Verifying the Menzerath-Altmann law in the verbal domain in 180 languages

Pegah Faghiri[1,2], Kim Gerdes[2], Sylvain Kahane[1,3]
[1]Paris Nanterre University, Modyco, CNRS [2]Paris-Saclay University, LISN, CNRS
[3]IUF - Institut Universitaire de France
pfaghiri@parisnanterre.fr, kim.gerdes@universite-paris-saclay.fr, skahane@parisnanterre.fr

Abstract

We present a large-scale evaluation of the Menzerath-Altmann law (MAL) in the verbal domain across 180 languages, using the Universal Dependencies (UD) treebank collection (v2.17). MAL predicts that as the number of constituents of a linguistic unit increases, their average size decreases. We propose a metric to estimate the MAL effect across corpora of widely varying sizes and define threshold-based categories to classify languages along a MAL preference cline. Crucially, we analyse the preverbal and postverbal domains separately, in addition to the standard bilateral MAL, and control for potential sampling bias by comparing results across language families (Indo-European vs. non-Indo-European) and syntactic types (VO, OV and no dominant order). Our results confirm MAL as a typologically widespread preference but not an absolute universal: several languages display a trivial or even opposite (anti-MAL) tendency. Furthermore, we uncover a significant asymmetry between the two sides of the verb: the MAL effect is stronger in the postverbal domain, while anti-MAL is stronger in the preverbal domain. VO languages tend to show a stronger MAL preference postverbally, whereas OV languages do so preverbally. These findings challenge the widespread assumption that length-based ordering constraints apply symmetrically on both sides of the verb and contribute new cross-linguistic evidence to the debate on the interaction between dependency length minimization and constituent size.

Keywords: Menzerath-Altmann law, quantitative typology, Universal Dependencies, verbal valency, constituent length, word order

1. Introduction

The current version of the Universal Dependencies (UD) collection of syntactic treebanks allows us to evaluate universal properties across a set of 180 languages (v2.17, **?**). In this paper, we investigate the typological validity of the so-called Menzerath-Altmann law (MAL), which concerns the nature of complexity in linguistic units in terms of the relationship between the size of a given unit and the size of its components (Menzerath, 1928, 1954; Altmann, 1980). Roughly speaking, the greater the number of components of a given unit, the shorter their respective length. In the verbal domain, for instance, this implies that the length of verbal arguments decreases as the valency, or number of verbal dependents, increases (Mačutek et al., 2017). We also aim to propose a reliable metric to evaluate this hypothesis in the unbalanced UD collection, which contains annotated corpora of significantly different sizes—including the so-called LOL languages ("Literate, Official, and with Lots of users"), with millions of tokens, but also a number of corpora from less-studied languages with fewer than a thousand tokens. Furthermore, in order to make typologically relevant claims based on an *ad hoc* language sample that was not designed for such purposes, we compare independent subsets of the collection separately to ensure that our conclusions are replicable beyond the sample as a whole and are not mere artefacts of sampling biases. First, with respect to language families, we consider the two main groups that are fairly equally represented in the collection, that is, Indo-European (IE) and non-Indo-European (non-IE) languages, separately. Second, we control for syntactic language types, namely OV, VO, and no dominant order (NDO), which we define using a token-based approach.[1]

Another important contribution of our study is that we examine preverbal and postverbal domains separately. That is, in addition to our general (or combined) MAL metrics, we calculate two seperate sets of metrics for left/preverbal and right/postverbal arguments, respectively LMAL and RMAL. It is important to brear in mind that constituent size is known to obey other constraints. In particular there is the well-studied Dependency Length Minimization (DLM), which is one of the most widespread accounts of length-based word order preferences across languages (Futrell et al., 2015; Ferrer i Can-

Authors are listed in alphabetical order.

[1]We defined the three types based on a VO score calculated for each language, as the ratio of the `obj` relations with a nominal object occuring after the verb. We labelled VO languages with a score greater than 0.67 (that is, languages in which the VO order is at least twice more widespread than OV), and OV languages with a score less than 0.33.

cho, 2004). According to DLM, natural languages tend to favour orders that minimize dependency lengths. This results in a short-before-long preference (SbL) in the postverbal domain (typical of VO languages) and the reverse, that is a long-before-before preference (LbS), in the preverbal domain (typical of OV languages). In other words, the left and right sides of the verb are assumed to behave as mirror images, showing the same DLM effect but in opposite directions (Hawkins, 2007). Unlike DLM, however, MAL has a direction-free definition. Consequently, that is, assuming that a direction-free length-based effect should apply similarly on each side of the verb, MAL is expected to hold equally on both sides. Relatedly, Chen et al. (2022) showed that the combined effect of MAL and HCS on the postverbal side is a cross-linguistic universal, but left the preverbal domain as an open question.

Our results show that MAL is a clearly widespread typological preference across the languages of the world. It is, however, not an absolute universal principle. Importantly, we find that while MAL is a strong cross-linguistic statistical preference, there are not only languages in which this preference is trivial, but also languages that display the opposite tendency. Furthermore, our results reveal a clear difference between the pre- and postverbal domains. On the one hand, the MAL preference is stronger in the postverbal domain; on the other hand, the anti-MAL preference is stronger in the preverbal domain. Importantly, our findings show that in the postverbal domain, VO languages are more likely to exhibit a strong RMAL preference, while, conversely, OV languages are more likely to exhibit a strong LMAL preference in the preverbal domain.

In the following, we first define MAL and briefly present the previous studies on which we build (Section 2). Next, we present our data extraction method using the UD annotation scheme (Section 3) and the metrics we use in this study (Section 4 and Section 5), while motivating our methodological choices. We then present our results (Section 6), including a closer look at individual languages with unexpected behaviour (Section 6.3), before concluding the paper (Section 7). The results for MAL, LMAL, and RMAL on the 180 languages of UD are given in Table 5 in Appendix A.

