.	.	.	.	.	.	.	.	.	.	.	.
capilouto14a	81;1.	male	.	.	.	.	.	.	.	.	.	.	.	.
capilouto15a	71;10.	male	.	.	.	.	.	.	.	.	.	.	.	.
capilouto16a	79;11.	female	.	.	.	.	.	.	.	.	.	.	.	.
capilouto17a	71;3.	female	.	.	.	.	.	.	.	.	.	.	.	.
capilouto18a	64;4.	female	.	.	.	.	.	.	.	.	.	.	.	.
capilouto19a	60;9.	male	.	.	.	.	.	.	.	.	.	.	.	.
capilouto20a	71;6.	female	.	.	.	.	.	.	.	.	.	.	.	.
capilouto21a	74;6.	male	.	.	.	.	.	.	.	.	.	.	.	.
capilouto23a	70;6.	male	.	.	.	.	.	.	.	.	.	.	.	.
capilouto24a	70;8.	male	.	.	.	.	.	.	.	.	.	.	.	.
capilouto26a	77;0.	male	.	.	.	.	.	.	.	.	.	.	.	.
capilouto28a	76;9.	male	.	.	.	.	.	.	.	.	.	.	.	.
capilouto29a	71;6.	male	.	.	.	.	.	.	.	.	.	.	.	.
capilouto31a	72;2.	female	.	.	.	.	.	.	.	.	.	.	.	.

*The compilers of the corpus only used the labels *Fluent* and *Non-fluent* to describe the type of aphasia for the Greek aphasic groups

Appendix 2. Detailed results per speaker

	English						Greek						Hungarian					
	Speaker ID	*Types*	*Tokens*	*α*	*β*	R^2	*Speaker ID*	*Types*	*Tokens*	*α*	*β*	R^2	*Speaker ID*	*Types*	*Tokens*	*α*	*β*	R^2
Fluent Aphasics	ACWT11a	75	200	1.127	1.428	0.988	02_combined	107	200	0.766	2.622	0.971	c01_combined	104	200	0.853	0.177	0.99
	adler23a	82	200	1.1	4.983	0.978	04_combined	115	200	0.596	−0.237	0.973	c02_combined	112	200	0.771	−0.288	0.992
	BU10a	86	200	1.029	2.352	0.919	15_combined	99	200	0.74	0.813	0.978	c05_combined	97	200	0.778	−0.218	0.987
	elman12a	103	200	1.372	9.658	0.96	19_combined	104	200	0.67	0.092	0.954	w02_combined	97	200	1.05	4.586	0.981
	kansas14a	93	200	0.993	0.701	0.994	20_combined	115	200	0.653	0.285	0.97	w04_combined	94	200	0.651	−0.402	0.938
	kansas23a	99	200	0.972	1.515	0.98	30_combined	125	200	0.658	2.198	0.931	.	.	.	.	.	.
	kurland01c	97	200	1.241	5.139	0.956	32_combined	116	200	0.552	−0.618	0.976	.	.	.	.	.	.
	kurland01d	89	200	1.26	5.966	0.983	33_combined	98	200	0.846	2.236	0.937	.	.	.	.	.	.
	scale11b	93	200	1.16	7.236	0.974	34_combined	148	200	0.562	0.413	0.954	.	.	.	.	.	.
	scale24a	80	200	1.116	7.696	0.977	37_combined	109	200	0.606	0	0.971	.	.	.	.	.	.
	tucson03a	86	200	1.132	3.836	0.969	38_combined	110	200	0.704	0.72	0.943	.	.	.	.	.	.
	tucson13a	90	200	0.844	0.461	0.99	39_combined	121	200	0.604	1.111	0.952	.	.	.	.	.	.
	whiteside10a	112	200	1.195	6.15	0.948		.	.	.	.	.	.	.	.	.	.	.
	williamson23a	75	200	1.016	2.115	0.976		.	.	.	.	.	.	.	.	.	.	.
	garrett01a	78	200	1.004	2.003	0.992		.	.	.	.	.	.	.	.	.	.	.
	kansas05a	88	200	1.172	5.174	0.987		.	.	.	.	.	.	.	.	.	.	.
	thompson03a	62	200	1.126	1.78	0.985		.	.	.	.	.	.	.	.	.	.	.
	thompson05a	92	200	1.135	4.312	0.951		.	.	.	.	.	.	.	.	.	.	.
	tucson15a	101	200	1.074	4.237	0.988		.	.	.	.	.	.	.	.	.	.	.
	elman14a	86	200	0.843	−0.353	0.993		.	.	.	.	.	.	.	.	.	.	.
	whiteside14a	102	200	0.898	1.056	0.966		.	.	.	.	.	.	.	.	.	.	.
	ACWT10a	75	200	0.867	1.698	0.982		.	.	.	.	.	.	.	.	.	.	.
	adler06a	89	200	1.003	1.293	0.983		.	.	.	.	.	.	.	.	.	.	.
	kansas12a	83	200	1.116	3.422	0.927		.	.	.	.	.	.	.	.	.	.	.
Non-Fluent Aphasics	adler13a	50	200	2.107	6.422	0.972	05_combined	107	200	0.759	1.255	0.947	b05_combined	91	200	0.756	−0.621	0.884
	adler16a	67	200	0.89	−0.433	0.98	09_combined	106	200	0.617	−0.834	0.991	b07_combined	97	200	0.776	−0.465	0.988
	adler25a	55	200	1.063	−0.115	0.989	11_combined	101	200	0.723	−0.506	0.945	a06_combined	106	200	0.685	−0.452	0.99
	BU07a	86	200	1.191	2.94	0.983	35_combined	109	200	0.662	−0.256	0.959	a07_combined	118	200	0.774	0.176	0.957
	BU08a	82	200	1.054	2.213	0.984	36_combined	111	200	0.8	6.855	0.949	a11_combined	101	200	0.805	0.107	0.93
	elman03a	94	200	1.05	2.329	0.976	.	.	.	.	.	.	.	.	.	.	.	.
	elman06a	71	200	1.603	12.515	0.987	.	.	.	.	.	.	.	.	.	.	.	.

