Xinying Chen, Kim Gerdes, Sylvain Kahane, Marine Courtin

# The Co-Effect of Menzerath-Altmann Law and Heavy Constituent Shift in Natural Languages

**Abstract:** The present paper tries to link the Menzerath-Altmann law (MAL) to the Heavy Constituent Shift (HCS) phenomenon and discuss their co-effect in human natural languages. We deduce a hypothesis based on MAL and HCS and then try to empirically verify it by investigating multiple language data from Surface-Syntactic Universal Dependencies (SUD). Our results show that the hypothesis is valid across the complete set of typologically diverse languages and the co-effect of MAL and HCS appears to be a very regular universal.

**Keywords:** Menzerath-Altmann law, Heavy Constituent Shift, Surface-Syntactic Universal Dependencies, co-effect, natural languages

## 1 Introduction

Menzerath's law, also known as the Menzerath-Altmann Law (MAL), predicts that the increase of the size of a linguistic construct results in a decrease of the average size of its components (Altmann & Schwibbe 1989; Hřebíček 1995; Cramer 2005a). Despite the success of the research on MAL, it seems that this powerful law is still not strongly connected to other traditional linguistic discussions that go beyond a mere application of the definitions to various linguistic units. In this study, we aim to address this point by linking MAL to the Heavy Constituent Shift (HCS) phenomenon and discuss their co-effect in human natural languages.

The article is organized as follows. Section 2 starts with a reminder of some studies on MAL. HCS is introduced in Section 3 and our co-effect hypothesis is presented in Section 4. Section 5 describes the methodology of this study and the data resource that the study is based on. Section 6 summarizes the results and the conclusion is provided in Section 7.

**Xinying Chen:** Xi'an Jiaotong University, chenxinying@mail.xjtu.edu.cn
**Kim Gerdes:** University Paris Saclay, kim@gerdes.fr
**Sylvain Kahane:** Université Paris Nanterre, sylvain@kahane.fr
**Marine Courtin:** Sorbonne Nouvelle, rinema56@gmail.com

# 2 Menzerath-Altmann Law

The Menzerath-Altmann law is one of the most discussed linguistic laws; the majority of related studies are focused on verifying this law in certain linguistic constructs with different texts and languages as well as trying to interpret the parameters (Altmann 1980 & 2014; Gustison et al. 2016; Cramer 2005b; Mikros & Jiří 2014), for example, examining whether longer words (in the number of syllables) have shorter syllables (in the number of graphemes for phonemes), or if longer clauses (in the number of words) have shorter words (in the number of syllables) in different human languages (Menzerath 1954; Kelih 2010). A small minority of studies discuss the language features that might influence the results of MAL, such as registers (Hou et al. 2020; Xu & He 2020). Meanwhile, MAL has started to transcend quantitative linguistics and is also gaining attention from other disciplines, such as biology (Li 2012; Ferrer-I-Cancho & Forns 2009).

# 3 Heavy Constituent Shift

HCS (Ross 1967; Stallings et al. 1998) is a well-known phenomenon of syntax. Based on the concept of “heavy constituents” that are composed of more words (and syllables) than “light constituents”, it states that heavier constituents tend to be shifted to the end of the clause. Here is the example 5.56 from Ross (1967: 306):

a. I'll ***give*** some **to my good friend from Akron**.
b. I'll ***give*** **to my good friend from Akron** some.

In this example, the constituent ‘to my good friend from Akron’ has six words and it is heavier than the constituent ‘some’ which only has one word. Therefore, it should be shifted to the end of the sentence, as in the further examples from the GUM English treebank of Universal Dependencies (Zeldes 2017): [1]

a. […] I might capture them and ***learn*** from them **the secrets which the moon had brought upon the night**. (fiction_moon-9)
b. […] the bartender will ***recount*** for the customer **the definition of the santorum neologism**. (interview_coktail-15)
c. […] a scenery made of sand and rocks which ***have*** vaguely **the shape of a castle**. (voyage_guadeloupe_17)

[1] These examples have been collected with the following grew-match request: pattern { X - [ obl | advmod ] -> B ; X - [ obj ] -> C ; X << B ; B << C } without { X -> D ; X << D }.