2. The Menzerath-Altmann law in the verbal domain

MAL predicts a negative correlation between the average length of constituents and their total number in any given domain. MAL has been previously studied in the verbal domain using the UD annotation scheme in Czech (Mačutek et al., 2017), as well as in version 2.3 of the UD collection covering 76 languages (Tanaka-Ishii, 2021). Chen et al. (2022) linked MAL to the Heavy Constituent Shift (HCS) phenomenon on the postverbal side and showed that the co-effect of MAL and HCS constitutes a very regular universal across 80 languages of SUD 2.7. Our study builds on the latter while proposing a number of substantial improvements in order to draw more reliable and typologically valid conclusions. First, we study MAL in the preverbal and postverbal domains separately, in addition to bilateral MAL. Second, our data extraction is more fine-grained than in Tanaka-Ishii's study (see Section 3). Third, we use the latest version of UD, which, with 180 languages, offers a greater diversity of languages from around the world. Finally, we analyse the data while taking into account both language families and language types, that is, VO, OV, or no dominant order (NDO).

3. Extracting the constituents of verbal constructions in UD

We assume a certain familiarity with the UD annotation scheme in this paper and simply recall that the syntactic structure of all sentences is encoded as a dependency tree. Here we focus on verbal constructions, that is, a verb and all the constituents that depend on it. We consider all lexical verbs and have decided not to restrict our study to main verbs (see (1b)) for at least two reasons: first, including all lexical verbs can double or triple the number of constructions, which is not negligible when the corpus is small; second, in some sentences, the most interesting verb in terms of construction may be subordinated because the main verb is merely a modal or an auxiliary.

All dependents of a verb (in the tree) cannot be considered constituents (of the verbal construction we are interested in). The following are excluded in our data extraction: punctuation signs (attached by the `punct` relation in UD), discourse markers (`discourse` relation; see *alas* in (1a)), juxtaposed clauses or parentheses (`parataxis` relation), conjuncts in a coordination (`conj` relation; (1c)), and coordinating conjunctions (`cc` relation; see *but* in (1a)). The `vocative` relation, which concerns addresses to a participant (*Guys, take it easy!*), is excluded but dislocated arguments (*These guys, I don't trust them.*) are kept in, especially because in some languages the boundary between governed units and dislocated units is difficult to draw. The `aux` relation (see *do* in (1a)), is excluded, considering that auxiliaries are part of the verb form, –compounds (see *out* in (1d)) are excluded as well. Markers of subordination such as subordinating conjunctions and adpositions (`mark` and `case` relations; see *when* in (1b)), which are analysed as

dependents in UD but are generally considered as heads of their constructions, are also discarded.[2]

Examples of English verbal constructions are given in (1) for illustration; the verb is underlined and the constituents are shown between square brackets; filtered elements are given in italics.

(1) a. *But* [I] , *alas* , *do* [not] know [how to see sheep] [through the walls of boxes]. en_littleprince

b. This means that *when* [you] move [a series or category field] [to the filter area and back], previously hidden items are again hidden. en_lines

c. [You] tell [me] [what they look like] *and I'll tell you what they are*. en_childes

d. [One minister] [reportedly] handed *out* [100 dollar 'gifts'] [to journalists attending a press conference for Allawi], [a practice that brings back bad memories to many Iraqis]. en_ewt

Another possibility taken into account is that a constituent may be split into two contiguous parts, for example in cases of extraction or extraposition. These cases are treated as two separate constituents. For instance, in (2a), the relative pronoun *what* is extracted and separated from its governor *do*, and in (2b), the relative clause *that go well with the Santorum cocktail* is extraposed and separated from its governor *dishes*.

(2) a. [What] *do* [you] like [to do] ? en_ewt

b. [What culinary dishes] *would* [you] recommend [that go well with the Santorum cocktail]? en_gum

For each language L and each $n \geq 1$, we extracted all the verbal constructions with n constituents. We then calculated the size of each constituent, excluding punctuation. In other words, we kept all relations except `punct`; if a constituent contains a coordination or a discourse marker, there is no reason to remove them.

We call MAL_n(L) the mean size of the constituents in all verbal constructions with n constituents. Note that we compute MAL_n(L) for a given language L only when we have at least 100 constructions with n constituents. We return to this point in Section 4. We define LMAL_n(L) and RMAL_n(L) in the same way. For LMAL, all constructions with n preverbal constituents (i.e. on the left) were selected, regardless of the other side of the verb. Likewise for RMAL, with all constructions having n postverbal constituents (i.e. on the right).

[2] In her study, Tanaka-Ishii (2021) did not exclude any dependent, except the punctuation.

4. Defining the MAL effect

We call λMAL(L) the function that maps MAL_n(L) to each n. It has been shown by Altmann (1980) that a function such as λMAL(L) should follow a power law. In other words, MAL_n(L) is expected to be roughly proportional to $n^{-\beta}$, where β is a positive value (generally smaller than 1), which we call the *MAL effect* for L. Note that if MAL_n(L) $\approx n^{-\beta}$, then $\log(\text{MAL}_n(\text{L})) \approx -\beta \log(n)$. In other words, $-\beta$ is the slope of the regression line in the log-log representation of λMAL(L). The greater β is, the faster the decrease and the stronger the MAL effect. Figure 1 gives the value of β (1.171) as well as the coefficient R^2 (0.892), which shows that λMAL(German) is close to a power law and that MAL holds for German.

