Annotation and Personality: Individual Differences in Sentence Boundary Detection

Anton Stepikhov[1] and Anastassia Loukina[2]

[1] Department of Russian, St. Petersburg State University,
11 Universitetskaya emb., 199034 St. Petersburg, Russia
a.stepikhov@spbu.ru
[2] Educational Testing Service,
660 Rosedale Rd, MS 11-R, Princeton, NJ 08541, USA
aloukina@ets.org

Abstract. The paper investigates the relationships between personality traits and expert manual annotation of unscripted speech. The analysis is based on the results of a psycholinguistic experiment in which the participants were asked to annotate sentence boundaries in transcriptions of Russian spoken monologues. Personality traits were measured using the Eysenck Personality Inventory and the Five Factor Personality Questionnaire. A multiple regression model showed that the traits measured by the Five Factor Personality Questionnaire along the scales 'unemotionality vs. emotionality' and 'practicality vs. playfulness' had significant effect on the average sentence length.

Keywords: sentence boundary detection, spontaneous speech, manual annotation, Russian, personality, five factor personality model, big five, Eysenck personality inventory.

1 Introduction

Expert manual annotation is usually considered to be the gold standard for sentence boundary detection in unscripted speech. These boundaries are crucial for natural language processing since a range of automatic procedures are based on this information. Nonetheless the extent of inter-annotator agreement may vary significantly [1,2,3]. For example, we earlier showed considerable disagreement in the number of boundaries and average length of sentences marked by expert annotators in Russian spontaneous speech [3]. In the present paper we explore the role of individual differences between annotators in the inter-annotator agreement.

Previous work sometimes sought to establish strict definition of a unit of segmentation as a way to increase inter-annotator agreement (cf. [4]). On the contrary, our basic assumption in this study is that sentence boundaries are inherently ambiguous and vary between speakers. This is illustrated by Russian written texts which allow substantial variability in placement of sentence boundaries. While several syntactically independent sentences may be combined into one by a semicolon or just a comma, main and subordinate clauses of the same sentence may be segmented as separate sentences. In some cases, a sentence boundary may even be placed between two syntactic

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 105–112, 2014.

constituents governed by the same head. Therefore sentence length in a written text is significantly determined by the author's individuality. We suggest that similar variability is also likely in spoken monologues and especially in expert annotations of sentence boundaries in speech since the latter can be seen as a transformation of an oral text into a written one.

Previous studies have looked into how the inter-annotator agreement may be affected by the choice of material. For example, when studying the effect of social characteristics of speaker and type of text we found higher inter-annotator agreement for texts produced by female speakers than for texts produced by male speakers and also higher agreement for more structured texts such as story telling than for descriptive texts [3].

In this paper we investigate whether the preference for different sentence length observed in previous studies may be related to the presence of different personality traits such as extraversion or neuroticism. We use two personality questionnaires to evaluate such traits: the Eysenck Personality Inventory (EPI) [5] adopted and validated for Russian by [6] and the Five Factor Personality Questionnaire, or the Big Five (FFPQ) [7] adopted and validated for Russian by [8].

We use linear regression to evaluate whether variation in sentence lengths between the annotators can partially be explained by differences in personality as measured by two personality questionnaires. The results of the analysis may be of a special interest for oral speech researchers, psycholinguists and those involved in development and adjustment of automatic sentence boundary detection models.

2 Experimental Design and Data Collection

2.1 Data

The study is based on expert manual annotation of spontaneous monologues. We used three texts taken from the corpus of transcribed spontaneous Russian monologues described in [2]. This corpus contains manual transcriptions of different types of monologues recorded by 32 native speakers of Russian. Each speaker was presented with several tasks: (1) to read a story with a plot and subsequently retell it from memory ('story'), (2) to read a descriptive narrative without a plot and retell it from memory ('description'), (3) to describe a series of pictures in a cartoon ('picture story'), (4) to describe a landscape painting ('picture description') , and finally (5) to comment on one of the suggested topics ('free comment'). All speakers were well acquainted with the person making the recording and used natural conversational style. All recordings were done in Russia.

For this study we selected 3 monologues from this corpus produced by the same male speaker. This speaker had a higher education and was 40 years old at the time of the recording. Since expert manual annotation depends on text genre [3] we included the monologues which covered three different tasks: "Description" (162 words), "Story" (225 words) and "Free comment" (312 words).