d. […] the adjustments and calculations ***take*** into account **the weighted nature of the data**. (academic_discrimination-51)
e. […] the only candidate who ***embodies*** both physically and philosophically **the growing diversity of the commonwealth**. (interview_libertarian-11)

This commonly observed language phenomenon has been noted by several linguists before Ross (1967). Here are three citations by French linguists from the 18th and 19th centuries, given in Kahane (2020):

"The [complements[2]] must be as close as possible to the governing word, which would not be the case if one were to put the longest [complement] first, which would move the shortest one too far away." (Buffier 1709: 313)

"When several complements fall on the same word, it is necessary to put the shortest one first after the completed word; then the shortest of those that remain and so on until the longest of all, which must be the last. It is important for the clarity of the expression, *cujus summa laus perspicuitas*[3], to move what serves as the complement as little as possible away from a word. However, when several complements contribute to the determination of the same term, they cannot all follow it immediately; and all that remains is to bring the one that we are forced to keep away from it as close as possible to it: this is what we do by putting first the one which is the shortest, and keeping the longest for the end." (Beauzée 1765: 7)

"When several complements fall on the same word, give the most concise form to the one immediately following the complete word and, as you go along, give the complements a more developed and extensive expression." (Weil 1844: 97 )

Note that HCS has first been observed for SVO languages such as French and English, where the complements are produced after the verb that governs them. The term heavy constituent shift has been coined by Ross in the framework of transformational grammar, with the idea that heavy constituents were shifted from some initial position to the final place. For Buffier and Beauzée, light complements must simply be produced before heavy complements. Weil introduced an additional idea: If you want to produce two complements in a given order, make the second one heavier than the first one. In other words, it is not because a complement is heavy that you put it in the second place, it is because it is in the second place that

[2] In the French tradition, complement means argument constituents as well as modifier constituents depending on the verb. We will keep this sense in the paper.

[3] 'whose highest praise is clarity' (a variation of the famous quote from Quintilian's The Orator's Education stating that the oratory's "basic virtue is clarity").

you make it heavier (and, again, there is absolutely no shift in this framing of the phenomenon).

There are still debates concerning the definition of 'heavy'. Although the theoretical discussion is valuable, for the empirical data analysis in this study, we take the operational definition of 'heavy', namely, having more words.

# 4 Co-effect Hypothesis: Combining the Heavy Constituent Shift and the Menzerath-Altmann Law

Both HCS and MAL are associated with constituent size. This suggests that HCS and MAL interfere with each other. We can deduce a hypothesis based on these two premises.

To be more specific, we investigate and compare the size of different constituents in two types of clauses that have either one or two complements to the right of the word X:[4]

1) ~XAB (the word X has two complements A and B to its right, and A precedes B)
2) ~XC (the word X has only one complement C to its right)

We will focus on words X, when it is the verbal head of a clause. *a*, *b*, *c* corresponds to the size (the number of words) of the constituents A, B, and C.

First, according to MAL, we can expect that the average size of two complements (case 1.) is smaller than the size of the unique element (case 2.):

I. $( a + b ) / 2 < c$

And then, according to HCS, we also expect that B is heavier than A:

II. $a < b$

When we combine I & II, we can get that:

[4] Note that this simplified definition allows any number of dependents to the left of X and does not take into account the presence and size of any elements to the left of X which might be part of the projection of X. We will see in Section 6 that taking into account possible elements to the left does not significantly alter the results.

III. $a = ( a + a ) / 2 < ( a + b ) / 2 < c \Rightarrow a < c$

We thus presume that we should observe "$a < c$" in empirical data, and if our hypothesis is validated, this language phenomenon can be seen as the co-effect of MAL and HCS in human natural languages.

# 5 Methodology

Constituents, from the viewpoint of dependency syntax, are projections of a node in the dependency tree, that is the node and all the nodes that it dominates.

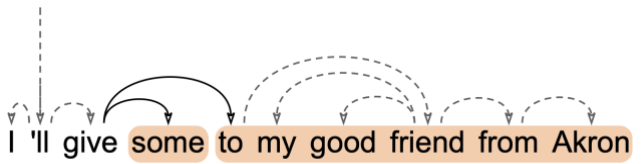


**Fig. 1:** The dependency tree of the sentence 'I'll give some to my good friend from Akron.'

As we can see in Fig. 1, there are two dependencies (bold lines) that fall on the right side of the verb ***give***. Each branch heads one constituent. These two constituents are the two complements of ***give***. In our study, the size of a constituent will be determined by the numbers of words it contains. The two complements of ***give*** on its right, have respectively size one and size six.