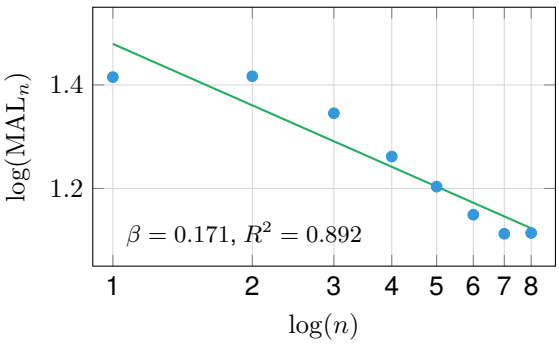


Figure 1: λMAL(German).

Two methodological clarifications are in order at this point. Firstly, there is the question of the statistical reliability of the estimated slope. Indeed, our study faces a methodological challenge in calculating the MAL effect as defined above for our sample, the UD collection, which includes corpora of different sizes, varying from fewer than 1k to more than 1000k tokens. As a result, our samples of extracted tokens, on the basis of which we aim to calculate one MAL effect per language, vary in size. This is a problem because we are estimating coefficients via linear regression modelling and the statistical reliability of these estimates depends on the sample size. In other words, our samples do not offer the same statistical power for the fitted models. To accommodate this issue, we apply a threshold to guarantee a baseline statistical power across samples. We have opted for a fairly high threshold (a minimum of 100 tokens per configuration) in order to maximize the reliability of our MAL metrics, at the cost of losing a number of languages with smaller corpora. Given that we are evaluating the null hypothesis that MAL is a universal preference, our rationale is to favour minimizing the risk of incorrectly detecting counterexamples, at the cost of mislabelling borderline cases.

Secondly, there is the choice of which values

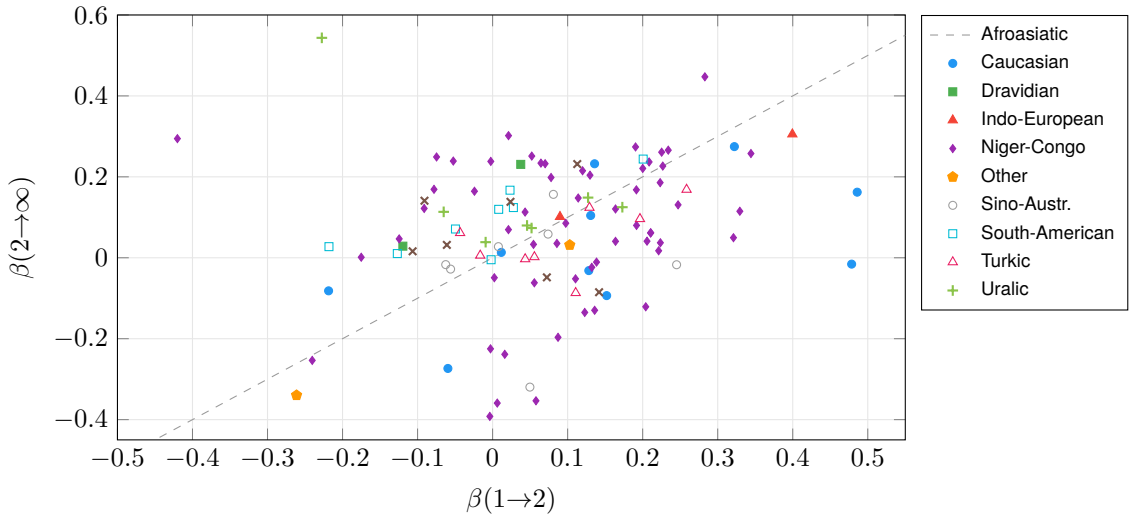


Figure 2: $\beta(1\to2)$ vs $\beta(2\to\infty)$

of n to include or exclude from our calculations. Tanaka-Ishii (2021) showed that, in her data, the values for n = 1 and n = 2 did not always align, and, accordingly, she took only the values for $n \geq 3$ when computing β into account. We do not follow her on this methodological decision, as we believe that the problem in her study may, at least partly, be due to the fact that the data were not filtered - that is, contrary to our data extraction, in her study all verbal dependents were taken into account without excluding any relations. Moreover, since our data is filtered, for many small corpora there may not be enough occurrences to calculate MAL_n for $n \geq 4$. This is even more problematic when we look at LMAL_n and RMAL_n. In other words, excluding MAL_n for $n \leq 2$ is more costly and less straightforward in our study. We therefore considered all options before making our own choice.

We call $\beta(j \to k)$ the slope of λMAL for $j \leq n \leq k$. If $k = j + 1$, $\beta(j \to k)$ is simply the slope between the points $(\log(\text{MAL}_j),\log(j))$ and $(\log(\text{MAL}_k),\log(k))$. In Figure 2, we compare $\beta(1\to2)$ with $\beta(2\to\infty)$. The considerable dispersion suggests that the value for MAL_1 is indeed poorly aligned with the other values. The cases where MAL_1 is aligned correspond to points on the diagonal $y = x$. Nevertheless, we observe that the value of MAL_1 can sometimes fall below the regression line for $n \geq 2$ (as is the case for German, Fig. 1) and sometimes above it (see Arabic, Fig. 3, or Czech, Fig. 4).

Given these observations, we decided to take $\beta(1\to\infty)$ as the metric for evaluating the MAL effect.[3] There are 131 languages in our sample

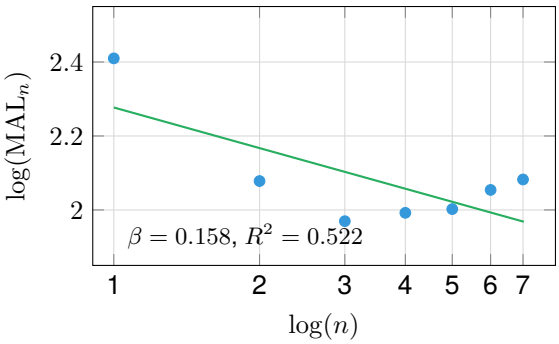


Figure 3: λMAL(Arabic).