2.2 Participants

Fifty native speakers of Russian (9 male and 41 female) took part in the experiment. All participants were students or professors in linguistics and/or modern languages with

a background in linguistics. The age of participants varied between 18 and 68 with a median age of 24. Forty nine participants were right-handed.

2.3 Personality Questionnaires

The participants were first asked to complete two personality questionnaires.

Eysenck Personality Inventory (EPI) consists of 57 yes/no questions and the results are interpreted along two scales: introversion vs. extraversion and stability vs. neuroticism. Each scale ranges from 0 to 24. There is also a separate lie-scale designed to identify participants who are being insincere and exclude them from the data.

Five Factor Personality Questionnaire (FFPQ) includes 75 items with five-level Likert scale (from -2 to 2 including 0). Each item has two opposite statements, and a respondent has to choose the closest score on the scale to one or another statement. The results of FFPQ are interpreted along five scales corresponding to five super-trait factors to describe personality: 1) introversion vs. extraversion, 2) separateness vs. attachment, 3) naturality vs. controlling, 4) unemotionality vs. emotionality, and 5) practicality vs. playfulness.[1] Each scale ranges from 15 to 75. Both questionnaires were administered on paper.

2.4 Sentence Boundary Annotation

After completing the questionnaires, the participants were given orthographic transcriptions of the 3 recordings described in Section 2.1 and asked to mark the sentence boundaries using conventional full stops or any other symbol of their choice (e.g. a slash). The participants did not have access to actual recordings and were asked to provide annotations based on text only. In addition, the transcriptions did not contain any punctuation or any other information that could indicate prosodic information such as graphic symbols of hesitation or filled pauses (like *eh, uhm*) or other comments (e.g. *[sigh], [laughter]*). Thus, we tried to focus on semantic and syntactic factors in boundary detection (see [2] for further discussion about the relative role of prosodic and semantic factors for this task). The experts were presumed to have a native intuition of what a sentence is and, thus, it was left undefined. There were no time-constraints.

In addition to the three main texts, the participants were also asked to annotate sentence boundaries in a control text included to make sure that the participants understood the task correctly. For this control task we selected a short written story (371 words) which had relatively simple syntax and almost unambiguous sentence boundaries. This text was processed in the same way as other monologues to remove punctuation and capitalisation and presented along with other texts.

3 Data Analysis and Results

We first computed scores for each scale of the two personality inventories giving us 7 personality scores per each participant. We also computed an average sentence length

[1] We follow [7] for factor names since this version of FFPQ was used as the basis for the Russian version.

in each text for every annotator. We then used multiple linear regression analysis to evaluate the effect of personality traits on average sentence length.

3.1 Inter-annotator Agreement in Control Text and Main Corpus

We first compared the inter-annotator agreement in the control text and spoken monologues. We used the method suggested in [2] to compute boundary confidence scores (BCS) for each boundary. This score reflects how many annotators agreed on the presence of a boundary at a particular location. Then the number of BCS with acceptable (60-100%) or low (less than 60%) agreement for each text was calculated.

We found that the number of positions with acceptable agreement was 2.5 times higher in the control text (60.4% of all boundaries marked by experts) than in spoken monologues (24.7% of all boundaries). Since this result was what we had expected, we concluded that all participants understood the task correctly and followed the same pattern while performing the annotation.

3.2 Descriptive Analysis of the Main Corpus

Table 1 shows the descriptive statistics for all personality traits as well as sentence length across different texts.

Table 1. Summary statistics of the acquired data ($N = 50$)

Personality scores					
Inventory	Scale	Mean	SD	Min	Max
EPI	introversion vs. extraversion	12.86	4.46	5.00	21.00
EPI	stability vs. neuroticism	14.74	4.22	5.00	22.00
FFPQ	introversion vs. extraversion	49.00	9.01	27.00	70.00
FFPQ	separateness vs. attachment	52.50	10.06	25.00	67.00
FFPQ	naturality vs. controlling	52.28	9.70	27.00	74.00
FFPQ	unemotionality vs. emotionality	53.22	10.41	22.00	70.00
FFPQ	practicality vs. playfulness	56.82	7.80	39.00	74.00
Average sentence length in different texts					
	Description	15.96	2.88	10.80	23.14
	Story	15.12	3.27	9.78	25.00
	Free comment	15.69	4.19	6.93	26.00
	Control text	14.20	2.08	10.91	21.82

The values of both scales of EPI (introversion vs. extraversion scale and stability vs. neuroticism scale) were highly correlated with the values of two corresponding scales of FFPQ – introversion vs. extraversion scale (Pearson's $r = 0.78$, $p < 0.01$[2]) and unemotionality vs. emotionality scale (Pearson's $r = 0.7$, $p < 0.01$). This shows that these two scales measure the same traits within each of the inventories. There were no correlations significant at $\alpha = 0.01$ between other traits in these questionnaires.