What we are investigating in this paper are two types of clauses, namely, ~XAB and ~XC. In which, ~ represents left branches of the tree. It should be noted that here we do not take into consideration the size of possible left tree branches. For instance, the following three sentences would all be considered as ~XAB clauses:

a. I definitely give some to my good friend from Akron.
   (including two left tree branches)
b. I give some to my good friend from Akron.
   (including one left tree branches)
c. Give some to my good friend from Akron.
   (including zero left tree branches)

And all the following three sentences would be considered as ~XC type clauses:

d. I probably did the job. (including two left tree branches)
e. I did the job. (including one left tree branches)
f. Do the job. (including zero left tree branches)

Furthermore, we strictly limited the numbers of the right tree branches to either one or two. For instance, the following clauses would not be considered in our analysis:

g. I tried. (including zero right tree branches)
h. I told her the truth eventually. (including three right tree branches)

To test our hypothesis, we chose the Surface-Syntactic version (SUD 2.7, Gerdes et al. 2018 & 2019) of the Universal Dependencies treebank set (Nivre et al. 2016). The dataset includes 183 treebanks in 104 languages from various typological groups, with a majority of Indo-European languages. For some languages, several treebanks have been developed. In this pilot study, we are more interested in the general picture, and we combine all the treebanks of a language into one collective treebank. Therefore, we take global measures across all trees of each language.

After clearly defining all the conditions, we first filter out ~XAB type and ~XC type clauses from each dependency treebank we study. We only look at X that are verbs and A, B, C that are subjects or complements. More specifically, we only look at the complements with the dependency tag 'subj', 'comp', 'mod', or 'udep' ('udep' is an underdetermined relation that subsumes both 'comp' and 'mod').[5] For each clause we collect, we compute the size of the constituents A, B or C and store them as *a*, *b*, or *c*, and then we calculate the mean value of all *a*, *b*, and *c* on the whole treebank. By comparing the mean value of *a* and *c*, we can either accept or reject our hypothesis.

For the numbers of ~XAB and ~XC clauses in each language, see Tab. 1 in the Appendix.

# 6 Results

We filter out languages with very sparse data that have less than 20 measures of *a* or *c*. This reduces the number of languages to 80. Our results in Fig. 2 show that all

[5] Of course we also take into account all possible extensions of these tags, such as 'comp:obj', 'compl:obl', 'comp:aux', etc.

languages appear above the diagonal. It reflects that our hypothesis "$a < c$" is verified across these typologically different languages (Gerdes et al. 2020).

The colors and shapes in Fig. 2 roughly represent language groups.[6]

- Indo-European languages: triangles
  - Indo-European-Romance: brown
  - Indo-European-Baltoslavic: purple
  - Indo-European-Germanic, including the English Creole *Naija*: olive
  - Other Indo-European: blue
- Sino-Austronesian: green stars
- Agglutinating languages: red plus signs
- Other languages: black squares

For this article, the actual values of *a*, *b* and *c* in each language are presented in Tab. 1 in the Appendix.

[6] We provide an interactive interface on https://typometrics.elizia.net/ allowing for an easy visual exploration of the data as in Fig. 2. The raw data can be found on https://github.com/typometrics/datapreparation.