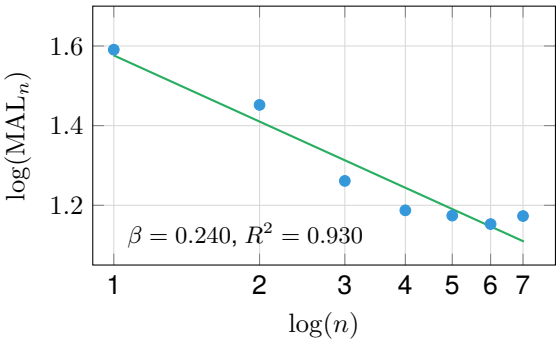


Figure 4: λMAL(Czech).

for which a value for $\beta(1\to\infty)$ is calculated for λMAL, and respectively 124 and 103 for λLMAL and λRMAL. These values mostly fall between -0.3

[3]We should nevertheless mention an additional issue with n = 1. While the number of configurations with n constituents decreases for $n \geq 2$, only 20 of the 131 languages that have at least 100 configurations for n = 2 and n = 3 have more data for n = 1 than for n = 2. Among them, 12 languages have fewer than 100 configurations with one constituent (n = 1), and we must take $\beta(2\to\infty)$ to evaluate the MAL effect.

and 0.3.[4] Given the range of values for $\beta(1\rightarrow\infty)$ in our data, we defined the following three categories. Here again, our aim was to minimize the risk of incorrectly detecting counterexamples, which resulted in a comparatively large borderline (or grey) zone.

- Language L with $\beta(1\rightarrow\infty) > 0.1$ is considered to have a *MAL effect*;
- Language L with $\beta(1\rightarrow\infty) < -0.1$ is considered to have an *Anti-MAL effect*.
- Language L with $-0.1 \leq \beta(1\rightarrow\infty) \leq 0.1$ is considered to be in the grey zone;

The majority of languages in our sample (62 languages) fall into the first category, that is, they show a MAL effect. However, there are also 10 languages that show an Anti-MAL effect. The remaining 59 languages are in the grey zone.

5. MAL compliance rate

In addition to the MAL effect, we define a simple regression-free measure, which we call the MAL compliance ratio. This metric gives the proportion of pairs where MAL_n decreases (for instance, $\text{MAL}_{n+1} \leq \text{MAL}_n$).

We calculated an average compliance ratio for MAL, LMAL, and RMAL for all languages with a corresponding value for $\beta(1\rightarrow\infty)$, and defined the following three MAL compliance types:

- Language L with a MAL compliance ratio > 0.67 is considered to have high MAL compliance;
- Language L with a MAL compliance ratio < 0.33 is considered to have low MAL compliance.
- Language L with $0.33 \leq$ MAL compliance ratio ≤ 0.67 is considered to have medium MAL compliance;

6. Results

6.1. MAL effect in the UD collection

Out of the 186 languages available in our data, we calculated a $\beta(1\rightarrow\infty)$ for 131. In this sample, we identify 62 languages with a MAL preference, 10 with an Anti-MAL preference, and 59 in the grey zone. Importantly, these three categories are fairly comparably represented across language families as well as across the main syntactic language types (Table 1). The only noticeable exception is that languages with no dominant order (ND0) are significantly more likely to be in the grey zone (and less likely to have a MAL preference) than VO or OV languages.[5]

Family / MAL type	IE	Non-IE	**Total**
Anti-MAL	6	4	**10**
Gray zone	22	37	**59**
MAL	38	24	**62**
Total	66	65	**131**

VO/OV / MAL type	NDO	OV	VO	**Total**
Anti-MAL	2	2	6	**10**
Gray zone	13	20	26	**59**
MAL	5	16	41	**62**
Total	20	38	73	**131**

Table 1: MAL type per language family and VO/OV type.

For RMAL (postverbal domain) and LMAL (preverbal domain), we calculated $\beta(1\rightarrow\infty)$ for 124 and 103 languages respectively, out of the total of 186 languages (Table 2 and Table 3). Interestingly, we find a clear difference between the pre- and postverbal domains. On the one hand, while we identify 62 languages (47%) showing a MAL effect, there are 81 languages (79%) showing a MAL effect in the postverbal domain (an RMAL effect), compared to only 39 languages (31%) showing a MAL effect in the preverbal domain (an LMAL effect). In other words, 79% of languages show a MAL effect in the postverbal domain, compared to 31% in the preverbal domain, while the combined MAL effect is found in 47% of languages. On the other hand, while we identify 10 Anti-MAL languages, there are 29 Anti-LMAL languages and 7 Anti-RMAL languages. In other words, 23% of languages show an Anti-MAL effect in the preverbal domain, compared to 7% in the postverbal domain, while 8% show a bilateral Anti-MAL effect.

We can therefore speak of both an RMAL bias and an Anti-LMAL bias—in other words, a preference for MAL in the postverbal domain and for Anti-MAL in the preverbal domain. Importantly, in our data, as with the combined MAL, the three LMAL and RMAL types are also fairly well distributed across language families (Table 2). However, the

[4] For λMAL, no $\beta(1\rightarrow\infty)$ values are below -0.3 and only 9 languages have a $\beta(1\rightarrow\infty)$ greater than 0.3. The situation is similar for λLMAL. For λRMAL, 48 languages have a value greater than 0.3, but for the sake of comparison with keep the same categorization.

