Abstract. We discuss a discrete and a continuous approach for modelling the distribution of morph length. The proposed models (1-displaced extended binomial distribution, beta function) have been successfully fitted to Spanish, Russian and Slovenian data records. It can be shown that the models are suitable for modelling word form types as well as word form tokens.

Keywords: Morph length, Spanish, Russian, Slovenian

0. Introduction

Unlike sentence, syllable and word lengths, the distribution of the morph length has so far received little attention in the literature (see Best 2001, 2005a). In the studies of Best it is assumed that the morph length basically follows the same laws as the word length, which can be justified linguistically by a variable proportional relationship or a generalized relation (see for example Eq. (1) and (3) in Best (2005 b)). It is now reasonable to try to fit these probabilistic models to morph length distributions that have proved to be adequate for the word length. In particular, Best (2001) has successfully fitted the 1-shifted Hyperpoisson distribution to the distribution of morph lengths in journalistic texts of a German newspaper and Rottmann (2003) has fitted the extended positive binomial distribution (EPB) to word length distributions (measured in the number of syllables) of Latvian and Lithuanian texts. Generally, it is an inductive attempt to find a suitable model based on the proportionality approach of Wimmer and Altmann (2005).

In the present paper, we show that the EPB is suitable to model the morph length distribution. We start with a data set of Saporta (1966), which is mentioned by Köhler and Altmann (2009: 78–79) as an open problem of quantitative linguistics. In addition, we provide new data sets from Russian and Slovenian, for which the EPB has also been proved appropriate. The morph length is measured in the number of phonemes and its determination is performed on the level of word form types and word form tokens. Since the morph length is a discrete random variable it is natural to model it by means of a discrete probability model. However, in quantitative linguistics, continuous models have also been considered to describe discrete observations, since “discrete” and “continuous” are considered to be mere conceptual properties see for example Tuzzi et al. (2012, section 3) and in particular Mačutek and Altmann (2007). Therefore we also present as an alternative to the discrete approach a continuous model based on the beta function to describe the morph length.
1. Spanish data record (Saporta, 1966)

Before re-analyzing the data from Saporta (1966: 69) it should be noted that the author does not provide further information about the performed morph segmentation. From the data record it is only evident that a class of zero-morphs is considered. To ensure the comparability of our study we do not consider zero morphemes as a separate class either in the re-analysis of the Spanish data or in our analysis of the Slovenian and Russian data. In Saporta (1966) the morph length has been measured in the number of phonemes. In Table 1, the observed morph frequency \( f_x \) is listed as a function of the phoneme number \( x \), wherein, as mentioned above, the class of the zero-morphs was omitted and the last five classes have been pooled.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f_x )</th>
<th>discrete, NP(_x)</th>
<th>continuous, y((x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59</td>
<td>59.00</td>
<td>25.40</td>
</tr>
<tr>
<td>2</td>
<td>97</td>
<td>114.95</td>
<td>139.04</td>
</tr>
<tr>
<td>3</td>
<td>307</td>
<td>265.89</td>
<td>284.86</td>
</tr>
<tr>
<td>4</td>
<td>387</td>
<td>378.45</td>
<td>370.30</td>
</tr>
<tr>
<td>5</td>
<td>327</td>
<td>370.34</td>
<td>353.59</td>
</tr>
<tr>
<td>6</td>
<td>261</td>
<td>263.57</td>
<td>257.81</td>
</tr>
<tr>
<td>7</td>
<td>143</td>
<td>140.68</td>
<td>140.41</td>
</tr>
<tr>
<td>8</td>
<td>64</td>
<td>57.21</td>
<td>52.06</td>
</tr>
<tr>
<td>9</td>
<td>19</td>
<td>17.81</td>
<td>10.26</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>5.08</td>
<td>0.47</td>
</tr>
</tbody>
</table>