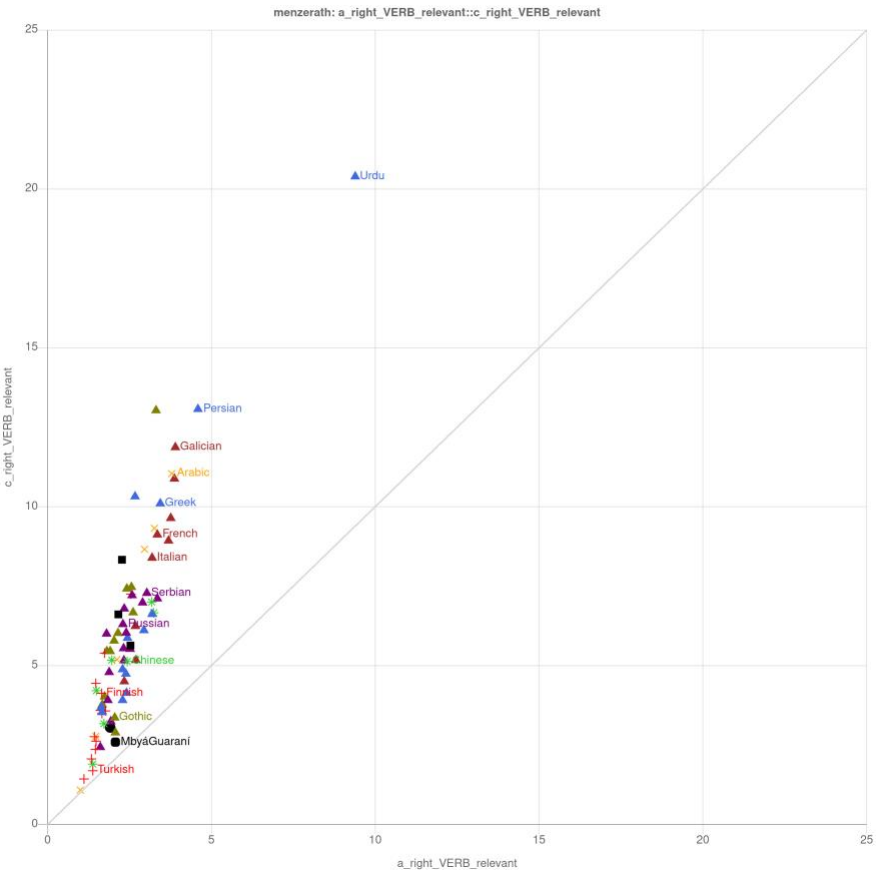


**Fig. 2:** The average size *c* of C constituents is bigger than the average size *a* of A constituents across the 80 languages of SUD 2.7 where we have at least 20 occurrences of corresponding structures.

# 7 Conclusion

Our results show that our hypothesis is valid across the complete set of typologically diverse languages that are present in SUD treebanks. The co-effect of Menzerath-Altmann Law and Heavy Constituent Shift appears to be a very regular universal.

Our pilot study shows that by making use of the recently available coherently annotated multilingual SUD, we can bridge MAL with traditional linguistic discussions such as the HCS, and therefore expand the scope of studies on MAL.

Meanwhile, there are still various details to be investigated in the future. For example, we need to explore what happens to the left of the governor, in particular

for verb-final languages. It might also be worthwhile to verify the measures for all kinds of clauses, not only the clauses that have a verbal head.

Note also that the data resource for languages is unevenly distributed. Some languages, such as German, English, Czech, Arabic, etc., have large treebanks, while treebanks of some languages have very limited sizes. We still have to evaluate in the future how much the sample size would affect the results. Also, even for the same language, treebanks annotated by different teams can vary from each other. We have to consider the effect of fusing treebanks in the future. Last but not least, we can gradually ease the control factors, reduce the constraints for selecting samples, to test the boundary conditions of the co-effect phenomenon.

**Acknowledgement:** This work is supported by the National Social Science Fund of China (18CYY031).

# References

Altmann, Gabriel. 1980. Prolegomena to Menzerath's law. In Rüdiger Grotjahn, editor, *Glottometrika*, 2, 1-10.

Altmann, Gabriel. 2014. Bibliography: Menzerath's law. *Glottotheory*, 5(1):121-123.

Altmann Gabriel & Michael H. Schwibbe. 1989. *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim/Zürich/New York: Olms.

Beauzée, Nicolas. 1767. *Grammaire générale ou Exposition raisonnée des éléments nécessaires du langage, pour servir de fondement à l'étude de toutes les langues* [General grammar or Rational exposition of necessary elements to serve as the foundation for the study of all languages], vol. 2 : Syntax. Barbou, Paris.

Buffier, Claude. 1709. *Grammaire françoise sur un plan nouveau* [French grammar on a new plan]. Paris: Le Clerc- Brunet-Leconte & Montalant.

Cramer, Irene M. 2005a. Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski, editors, *Quantitative Linguistics. An International Handbook*, 659-688. De Gruyter, Berlin / New York.