(*Continued*)

Appendix 2. (*Continued*)

	English						Greek						Hungarian					
	Speaker ID	*Types*	*Tokens*	*α*	*β*	R^2	*Speaker ID*	*Types*	*Tokens*	*α*	*β*	R^2	*Speaker ID*	*Types*	*Tokens*	*α*	*β*	R^2
	fridriksson03a	77	200	1.174	4.102	0.977	.	.	.	.	.	.	.	.	.	.	.	.
	fridriksson10a	48	200	0.817	−0.71	0.995	.	.	.	.	.	.	.	.	.	.	.	.
	fridriksson12a	75	200	1.276	4.885	0.972	.	.	.	.	.	.	.	.	.	.	.	.
	kempler03a	103	200	0.81	−0.562	0.991	.	.	.	.	.	.	.	.	.	.	.	.
	kempler04a	75	200	0.977	1.806	0.965	.	.	.	.	.	.	.	.	.	.	.	.
	kurland10b	103	200	0.773	0.177	0.979	.	.	.	.	.	.	.	.	.	.	.	.
	scale01a	75	200	1.178	4.35	0.955	.	.	.	.	.	.	.	.	.	.	.	.
	scale15b	86	200	0.988	1.978	0.979	.	.	.	.	.	.	.	.	.	.	.	.
	scale25a	75	200	1.136	4.043	0.956	.	.	.	.	.	.	.	.	.	.	.	.
	scale26a	84	200	0.717	−0.607	0.981	.	.	.	.	.	.	.	.	.	.	.	.
	tap13a	76	200	1.208	4.073	0.969	.	.	.	.	.	.	.	.	.	.	.	.
	tap19a	72	200	1.191	1.203	0.985	.	.	.	.	.	.	.	.	.	.	.	.
	tcu02a	69	200	1.362	4.112	0.987	.	.	.	.	.	.	.	.	.	.	.	.
	tcu03a	91	200	0.917	0.545	0.985	.	.	.	.	.	.	.	.	.	.	.	.
	tcu07a	75	200	1.124	3.306	0.982	.	.	.	.	.	.	.	.	.	.	.	.
	tcu08a	64	200	0.892	−0.011	0.992	.	.	.	.	.	.	.	.	.	.	.	.
	whiteside15a	87	200	0.899	0.296	0.951	.	.	.	.	.	.	.	.	.	.	.	.
	wright206a	70	200	1.18	2.391	0.979	.	.	.	.	.	.	.	.	.	.	.	.
Healthy Controls	capilouto02a	90	200	0.91	0.963	0.997	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto03a	98	200	0.841	0.229	0.994	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto04a	92	200	0.893	0.633	0.962	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto05a	106	200	0.84	0.275	0.971	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto06a	104	200	0.903	0.981	0.983	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto07a	84	200	0.869	0.406	0.983	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto08a	107	200	0.954	1.127	0.988	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto09a	104	200	0.742	0.014	0.986	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto10a	101	200	0.937	1.655	0.994	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto11a	88	200	0.917	0.753	0.99	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto12a	86	200	0.954	1.134	0.986	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto13a	102	200	0.931	1.341	0.974	.	.	.	.	.	.	.	.	.	.	.	.
	capilouto14a	91	200	0.807	0.376	0.978	.	.	.	.	.	.	.	.	.	.	.	.

capilouto15a	92	200	0.964	1.94	0.992	.	.	.	.	.	.	.	.	.	.	.	.
capilouto16a	111	200	0.811	0.728	0.984	.	.	.	.	.	.	.	.	.	.	.	.
capilouto17a	112	200	0.973	1.974	0.989	.	.	.	.	.	.	.	.	.	.	.	.
capilouto18a	103	200	0.914	0.733	0.984	.	.	.	.	.	.	.	.	.	.	.	.
capilouto19a	88	200	0.802	−0.266	0.959	.	.	.	.	.	.	.	.	.	.	.	.
capilouto20a	108	200	0.913	0.776	0.988	.	.	.	.	.	.	.	.	.	.	.	.
capilouto21a	92	200	0.88	1.107	0.986	.	.	.	.	.	.	.	.	.	.	.	.
capilouto23a	115	200	0.824	0.106	0.977	.	.	.	.	.	.	.	.	.	.	.	.
capilouto24a	100	200	0.898	0.809	0.983	.	.	.	.	.	.	.	.	.	.	.	.
capilouto26a	100	200	0.853	0.43	0.992	.	.	.	.	.	.	.	.	.	.	.	.
capilouto28a	101	200	0.917	0.718	0.973	.	.	.	.	.	.	.	.	.	.	.	.
capilouto29a	107	200	0.954	1.625	0.984	.	.	.	.	.	.	.	.	.	.	.	.
capilouto31a	107	200	0.931	1.141	0.986	.	.	.	.	.	.	.	.	.	.	.	.