**EPB d.** \( n = 14, \quad p = 0.2625, \quad \alpha = 0.964, \quad C = 0.0097 \)

**Beta f.** \( m = 0, \quad M = 11, \quad C = 0.000295, \quad a = 3.2027, \quad b = 4.9355, \quad R^2 = 0.9736 \)

Based on the unified theory (cf. Wimmer, Altmann 2005) we suppose that the occurrence probabilities of magnitudes of properties develop in such a way that neighbouring classes stay in proportional relation to each other. Practically,

\[
P_x = f(x)P_{x-1}
\]

where \( f(x) \) is a proportionality function changing with \( x \). In many cases it is simply the ratio \( f(x) = g(x)/h(x) \), where \( g(x) \) contains a language constant and the function of change performed by the speaker, and \( h(x) \) is the controlling function of the hearer. In our case we consider the possibility that the language constant is \( a \), and the effect of the speaker is \(-bx\),

\[
3 \quad \text{The question of whether so-called zero-morphemes must be considered in the analysis of morphs must be decided prior to the examination. In our analysis no zero-morphemes are considered. However, one should bear in mind that considering a “zero” class causes a number of serious methodological and theoretical problems. The situation is quite similar to that of zero-syllabic words in Slavic languages – a problem which has been discussed at great length in Antić, Kelih and Grzybek (2006). When considering zero-classes in statistical modelling, in particular the following must be considered: the number of zero-morphemes is probably very low in any language and this low frequency presumably causes a “high jump” from this class to the class of morphemes of length 1, which in turn can lead to difficulties in modelling.}
\]
while the control function of the hearer is \( h(x) = cx \). Inserting these assumptions into (1) we obtain

\[
(2) \quad P_x = f(x)P_{x-1} = \frac{g(x)}{h(x)} P_{x-1} = \frac{a-bx}{cx} P_{x-1} = \frac{a}{b-x} \frac{b}{c} P_{x-1}.
\]

By substituting \( a/b = n + 1 \) and \( b/c = p/q \), we obtain the well-known recurrence formula

\[
(3) \quad P_x = \frac{n-x+1}{x} \frac{p}{q} P_{x-1},
\]

whose solution is the binomial distribution

\[
(4) \quad P_x = \binom{n}{x} p^x q^{n-x}, \quad x = 0,1,\ldots,n.
\]

Now, since morphs of zero lengths are excluded, we truncate (4) at \( x = 1 \) and obtain the truncated binomial distribution

\[
(5) \quad P_x = \binom{n}{x} \frac{p^x q^{n-x}}{1-q^n}, \quad x = 1,2,\ldots,n.
\]

Since the first frequency, i.e. \( P_1 \), yields in each case a special value, we displace the distribution one step to the right and define \( P_1 \) separately by setting \( P_1 = 1 - \alpha \). This yields a modified distribution that can be called 1-displaced extended binomial distribution (1-displaced EPB), defined as

\[
(6) \quad P_x = \begin{cases} 
1 - \alpha, & x = 1 \\
\alpha \left( \frac{n}{x-1} \right) p^{x-1} q^{n-x+1} \frac{1}{1-q^n}, & x = 2,3,\ldots,n+1
\end{cases}
\]

where \( 0 < \alpha, p < 1, q = 1 - p \).

By means of the software Altmann-Fitter (1997), the EPB has been proved to be suitable for the Spanish data.

The probability \( P_1 = 1 - \alpha \) is set equal to the corresponding relative frequency, i.e. the estimator \( \hat{\alpha} \) satisfies

\[
(7) \quad 1 - \hat{\alpha} = \frac{f_1}{N},
\]

where \( N = \sum x f_x \) denotes the number of all morphs (sample size). The parameters \( p \) and \( n \) are iteratively determined.
By fitting the 1-shifted EPB to the data in Table 1, we obtain the theoretical
frequencies in the third column. The penultimate line of Table 1 lists the optimal parameter
values \( n, p, \alpha \) and the contingency coefficient \( C = \chi^2 / N \).

Since the sample size is very large (here as well as in all following data sets) the chi-
square value is also high and, as is often done in quantitative linguistics in this case (see for
example Rottmann (2003: 53)), one can use the coefficient \( C \) as a criterion to decide on the
goodness of fit. A fit is considered good if \( C \leq 0.01 \) and satisfactory if \( C \leq 0.02 \). Therefore
the fit of the EPB in Table 1 can be considered satisfactory.