[5] More precisely, the difference is significant between NDO and VO languages for the gray zone (respectively, 65% vs. 36%, Fisher's Exact test $p<0.05$), as well for the MAL type (25% vs. 56%, Fisher's Exact test $p<0.05$).

Family RMAL type	IE	Non-IE	Total
Anti-MAL	4	3	7
Gray zone	5	10	15
MAL	51	30	81
Total	60	43	103

Family LMAL type	IE	Non-IE	Total
Anti-MAL	15	14	29
Gray zone	28	28	56
MAL	20	19	39
Total	63	61	124

Table 2: RMAL and LMAL per language family.

same is not true for syntactic language types (Table 3). Nevertheless, the differences we observe in LMAL and RMAL types for OV and VO languages show a familiar pattern and are not suprising.

VO/OV RMAL type	NDO	OV	VO	Total
Anti-MAL	3	2	2	7
Gray zone	2	4	9	15
MAL	14	10	57	81
Total	19	16	68	103

VO/OV LMAL type	NDO	OV	VO	Total
Anti-MAL	2	4	23	29
Gray zone	12	19	25	56
MAL	4	15	20	39
Total	18	38	68	124

Table 3: RMAL and LMAL per VO/OV type.

In the postverbal domain, we find that VO languages show a clearer bias towards RMAL than OV languages. In our sample of 103 languages with enough data on the postverbal domain, we have 68 VO, 16 OV, and 19 NDO languages. Out of the 81 languages showing an RMAL effect, that is, with a significant positive RMAL effect in the postverbal domain, 57 are VO and 10 are OV languages. That is, 83% of our VO languages show a RMAL effect compared to 63% of our OV languages. Moreover, the Anti-RMAL preference is particularly rare among VO languages (2 out of 68 languages, 3%, compared to 2 out of 16, 13%, for OV languages).

Conversely, in the preverbal domain, we observe a larger proportion of OV languages with an LMAL preference than VO languages: 39% (15 out of 38) of OV languages show a LMAL effect, compared to 29% (20 out of 68) of VO languages. Importantly, the Anti-LMAL preference is much more frequent in VO languages than in OV languages: 34% (23 out of 68) of VO languages show an Anti-LMAL effect compared to 11% (4 out of 38) of OV languages.

In other words, on the one hand, VO languages are significantly more likely to show an RMAL preference than OV languages (83% vs. 63%), while also being less likely to show an Anti-RMAL preference (3% vs. 13%). On the other hand, OV languages are more likely to show an LMAL preference than VO languages (39% vs. 29%) and are also less likely to show an Anti-LMAL preference (11% vs. 34%). This latter observation is the most interesting, since the difference between the two language types is the largest. We observe that VO languages are roughly three times more likely to exhibit the Anti-LMAL preference than OV languages and this difference is statistically significant (Fisher's Exact test, $p<0.01$).

These differences are indeed not surprising, given that an asymmetry in length-based effects between head-final/OV and head-initial/VO languages, for directional preferences such as DLM, is well established in the literature. They are also in line with previous observations on the difference between the effect of phrasal length in pre- and postverbal domains (Faghiri and Samvelian, 2020). We return to this point in our concluding remarks.

6.2. MAL compliance in the UD collection

Out of our 131 languages with a MAL effect, 79 show high MAL compliance, 23 show low MAL compliance, and 29 show medium MAL compliance (Table 4). The mean MAL compliance ratios in these three groups are 0.90 (high), 0.13 (low), and 0.50 (middle), respectively. This classification is less strict (and hence less precise), resulting in a smaller number of borderline languages and a greater number of languages at the opposite end of the cline with low MAL compliance, compared to Anti-MAL languages.

In particular, we observe that NDO languages with a notably high number of languages in the grey zone (65%), compared to VO and OV languages (36% and 53%, respectively), based on the MAL effect, show a more balanced distribution with MAL compliance types: 30% are labeled middle, 45% high and 25% low. Borderline languages are still more frequent in NDO than in VO and OV languages, but the distribution of the three MAL compliance types is closer in NDO to what we observe for VO and OV languages, respectively, 19% (middle), 64.5% (high) and 16.5% (low), and 24% (middle), 60% (high) and 16% (low).

Importantly, the average compliance ratio values suggest that both MAL and Anti-MAL languages in our sample are clearly located at the edges of the

cline. Based on these observations, we can expect the number of Anti-MAL languages to increase with a less strict measure and/or classification of the MAL effect.

	MAL compliance			
Language family	high	middle	low	**Total**
IE	38	16	12	**66**
Non-IE	41	13	11	**65**
Total	79	29	23	**131**

	MAL compliance			
VO/OV type	high	middle	low	**Total**
NDO	9	6	5	**20**
OV	23	9	6	**38**
VO	47	14	12	**73**
Total	79	29	23	**131**

Table 4: MAL compliance per language family and VO/OV type.

Note that we observe a similar pattern for RMAL and LMAL, which we do not present here for reasons of space.

6.3. Zooming into languages with unexpected behaviour

It is not possible to present the results for all the languages that contradict a universal MAL hypothesis. All the results are presented in Table 5 in Appendix A. An interactive version with the possibility to sort the languages according to the different values is available online at https://typometrics.elizia.net/menzerath/. Here, we present some notable cases of Anti-MAL languages, as well as some cases of languages with inconsistent behaviour in the pre- and/or postverbal domains.

One interesting example of an Anti-MAL language is Old East Slavic, which not only has an Anti-MAL effect (that is, the mean length of constituents increases with the number of constituents), but moreover displays an Anti-MAL effect that grows stronger with the number of constituents (Figure 5).