Cramer, Irene M. 2005b. The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics*, 12(1): 41-52.

Ferrer-I-Cancho Ramon & Núria Forns. 2009. The self-organization of genomes. *Complexity*. 15 (5): 34–36.

Gustison, Morgan L., Stuart Semple, Ramon Ferrer-i-Cancho, & Thore J. Bergman. 2016. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences*, 113(19): E2750-E2758.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane & Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal Dependencies Workshop 2018*. November 1, Brussels.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane & Guy Perrier. 2019. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. In *Treebanks and Linguistic Theories (TLT 2019), Syntaxfest*. August 28-29, Paris.

Gerdes, Kim, Sylvain Kahane & Xinying Chen. 2020. Typometrics From Implicational to Quantitative Universals in Word Order Typology. *Glossa*, forthcoming.

Hou, Renkui, Chu-Ren Huang, Kathleen Ahrens & Yat-Mei Sophia Lee. 2020. Linguistic characteristics of Chinese register based on the Menzerath-Altmann law and text clustering. *Digital Scholarship in the Humanities*, 35(1): 54-66.

Hřebíček, Luděk. .1995. *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law.* Wissenschaftlicher Verlag Trier.

Kahane, Sylvain. 2020. How dependency syntax found its modern form in the French Encyclopedia: from Buffier (1709) to Beauzée (1765). In Nicolas Mazziotta & András Imrényi, editors, *History of dependency-based approaches to grammatical theory*, Benjamins.

Kelih, Emmerich. 2010. Parameter interpretation of Menzerath law: evidence from Serbian. In Peter Grzybek, Emmerich Kelih & Ján Mačutek, editors, *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives*, 71-79. Praesens, Wien.

Li, Wentian. 2012. Menzerath's law at the gene-exon level in the human genome. *Complexity*. 17 (4): 49–53.

Mačutek, Jan, Radek Čech & Jiří Milička. 2017. Menzerath-Altmann Law in Syntactic Dependency Structure. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 100-107. Linköping Electronic Conference Proceedings.

Menzerath, Paul. 1954. *Die Architektonik des deutschen Wortschatzes*. Dümmler, Bonn.

Mikros, Georgios, & Jiří Milička. 2014. Distribution of the Menzerath's law on the syllable level in Greek texts. In Fengxiang Fan, Emmerich Kelih, Reinhard Köhler, Ján Mačutek & Eric S. Wheeler, editors, *Empirical approaches to text and language analysis*, 180-189. RAM-VERLAG, Lüdenscheid.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC2016)*, 1659-1666. May 23-28, Portorož (Slovenia).

Ross, John Robert. 1967. *Constraints on variables in syntax*. PhD thesis. MIT.

Stallings, Lynne M., Maryellen C. MacDonald & Padraig G. O'Seaghdha. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39(3): 392-417.

Weil, Henri. 1844. *De l'ordre des mots dans les langues anciennes comparées aux langues modernes* [About word order in ancient languages in comparison to modern languages]. Crapelet, Paris.

Xu, Lirong, & Lianzhen He. 2020. Is the Menzerath-Altmann law specific to certain languages in certain registers?. *Journal of Quantitative Linguistics*, 27(3): 187-203.

Zeldes, Amir. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3): 581-612.