Our continuous approach is based on the assumption that the relative rate of change of
the morph frequency \( y \) is proportional to the rate of change of the number \( x \) of phonemes, i.e.

\[
\frac{dy}{y} \sim dx.
\]

As in the discrete case we assume that the proportionality is not given by a constant but by a
function \( g(x) \), involving impacts of speaker and hearer. This leads to

\[
\frac{dy}{y} = g(x)dx.
\]

Now the function \( g(x) \) is composed of a difference of speaker and hearer portions. The former
can be expressed as

\[
\frac{a}{x - m},
\]

where \( a \) is the speaker-force and \( m \) the minimum value of \( x \). The farther away \( x \) is from \( m \), the
less the impact of the speaker is. Similarly, the hearer portion is expressed as

\[
\frac{b}{M - x},
\]

where \( b \) is the permanent force of the hearer and \( M \) the maximum value of \( x \). The farther away
\( x \) is from \( M \), the stronger the impact of the hearer is. The two forces are considered to be in
equilibrium.

Expressing the function \( g(x) \) in (9) by (10) and (11), we get the relation

\[
\frac{dy}{y} = \left( \frac{a}{x - m} - \frac{b}{M - x} \right)dx
\]

which has the simple solution

\[
y = C(x - m)^a (M - x)^b \quad \text{for} \quad m \leq x \leq M.
\]

This is a function with five parameters \( C, m, M, a \) and \( b \) \((C > 0; m < M; a, b > -1)\) which can
have many different shapes; for details see Altmann and Grotjahn (1988) and Köhler and
Altmann (1986). The parameters \( m \) and \( M \) can be directly estimated from the observed
Models of morph length

To ensure that the range \([m, M]\) of the model contains the observed \(x\)-values, it must hold \(m \leq x_{\text{min}}\) and \(M \geq x_{\text{max}}\), where \(x_{\text{min}}\) and \(x_{\text{max}}\) are the minimum and maximum of the observed phoneme numbers, respectively. In the present article we have always chosen

\[
m = x_{\text{min}} - 1 \quad \text{and} \quad M = x_{\text{max}} + 1.
\]

We fitted (13) to the data in Table 1, where \(m\) and \(M\) have been chosen as in (14) and \(C, a, b\) are considered as three freely selectable parameters \((C > 0; a, b > -1)\) which have been optimized iteratively. The optimal values of \(C, a\) and \(b\) and the coefficient of multiple determination \(R^2\), used as a measure for the “goodness of fit”, are listed in the last line of Table 1. The coefficient \(R^2\), also known as “proportion of variance explained”, is defined by

\[
R^2 = 1 - \frac{\sum_x (f_x - y(x))^2}{\sum_x (f_x - \bar{f})^2}
\]

where \(\bar{f} = \frac{1}{N(x_{\text{max}} - x_{\text{min}} + 1)} \sum_x f_x = \frac{N}{x_{\text{max}} - x_{\text{min}} + 1} \) is the mean of the observed frequencies.

The frequencies predicted by the continuous model, i.e. the values \(y(x)\) obtained from (13) for the optimal parameter values, are shown in the last column of Table 1. A fit is considered very good if \(R^2 > 0.9\) (Altmann 1997), so the fit of the continuous model can be considered as very good.

As an alternative to the continuous model above where \(C\) is assumed to be a free parameter, we could consider \(C\) as the normalizing constant. In this case the curve (13) becomes the density of the beta distribution on the interval \([n, M]\), in which \(C\) is expressed in terms of the other four parameters as

\[
C = \frac{\Gamma(a + b + 2)}{\Gamma(a + 1) \Gamma(b + 1)(M - m)^{a + b + 1}},
\]

where \(\Gamma\) denotes the gamma function. This density with the four parameters \(m, M, a\) and \(b\) could be fitted to the observed data (where the optimal values of \(a\) and \(b\) in the last line of Table 1 could be used as starting values for the iterative fitting). However, we will not pursue this idea in the present article.

2. New data record: Russian

The focus of the analysis to be carried out is the morph length in Russian. For this purpose we study the Russian novel “Kak zakaljalas’ stal’ (KZS)” by N. Ostrovskij; see Kelih (2009a, 2009b). For comparison purposes we also examine the Slovenian translation of this text, i.e. we study the morph length in a South Slavic and in an East Slavic language. In each case only the first chapter is considered. The investigation steps are:

1. These texts are subjected to a tagging procedure, in which chapter headings, abbreviations, digits etc. are processed by a common principle (cf. Antić, Kelih and Grzybek (2006)), i.e. omitted.
2. The word form is determined according to orthographic criteria (Kelih (2007)), i.e. each sequence which is separated by a space is regarded as one word form. The hyphen is considered to have a delimiting function.