Starting with the preverbal domain, we find 29 Anti-LMAL languages, of which 17 are RMAL – and they are mostly in the grey zone when bilateral MAL is considered. Two of these languages are noteworthy: Occitan (Romance, France) is MAL, while Naija (English-based pidgincreole, Nigeria) is Anti-MAL. Naija is also remarkable in that it is the most strongly Anti-LMAL language in our sample, following a near-perfect power law (Figures 6 and 7).

As for the postverbal domain, we find the following seven Anti-RMAL languages: Khoekhoe, Bambara, Egyptian, Old East Slavic, Western Armenian, Gothic, and Latin. It is worth noting that among them, four are corpora of ancient languages: Egyptian, Old East Slavic, Gothic, and Latin. Three are both Anti-MAL and Anti-LMAL as well (Bambara, Old East Slavic, and Latin), the others are in the grey zone (for MAL, LMAL or both). In other words, none of the seven languages shows a MAL or a LMAL preference.

Exploring the genetic and/or areal similarities and differences among these individual languages is beyond the scope of this paper, but it constitutes a necessary avenue for future research in understanding length-based effects across languages and linguistic domains.

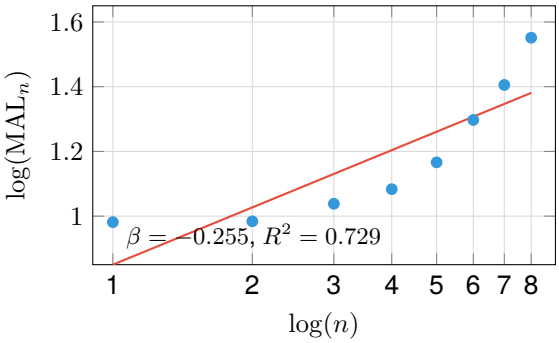


Figure 5: λMAL(OldEastSlavic).

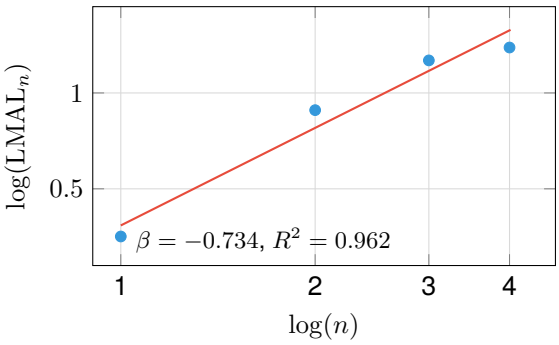


Figure 6: λLMAL(Naija).

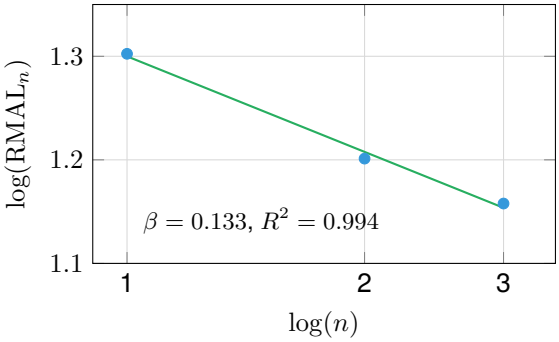


Figure 7: λRMAL(Naija).

7. Conclusion

In this paper, we presented an exhaustive token-based evaluation of MAL in the verbal domain across the UD collection of 180 languages. We proposed a classification based on a metric that we defined in order to estimate the MAL effect for any given language satisfying a minimum threshold of occurrences for relevant configurations. We obtained a reliable classification of 131 languages, well distributed between IE and non-IE languages as well as between VO and OV languages. This classification allowed us to confirm MAL as a typologically widespread preference across the languages of the world, but importantly, it also showed that MAL is not an absolute universal principle. Not only did we find a number of languages in which MAL is absent or trivial, but we also found languages on the opposite end of the cline, displaying a clear MAL dispreference.

The second contribution of our study is the finding that, contrary to what has been previously assumed, MAL does not apply equally on both sides of the verb. MAL is stronger on the right side than on the left side, while Anti-MAL is stronger on the left side. This extends the findings of Chen et al. (2022), who showed the co-effect of MAL and HCS to be universal on the postverbal side, by revealing that the picture is more complex on the preverbal side. This is a crucial finding, as it implies that length-based effects are not necessarily symmetric in the pre- and postverbal domains, contrary to the widespread assumption. Not only do our results show a clear difference between the pre- and postverbal domains, but they also show that VO and OV languages have different preference profiles in these domains. VO languages are more likely to show a strong MAL preference in the postverbal domain, while OV languages are more likely to show a strong MAL preference in the preverbal domain. Importantly, similar observations on the differences between pre- and postverbal domains have previously been made for length-based ordering preferences in individual languages such as Persian (Faghiri and Samvelian, 2020) or Mandarin (Yao, 2018). Moreover, there are alternative psycholinguistic production-based accounts of mirror-image word order preferences (SbL and LbS) in VO and OV languages that posit differences between the two domains. The latter claim that verbs exert strong influence over the constituents that follow them and, accordingly, assume that formal constraints are stricter in the postverbal domain than in the preverbal domain (Stallings et al., 1998). Our study sheds new light on this ongoing debate. All detailed results and interactive visualisations are available at https://typometrics.elizia.net/menzerath/.

8. Bibliographical References

Gabriel Altmann. 1980. Prolegomena to menzerath's law. In *Glottometrika 2*, pages 1–10, Bochum. Brockmeyer.

Xinying Chen, Kim Gerdes, Sylvain Kahane, and Marine Courtin. 2022. The co-effect of Menzerath-Altmann law and heavy constituent shift in natural languages.

Pegah Faghiri and Pollet Samvelian. 2020. Word order preferences and the effect of phrasal length in SOV languages: evidence from sentence production in Persian. *Glossa: a journal of general linguistics*, 5(1):86.