# Appendix

**Tab. 1:** Values of *a*, *b*, *c* and the numbers of selected clauses in each language.

| **Language** | ***a*** | ***b*** | ***c*** | **Number of ~XAB trees** | **Numbers of ~XC trees** |
|---|---|---|---|---|---|
| Afrikaans | 3.31 | 13.37 | 13.03 | 211 | 1281 |
| Akkadian | 1.9 | 4.4 | 1.71 | 10 | 276 |
| Akuntsu | 0 | 0 | 1.2 | 0 | 10 |
| Albanian | 2.53 | 6.16 | 5.98 | 19 | 51 |
| Amharic | 1.0 | 1.08 | 1.08 | 49 | 326 |
| AncientGreek | 2.41 | 5.2 | 4.14 | 8725 | 25192 |
| Apurinã | 1.08 | 2.67 | 1.61 | 12 | 54 |
| Arabic | 3.79 | 13.43 | 11.04 | 27627 | 25993 |
| Armenian | 3.19 | 8.91 | 6.62 | 182 | 2274 |
| Assyrian | 1.14 | 1.43 | 3.46 | 7 | 26 |
| Bambara | 2.53 | 9.27 | 5.63 | 168 | 1122 |
| Basque | 1.91 | 3.4 | 3.05 | 551 | 3997 |
| Belarusian | 2.33 | 5.07 | 5.16 | 4060 | 14177 |
| Bhojpuri | 7.1 | 9.9 | 9.95 | 10 | 74 |
| Breton | 2.29 | 4.44 | 3.91 | 228 | 480 |
| Bulgarian | 2.52 | 5.56 | 5.52 | 2491 | 9637 |
| Buryat | 1.0 | 12.33 | 5.2 | 3 | 81 |
| Cantonese | 1.96 | 5.21 | 5.17 | 85 | 647 |
| Catalan | 3.87 | 10.7 | 10.88 | 9561 | 22072 |
| Chinese | 2.43 | 5.14 | 5.13 | 564 | 21956 |
| Chukot | 1.11 | 1.94 | 1.43 | 54 | 254 |

| | | | | | |
|---|---|---|---|---|---|
| ClassicalChinese | 1.38 | 2.34 | 1.9 | 1895 | 27462 |
| Coptic | 2.12 | 7.84 | 5.19 | 1634 | 2163 |
| Croatian | 2.9 | 7.22 | 6.98 | 2290 | 9913 |
| Czech | 2.58 | 7.14 | 7.21 | 32884 | 97789 |
| Danish | 2.15 | 6.44 | 6.02 | 2456 | 4502 |
| Dutch | 2.42 | 7.19 | 7.42 | 3039 | 7525 |
| English | 2.61 | 6.3 | 6.67 | 13502 | 36424 |
| Erzya | 1.43 | 2.88 | 2.76 | 291 | 972 |
| Estonian | 1.74 | 4.81 | 5.39 | 9123 | 17000 |
| Faroese | 1.74 | 6.18 | 4.02 | 1144 | 1892 |
| Finnish | 1.65 | 3.99 | 4.13 | 10145 | 22200 |
| French | 3.35 | 9.74 | 9.12 | 21348 | 47025 |
| Gaelic | 2.44 | 8.51 | 5.87 | 2160 | 1247 |
| Galician | 3.9 | 12.0 | 11.87 | 2542 | 7803 |
| German | 2.56 | 7.27 | 7.47 | 30612 | 36399 |
| Gothic | 2.05 | 4.77 | 3.36 | 1598 | 3739 |
| Greek | 3.44 | 10.36 | 10.1 | 1348 | 3094 |
| Hebrew | 3.26 | 10.29 | 9.32 | 3527 | 6222 |
| Hindi | 5.71 | 6.71 | 16.93 | 17 | 4741 |
| HindiEnglish | 2.29 | 4.94 | 4.88 | 250 | 932 |
| Hungarian | 2.54 | 8.82 | 7.25 | 386 | 1334 |
| Icelandic | 1.81 | 6.55 | 5.45 | 23643 | 35949 |
| Indonesian | 3.18 | 7.66 | 7.0 | 2801 | 9876 |
| Irish | 2.94 | 6.94 | 6.11 | 2654 | 1706 |
| Italian | 3.19 | 9.18 | 8.39 | 12833 | 34151 |