3. In the created lists of word form types the number of morphs per word form is determined manually. Further details of the segmentation are given in the language-specific analyses below.

4. Zero-morphemes are not considered in the segmentation. The length of a morpheme is measured in the number of phonemes.

5. Only word forms with a frequency greater than one are analyzed.

6. The morph length is determined at the level of word form types and word form tokens. Thus the behaviour of the morph length can be modelled and compared both on the paradigmatic and syntactic level.

The morphological segmentation causes a series of theoretical problems, since – as with other linguistic units – a whole range of different definitions of morph or morpheme is provided. Slovenian as well as Russian are both, typologically, highly inflectional languages, so the morphological segmentation can be done by means of the same analytical procedure.

For the present analysis a pragmatic approach has been chosen. The segmentation is based on Russian morphological dictionaries, which provide in-depth morphological information about Russian word forms. For Russian the (exemplary) morphological dictionary by Kuznecova and Efremova (1986), which provides a detailed morphological segmentation of more than 52,000 Russian lexemes (for details see Kempgen (1999)), was used. As another reference source for Russian we used Tichonov (2002). To give at least one example of the performed segmentation: the Russian verb form *vyneslas’* (3.P. f. Sg. Past Tense, reflexive) is segmented into \{vy\} (prefix) – \{nes\} (root) – \{l\} (suffix, marking past tense) – \{a\} (suffix marking feminine) –\{s’\} (postfix, marking of reflexivity), resulting in one morph with three phonemes, three morphs with one phoneme and one morph with two phonemes.

The fitting of both models to the Russian data is illustrated in Table 2 for the data at the level of word form types and in Table 3 for word form tokens. One empirical fact of the analyzed texts is worth mentioning: in both cases the most frequent morph has the length of one phoneme, which can presumably be explained by the fact that Russian is a highly inflectional language, where just one phoneme can be the carrier of different grammatical information. A rather similar picture can be seen in the Slovenian data.

<table>
<thead>
<tr>
<th>(x)</th>
<th>(f_x)</th>
<th>discrete, (N_{P_x})</th>
<th>continuous, (y(x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>623</td>
<td>623.00</td>
<td>614.73</td>
</tr>
<tr>
<td>2</td>
<td>223</td>
<td>239.92</td>
<td>281.42</td>
</tr>
<tr>
<td>3</td>
<td>242</td>
<td>218.26</td>
<td>160.04</td>
</tr>
<tr>
<td>4</td>
<td>95</td>
<td>115.82</td>
<td>96.73</td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>39.51</td>
<td>58.32</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>8.99</td>
<td>33.24</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1.36</td>
<td>16.43</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.14</td>
<td>5.51</td>
</tr>
</tbody>
</table>

EPB 1-d. \(n = 9, p = 0.1853, \alpha = 0.5004, C = 0.0033\)

Beta f. \(M = 0, M = 9, C = 32.6042, a = -0.8552, b = 1.4123, R^2 = 0.9649\)

---

This approach results from the fact that elsewhere a separate analysis of the Hapax Legomena has been made (Kelih 2011) and the same word-form lists are used here.
Table 3
Russian morph lengths (Tokens)

<table>
<thead>
<tr>
<th>x</th>
<th>f_x</th>
<th>discrete, NP_x</th>
<th>continuous, y(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2518</td>
<td>2518.00</td>
<td>2498.78</td>
</tr>
<tr>
<td>2</td>
<td>1200</td>
<td>1190.86</td>
<td>1331.57</td>
</tr>
<tr>
<td>3</td>
<td>986</td>
<td>907.98</td>
<td>743.09</td>
</tr>
<tr>
<td>4</td>
<td>290</td>
<td>423.07</td>
<td>399.66</td>
</tr>
<tr>
<td>5</td>
<td>168</td>
<td>134.41</td>
<td>196.05</td>
</tr>
<tr>
<td>6</td>
<td>38</td>
<td>30.74</td>
<td>81.32</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>5.21</td>
<td>24.42</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>0.73</td>
<td>3.28</td>
</tr>
</tbody>
</table>

EPB 1-d. n = 13, p = 0.1127, \( \alpha = 0.5168, \) C = 0.0117
Beta f. \( m = 0, M = 9, C = 6.9700, a = -0.3632, b = 2.8286, R^2 = 0.9832 \)

All fittings of this section can be considered good or satisfactory. A first indication is given that the EPB distribution and the used beta function are suitable to model Russian morph lengths.