[2] All p–values in this section are adjusted for multiple comparison using Bonferroni correction.

Average sentence length across speakers varied between 7 and 26 words with an average length of about 15 words. We also observed correlations between average sentence length across different texts for the same annotator (r varied between 0.44 and 0.62, $p < 0.01$).

3.3 Multiple Regression Models

We first used multiple linear regression to model whether variability in sentence length can be explained by two personality traits measured by EPI. We used the sum of the mean values of sentence length in description, story and free comment as a response variable. This approach allowed us to estimate the effect of personality traits across texts of different types.

Upon building and examining this model, we found that the fit was unduly affected by two influential cases (Cook's $D > 4/50$). We removed these from the data. All further analyses were performed using only the data from the remaining 48 annotators.

The model based on EPI scales did not show any statistically significant association between personality traits and annotation results: multiple $R^2 = 0.006$, adjusted $R^2 = -0.04$, $F_{(2,45)} = 0.15$, $p = 0.87$ (see also Table 2).

Table 2. Raw (B) and standardised (β) coefficients, and p-values of EPI scales in a multiple regression model ($N = 48$)

Independent variable	*B*	Standardised β	*p*-value
introversion vs. extraversion	0.14	0.08	0.60
stability vs. neuroticism	0.02	0.01	0.95

In contrast, multiple linear regression built using five FFPQ scales' scores as independent variables showed significant effect of personality traits on sentence length (multiple $R^2 = 0.24$, adjusted $R^2 = 0.15$, $F_{(5,42)} = 2.61$, $p = 0.04$). Estimated β and p-values for the independent variables are given in Table 3.

Table 3. Raw (B) and standardised (β) coefficients, and p-values of FFPQ scales in a multiple regression model ($N = 48$)

Independent variable	*B*	Standardised β	*p*-value
introversion vs. extraversion	0.14	-0.03	0.84
separateness vs. attachment	0.13	0.15	0.30
naturality vs. controlling	0.00	0.02	0.92
unemotionality vs. emotionality	-0.28	-0.39	0.01*
practicality vs. playfulness	0.42	0.46	0.01*

We then modified our model excluding three insignificant variables from it. After transformation the new model achieved multiple $R^2 = 0.21$, adjusted $R^2 = 0.18$, $F_{(2,45)} = 6.05$ and $p = 0.005$ (see also Table 4). Analysis of deviance of the first and

Table 4. Raw (B) and standardised (β) coefficients, and p-values of two FFPQ scales in the modified multiple regression model ($N = 48$)

Independent variable	*B*	Standardised β	*p*-value
unemotionality vs. emotionality	-0.28	-0.39	0.01
practicality vs. playfulness	0.44	0.45	0.003

the modified models revealed that goodness of fit did not change after transformation ($p = 0.7$).

Finally, we fitted the same models to the control written text. None of the models achieved statistical significance. The model based on 'unemotionality vs. emotionality' and 'practicality vs. playfulness', which was the best performing model on spoken monologues, showed the following performance on the control text: multiple $R^2 = 0.10$, adjusted $R^2 = 0.065$, $F_{(2,45)} = 0.26$ and $p = 0.08$.

4 Discussion and Conclusions

We have examined the relationship between expert manual annotation of sentence boundaries in unscripted speech and expert personality. We found that, in agreement with previously reported results [2], the annotators differed in average sentence length. At the same time, there was a significant correlation between mean length of annotated sentences for each annotator across different texts. This suggests that there may be a tendency for each annotator to prefer sentences of a particular length which remains constant across different texts.

We modelled average sentence size in the annotations as a linear function of personality traits of annotators computed using two different personality questionnaires. Multiple regression analysis showed that about 18% of variability in sentence length can be explained by two personality traits which are described by the scales 'unemotionality vs. emotionality' and 'practicality vs. playfulness' of the Big Five. Of these two 'practicality vs. playfulness' had somewhat stronger effect than 'unemotionality vs. emotionality'.

The regression model showed that all other things being equal less emotional people tend to divide oral text into longer sentences than more emotional ones. At the same time, people who are more open to new experiences prefer longer sentences than more practical ones.