| | | | | | |
|---|---|---|---|---|---|
| Japanese | 4.29 | 1.88 | 2.74 | 17 | 61 |
| Karelian | 1.77 | 3.74 | 3.57 | 69 | 138 |
| Kazakh | 0 | 0 | 1.37 | 0 | 19 |
| Khunsari | 0 | 0 | 3.4 | 0 | 5 |
| Komi | 1.66 | 4.59 | 3.49 | 80 | 371 |
| Komi-Permyak | 1.1 | 2.0 | 3.04 | 10 | 53 |
| Korean | 0 | 0 | 2.59 | 0 | 233 |
| Kurmanji | 1.64 | 7.07 | 3.69 | 28 | 304 |
| Latin | 2.7 | 7.19 | 5.17 | 14020 | 43318 |
| Latvian | 2.32 | 6.01 | 5.54 | 3130 | 15899 |
| Lithuanian | 3.35 | 7.33 | 7.11 | 709 | 5271 |
| Livvi | 1.61 | 4.63 | 3.59 | 38 | 82 |
| Maltese | 2.96 | 7.9 | 8.66 | 873 | 3547 |
| Manx | 2.39 | 7.05 | 4.74 | 223 | 88 |
| Marathi | 2.8 | 1.6 | 3.1 | 5 | 20 |
| MbyáGuaraní | 2.07 | 3.07 | 2.59 | 43 | 300 |
| Moksha | 1.46 | 2.17 | 2.36 | 24 | 88 |
| Mundurukú | 0 | 0 | 2.11 | 0 | 19 |
| Naija | 1.67 | 4.53 | 3.77 | 10948 | 35429 |
| Nayini | 0 | 0 | 5.0 | 0 | 3 |
| NorthSami | 1.48 | 2.61 | 2.62 | 822 | 1669 |
| Norwegian | 2.03 | 6.1 | 5.78 | 14499 | 29070 |
| Old Turkish | 1.0 | 16.0 | 5.0 | 1 | 1 |
| OldChurchSlavonic | 1.61 | 3.67 | 2.43 | 1666 | 4045 |
| OldEastSlavic | 1.93 | 3.61 | 3.24 | 4560 | 8697 |

| | | | | | |
|---|---|---|---|---|---|
| OldFrench | 2.34 | 5.64 | 4.5 | 3952 | 11790 |
| Persian | 4.59 | 11.33 | 13.07 | 228 | 9922 |
| Polish | 1.88 | 4.84 | 4.79 | 10347 | 27611 |
| Portuguese | 3.69 | 9.56 | 8.93 | 10412 | 26994 |
| Romanian | 2.68 | 6.68 | 6.24 | 18678 | 45215 |
| Russian | 2.3 | 6.23 | 6.3 | 19485 | 72153 |
| Sanskrit | 1.66 | 3.18 | 3.53 | 204 | 1001 |
| Serbian | 3.03 | 7.22 | 7.28 | 1314 | 4870 |
| SkoltSami | 1.36 | 2.09 | 3.34 | 11 | 50 |
| Slovak | 1.84 | 4.01 | 3.91 | 1687 | 6982 |
| Slovenian | 1.8 | 5.98 | 6.0 | 2215 | 9008 |
| Soi | 0 | 0 | 6.0 | 0 | 1 |
| South Levantine Arabic | 1.47 | 3.4 | 2.77 | 30 | 56 |
| Spanish | 3.76 | 10.3 | 9.64 | 19650 | 43411 |
| Swedish | 1.91 | 6.01 | 5.45 | 5196 | 8793 |
| SwedishSign | 2.07 | 3.52 | 2.89 | 27 | 88 |
| SwissGerman | 1.57 | 5.29 | 5.93 | 7 | 30 |
| Tagalog | 1.72 | 3.1 | 3.17 | 72 | 69 |
| Tamil | 1.0 | 1.0 | 1.08 | 1 | 39 |
| Telugu | 0 | 0 | 1.32 | 0 | 25 |
| Thai | 3.24 | 6.94 | 6.66 | 781 | 2326 |
| Tupinambá | 8.0 | 5.0 | 0 | 1 | 0 |
| Turkish | 1.38 | 2.79 | 1.69 | 101 | 1561 |
| TurkishGerman | 1.47 | 4.96 | 4.44 | 212 | 606 |

| | | | | | |
|---|---|---|---|---|---|
| Ukrainian | 2.4 | 6.08 | 6.03 | 1946 | 5985 |
| UpperSorbian | 2.34 | 6.11 | 6.79 | 122 | 244 |
| Urdu | 9.39 | 7.32 | 20.39 | 31 | 1737 |
| Uyghur | 1.34 | 2.54 | 2.06 | 50 | 102 |
| Vietnamese | 1.49 | 4.45 | 4.21 | 1226 | 3946 |
| Warlpiri | 1.0 | 1.0 | 1.31 | 2 | 16 |
| Welsh | 2.67 | 9.84 | 10.32 | 798 | 656 |
| Wolof | 2.16 | 7.14 | 6.61 | 881 | 3890 |
| Yoruba | 2.27 | 9.66 | 8.33 | 160 | 376 |