Ramon Ferrer i Cancho. 2004. Euclidean distance between syntactically linked words. *Phys. Rev. E*, 70:056135.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

John A. Hawkins. 2007. Processing typology and why psychologists need to know about it. *New Ideas in Psychology*, 25(2):87–107. Modern Approaches to Language.

Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-altmann law in syntactic dependency structure. In *Proceedings of the fourth international conference on dependency linguistics (Depling)*, pages 100–107.

Paul Menzerath. 1928. Über einige phonetische probleme. In *Actes du premier Congres international de linguistes*, pages 104–105. Sijthoff Leiden.

Paul Menzerath. 1954. *Die architektonik des deutschen wortschatzes*. Dümmler: Bonn, Germany.

Lynne M. Stallings, Maryellen C. MacDonald, and 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.

Kumiko Tanaka-Ishii. 2021. Menzerath's law in the syntax of languages compared with random sentences. *Entropy*, 23(6):661.

Yao Yao. 2018. NP weight effects in word order variation in Mandarin Chinese. *Lingua Sinica*, 4(1):5.

9. Language Resource References

A. MAL results for all languages

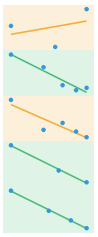
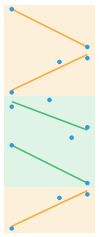
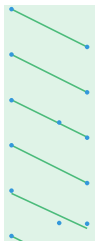
Table 5: MAL, LMAL and RMAL mini log-log plots and $\beta(1 \to \infty)$ for all languages. Green = MAL ($\beta > 0.1$), red = Anti-MAL ($\beta < -0.1$), orange = grey zone.

Language	Family	Type	MAL	β	LMAL	β	RMAL	β
Abkhaz	Caucasia	OV		-0.07		+0.07	—	—
Afrikaans	IE	OV		+0.18		+0.19		+0.55
Akkadian	Afroasia	OV		+0.05		+0.09	—	—
Albanian	IE	VO		-0.05		+0.12		+0.47
Amharic	Afroasia	OV		+0.07		+0.30		-0.00
AncientGreek	IE	mix		+0.05		-0.06		+0.01
AncientHebrew	Afroasia	VO		+0.03		-0.13		+0.03
Arabic	Afroasia	VO		+0.16		-0.12		+0.10
Arhuaco	South-Am	OV		+0.13		+0.25	—	—
Armenian	IE	mix		+0.06		-0.05		+0.15
Bambara	Nig-Co.	OV		-0.29		-0.51		-0.20
Basque	Other	OV		+0.06		+0.18		+0.11
Bavarian	IE	mix		+0.13		+0.09		+0.75
Beja	Afroasia	OV		+0.38		+0.40	—	—
Belarusian	IE	VO		+0.18		+0.09		+0.34
Bhojpuri	IE	OV		-0.04		+0.09	—	—
Borôro	South-Am	OV		+0.13		+0.23		+0.10
Breton	IE	VO		+0.12	—	—		+0.20
Buglere	South-Am	OV		-0.01		-0.05	—	—
Bulgarian	IE	VO		+0.17		+0.17		+0.33
Buryat	Other	OV		+0.16		+0.19	—	—
Cantonese	Sino-Aus	VO		-0.05		-0.24		+0.39
CappadocianGreek	IE	VO	—	—	—	—		+0.24
Catalan	IE	VO		+0.26		+0.25		+0.36
Chhintange	Sino-Aus	OV		+0.21		+0.19	—	—
Chinese	Sino-Aus	VO		-0.04		-0.15		+0.53
Chukot	Other	mix		-0.05		+0.01	—	—
ClassArmenian	IE	mix		+0.00		-0.07		+0.01
ClassChinese	Sino-Aus	VO		+0.07		+0.04		+0.01
Coptic	Afroasia	VO		-0.15		-0.49		-0.03
Croatian	IE	VO		+0.27		+0.15		+0.35
Czech	IE	VO		+0.24		+0.22		+0.42
Danish	IE	VO		+0.13		+0.26		+0.31
Dutch	IE	mix		+0.09		+0.15		+0.49

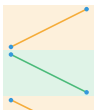
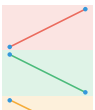
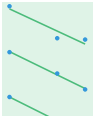
continued on next page

Language	Family	Type	MAL	β	LMAL	β	RMAL	β
Egyptian	Afroasia	VO		+0.08	—	—		-0.20
English	IE	VO		+0.06		-0.30		+0.19
Erzya	Uralic	mix		-0.02		-0.00		+0.27
Estonian	Uralic	VO		+0.21		+0.18		+0.45
Faroese	IE	VO		-0.08		-0.01		+0.02
Finnish	Uralic	VO		+0.11		+0.06		+0.38
French	IE	VO		+0.25		+0.01		+0.32
FrisianDutch	IE	mix		+0.33	—	—	—	—
Gaelic	IE	VO		+0.35	—	—		+0.24
Galician	IE	VO		+0.39		+0.24		+0.50
Gbaya	Nig-Co.	VO	—	—	—	—		+0.17
Georgian	Caucasia	mix		+0.15		-0.04		+0.44
German	IE	mix		+0.17		+0.20		+0.54
Gheg	IE	VO		+0.11		-0.04		+0.31
Gothic	IE	VO		-0.12		-0.07		-0.12
Greek	IE	VO		+0.21		-0.04		+0.32
Guajajára	South-Am	VO		+0.08	—	—		+0.18
Haitian	IE	VO		+0.19		-0.29		+0.30
Hausa	Afroasia	VO		-0.16		-0.40		+0.07
Hebrew	Afroasia	VO		+0.19		+0.20		+0.32
HighlandNahuatl	South-Am	VO		+0.23		+0.10		+0.43
Hindi	IE	OV		+0.10		+0.03		+1.03
Hungarian	Uralic	mix		+0.05		+0.08		+0.24
Icelandic	IE	VO		-0.03		-0.17		+0.22
Indonesian	Sino-Aus	VO		+0.11		-0.07		+0.31
Irish	IE	VO		+0.05		-0.31		+0.15
Italian	IE	VO		+0.15		+0.00		+0.22
Japanese	Other	OV		+0.03		+0.02	—	—
Javanese	Sino-Aus	VO		+0.22		-0.00		+0.21
Karo	South-Am	OV		+0.05		+0.07	—	—
Kazakh	Turkic	OV		+0.06		+0.07	—	—
Khoekhoe	Other	OV		-0.04		+0.07		-0.37
Kiche	South-Am	VO		+0.05	—	—		+0.21
Komi	Uralic	mix		-0.04		+0.01		+0.20
Korean	Other	OV		+0.13		+0.12		-0.06
Kurmanji	IE	OV		+0.14		-0.13	—	—
Kyrgyz	Turkic	OV		+0.15		+0.15	—	—
Latin	IE	mix		-0.23		-0.14		-0.10
Latvian	IE	VO		+0.11		-0.02		+0.30