Since the absolute values of standardised regression coefficients for these scales are very close (Table 4), their mutual effect, when both scales have the same values, is close to zero; i.e. very emotional and open people annotate texts in the same way as unemotional and practical ones do. And, vice versa, the greater the difference between the scores of these scales the stronger the effect of the factor with higher scores. This would imply that very emotional but practical people divide texts into shorter sentences, and unemotional but open ones into longer utterances.

One possible explanation for the observed personality effect may be that it affects the annotator's attitude towards the task which in turn results in a difference in sentence

length. For example, personality may affect the diligence with which the annotator approached the task. In this case we would expect a similar effect of personality traits on sentence length in all texts. However, our analysis showed that personality scores had no effect on sentence length in the control text with almost unambiguous syntactic boundaries. This suggests that characteristics such as attentiveness to the task which may come along with the annotator's personality, do not determine choices of boundary placement alone, but that there are more fundamental interactions between personality and segmentation in silent reading.

The use of questionnaires to measure personality has a number of drawbacks. Firstly, there is a probability of response bias due to respondent's wish to "erect a favorable social façade" [9] or possible deliberate faking. An informant may as well be unwittingly defensive or may not be sufficiently self-observant. Secondly, the five-factor model is not all-encompassing – there is a level of psychological understanding that simply cannot be reached and revealed by this approach [9]. Thus, there may be other individual traits which also influence the segmentation, e.g. divergent semantic sensibilities of annotators about the meaning of the text as a whole or their way of thought. However, these limitations do not affect the main result of our study: the variability between the annotators can partially be explained by their individual characteristics such as their responses to a personality questionnaire.

Other individual characteristics that may affect the annotation are the verbal and/or visual memory of the annotator. For example [10] showed that higher working memory capacity leads to larger implicit prosodic chunks in silent reading. This fact may be extrapolated to segmentation of a text into sentences in the process of expert manual labelling. Therefore in the future we plan to explore the relations between annotators' memory abilities and segmentation.

Acknowledgments. The paper has benefited greatly from the valuable comments of Dr. Walker Trimble. We also thank all the participants of the experiments. Finally, we would like to acknowledge Dr. Keelan Evanini, Dr. Su-Youn Yoon and the two anonymous reviewers for their comments and feedback.

References

1. Liu, Y., Chawla, V.N., Harper, M.P., Shriberg, E., Stolcke, A.: A study in machine learning from imbalanced data for sentence boundary detection in speech. Computer Speech and Language 20(4), 468–494 (2006)
2. Stepikhov, A.: Resolving Ambiguities in Sentence Boundary Detection in Russian Spontaneous Speech. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 426–433. Springer, Heidelberg (2013)
3. Stepikhov, A.: Analysis of expert manual annotation of the russian spontaneous monologue: Evidence from sentence boundary detection. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS (LNAI), vol. 8113, pp. 33–40. Springer, Heidelberg (2013)
4. Foster, P., Tonkyn, A., Wigglesworth, G.: Measuring spoken language: A unit for all reasons. Applied Linguistics 21(3), 354–375 (2000)
5. Eysenck, H.J., Eysenck, S.B.G.: Manual of the Eysenck Personality Inventory. University of London Press, London (1964)

6. Shmelev, A.G.: Test-oprosnik Ajzenka. In: Bodalev, A.A., Karpinskaya, et al. (eds.) Praktikum po Psikhodiagnostike. Psikhodiagnosticheskie Materialy, pp. 11–16. MGU, Moscow (1988) (in Russ.)
7. Tsuji, H., Fujishima, Y., Tsuji, H., Natsuno, Y., Mukoyama, Y., Yamada, N., Morita, Y., Hata, K.: Five-factor model of personality: Concept, structure, and measurement of personality traits. Japanese Psychological Review 40(2), 239–259 (1997)
8. Khromov, A.B.: P'atifactornyj oprosnik lichnosti: Uchebno-metodicheskoe posobie. Izd-vo Kurganskogo gosudarstvennogo universiteta, Kurgan (2000) (in Russ.)
9. Block, J.: The Five-Factor Framing of Personality and Beyond: Some Ruminations. Psychological Inquiry 21(1), 2–25 (2010)
10. Swets, B., Desmet, T., Hambrick, D.Z., Ferreira, F.: The role of working memory in syntactic ambiguity resolution: A psychometric approach. Journal of Experimental Psychology: General 136(1), 64–81 (2007)