3. New data record: Slovenian

Unlike the Russian, for the segmentation of Slovenian word forms there are not so many reference books (morphological and word formation dictionaries) available. Nevertheless the morphological segmentation was performed in analogy to the Russian analysis.

The prefixes, stems and suffixes have been identified step by step (as described in Toporišič (2000: 149) and SSKJ (1970ff), which were used as the main resources for the determination of the morphs). Other word formation issues were resolved with the help of Stramljič-Breznik’s (2004) analysis. The “depths” of the morpheme identification are identical to the segmentation of the Russian data. Tables 4 and 5 show the corresponding fittings for the two Slovenian data records.

Table 4
Slovenian morph lengths: word form types

<table>
<thead>
<tr>
<th>x</th>
<th>f_x</th>
<th>discrete, NP_x</th>
<th>continuous, y(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>453</td>
<td>453.00</td>
<td>440.86</td>
</tr>
<tr>
<td>2</td>
<td>258</td>
<td>276.86</td>
<td>313.96</td>
</tr>
<tr>
<td>3</td>
<td>284</td>
<td>251.10</td>
<td>216.51</td>
</tr>
<tr>
<td>4</td>
<td>136</td>
<td>147.08</td>
<td>139.46</td>
</tr>
<tr>
<td>5</td>
<td>64</td>
<td>62.53</td>
<td>79.95</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>20.56</td>
<td>36.85</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>6.89</td>
<td>9.91</td>
</tr>
</tbody>
</table>

EPB 1-d. n = 33, p = 0.0536, \( \alpha = 0.6281, \) C = 0.0054
Beta f. \( m = 0, M = 8, C = 11.4039, a = -0.0720, b = 1.8782, R^2 = 0.9486 \)
Table 5
Slovenian morph lengths: word form tokens

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f_x$</th>
<th>discrete, $NP_x$</th>
<th>continuous, $y(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1963</td>
<td>1963.00</td>
<td>1970.92</td>
</tr>
<tr>
<td>2</td>
<td>2170</td>
<td>2050.24</td>
<td>2145.73</td>
</tr>
<tr>
<td>3</td>
<td>1178</td>
<td>1284.23</td>
<td>1213.25</td>
</tr>
<tr>
<td>4</td>
<td>424</td>
<td>509.47</td>
<td>430.79</td>
</tr>
<tr>
<td>5</td>
<td>188</td>
<td>143.60</td>
<td>89.26</td>
</tr>
<tr>
<td>6</td>
<td>62</td>
<td>30.58</td>
<td>7.84</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>5.88</td>
<td>0.10</td>
</tr>
</tbody>
</table>

EPB 1-d. $n = 21$, $p = 0.0589$, $\alpha = 0.6721$, $C = 0.0098$

Beta f. $m = 0$, $M = 8$, $C = 0.0041$, $a = 1.6186$, $b = 6.7270$, $R^2 = 0.9971$

Both considered models have proved to be suitable also for Slovenian data.

4. Concluding remarks

In the present study we consider a discrete model for the morph length distribution for Spanish, Slovenian and Russian data, which requires three parameters. It should be mentioned at this point that, considering a language individually, more suitable discrete models than the EPB come into play. However, the EPB is the only distribution (out of the stock of about 200 discrete distributions available with the Altmann-Fitter (1997)) which could be fitted to all five data records. Since the data sets in Tables 1–5 have very small classes for large lengths, we have pooled the classes so that the minimum class size is 50.

Choosing a continuous function, we restricted ourselves to the beta function. The use of this continuous function is satisfactory for all analyzed languages. In any case, it could be shown that for the newly analyzed data (Slovenian, Russian) the same probabilistic model can be employed, both for the type as well as the token level. It appears that in addition to the 1-shifted Hyperpoisson distribution (cf. Best (2005a: 258)) also the EPB distribution and the Bbeta function can be considered as good models for the distribution of the morph length. Since only a small number of languages and texts have been investigated to date, further studies of other languages are indispensable in future. Furthermore, the impact of the used (different) morph segmentation should be analyzed in detail.

References


Tuzzi, Arjuna; Popescu, Ioan-Iovitz; Zörnig Peter; Altmann Gabriel (2012): Aspects of the behaviour of parts-of-speech in Italian texts (In this volume).


Software