continued on next page

Language	Family	Type	MAL	β	LMAL	β	RMAL	β
Ligurian	IE	VO		-0.07		+0.05		+0.55
Lithuanian	IE	VO		+0.15		-0.08		+0.38
LowSaxon	IE	mix		+0.10		+0.14		+0.51
Maghrebi	Afroasia	VO		+0.12		+0.15		+0.17
Maltese	Afroasia	VO		+0.30		-0.07		+0.41
Manx	IE	VO	—	—	—	—		+0.39
Marathi	IE	OV		+0.12		+0.10	—	—
MbyáGuaraní	South-Am	mix		-0.01		+0.00	—	—
MiddleFrench	IE	VO		+0.12		-0.29		+0.15
Moksha	Uralic	VO		+0.07		+0.06	—	—
Naija	IE	VO		-0.25		-0.73		+0.13
Nhengatu	South-Am	VO		+0.03		+0.14		+0.10
NorthSami	Uralic	VO		+0.03		-0.00		+0.19
Norwegian	IE	VO		+0.22		+0.18		+0.39
Occitan	IE	VO		+0.21		-0.15		+0.33
OldChurchSlavonic	IE	VO		-0.13		-0.17		-0.08
OldEastSlavic	IE	mix		-0.26		-0.21		-0.16
OldFrench	IE	VO		-0.05		-0.18		+0.15
OldProvençal	IE	VO		-0.20		-0.44		-0.01
OttomanTurkish	Turkic	OV		+0.06		+0.03	—	—
Persian	IE	OV		-0.01		-0.09		+0.68
Pesh	Other	OV		+0.10		-0.04	—	—
Polish	IE	VO		+0.14		+0.12		+0.26
Pomak	IE	VO		+0.05		+0.10		+0.19
Portuguese	IE	VO		+0.13		+0.13		+0.17
Romanian	IE	VO		+0.09		-0.11		+0.21
Russian	IE	VO		+0.11		+0.07		+0.32
Sanskrit	IE	OV		-0.02		+0.03		+0.42
Serbian	IE	VO		+0.17		+0.23		+0.37
Sicilian	IE	VO		+0.32		-0.08		+0.36
Sindhi	IE	OV		-0.05		+0.05		+0.84
SkoltSami	Uralic	mix		+0.03		-0.10	—	—
Slovak	IE	VO		+0.14		+0.24		+0.34
Slovenian	IE	VO		+0.23		+0.06		+0.42
Spanish	IE	VO		+0.21		+0.22		+0.26
Swedish	IE	VO		+0.11		+0.10		+0.35
SwissGerman	IE	OV		-0.00		+0.17		+1.01
Tamil	Dravid.	OV		+0.35		+0.38	—	—
Tangkhul	Sino-Aus	OV		+0.11		+0.09	—	—

continued on next page

Language	Family	Type	MAL	β	LMAL	β	RMAL	β
Tatar	Turkic	OV		-0.08		-0.13	—	—
Teko	South-Am	mix		+0.10		+0.13	—	—
Telugu	Dravid.	OV		+0.09		+0.09	—	—
Thai	Sino-Aus	VO		+0.09		-0.22		+0.30
Turkish	Turkic	OV		+0.14		+0.14		-0.06
TurkishGerman	Turkic	OV		+0.05		+0.13		+0.40
Ukrainian	IE	VO		+0.19		-0.04		+0.32
UpperSorbian	IE	VO		+0.50		+0.50		+0.72
Urdu	IE	OV		+0.14		+0.05		+0.76
Uyghur	Turkic	OV		+0.02		+0.04	—	—
Uzbek	Turkic	OV		+0.14		+0.16	—	—
Vietnamese	Sino-Aus	VO		+0.01		-0.20		+0.24
Welsh	IE	VO		+0.63		-0.02		+0.70
WestArmenian	IE	mix		-0.03		-0.10		+0.17
WestNahuatl	South-Am	VO		+0.04		+0.03		-0.13
Wolof	Nig-Co.	VO		+0.06		-0.32		+0.22
Wu	Sino-Aus	VO		-0.00		-0.05	—	—
Xibe	Other	OV		-0.13		-0.13	—	—
Yiddish	IE	VO		+0.04		-0.10		+0.29
Yoruba	Nig-Co.	VO		+0.34		-0.21		+0.49
Yupik	Other	mix		-0.05	—	—	—	—
Zaar	Afroasia	VO		+0.01		-0.28		+0.35