Journal of Advanced Research in Social Sciences and Humanities
Volume 4, Issue 2 (57-65)
DOI: https://dx.doi.org/10.26500/JARSSH-04-2019-0203

Analysis of formal characteristics of text in the CPACT Research: Enhancing the LIWC linguistic processing for the Czech language

DALIBOR KUCERA[*], **JIRI HAVIGER**

[1] University of South Bohemia in Ceske Budejovice, Czech Republic, EU
[2] University of Hradec Kralove, Czech Republic, EU

Abstract

Aim: This paper describes how psycholinguistic and psychodiagnostic fields have adopted quantitative text analysis to process spoken Czech. This method employs computer-assisted linguistic procedures to categorize and quantify formal characteristics (such as morphology, semantics, etc.) of recorded texts.
Method: The study's sample size is 200 people who were selected using age, gender, and level of education to reflect the same proportion of representation of the target groups as is found in the total Czech population. The processes of lemmatization (the identification of a lexical unit as a dictionary entry) and unambiguity (the removal of ambiguity in interpreting a particular word or homonymy) are used in formal text analysis.
Findings: In total, CPACT studies use 212 linguistic variables, which is a substantial number. So the output is much larger than the Linguistic Processes module in the LIWC 2015 program, which processes 29 grammatical/summary variables. The linguistic variables processed by LIWC are limited, but the grammatical categories and subcategories used in the CPACT study allow for a much more in-depth exploratory study.
Implications/Novel Contribution: The results of this study provide new information on the experimental application of quantitative psycholinguistic analysis to formal parameters. It's a fascinating strategy, and it yields many interesting hypotheses and study directions. Research into this area, whether by linguists or psychologists, has the potential to reveal surprising new insights into the makeup and dynamics of human communication.

Keywords: Psycholinguistics, Psycho Diagnostics, Computational Linguistics, Text, Personality, CPACT

Received: 11 February 2019 / **Accepted:** 7 March 2019 / **Published:** 22 April 2019

INTRODUCTION

The most common label for the current school of thought that focuses on the psychological aspects of texts is "psycholinguistics," a science that bridges the gap between psychology and linguistics. The study of language encompasses a wide variety of issues, including but not limited to: speech production and reception; mental structures; the correlation between linguistic ability and performance; the mental representation of linguistic constructs; the interface between language and cognition; the process of learning a language; the acquisition of that language; the relationship between language and cognitive processes; and sometimes even the biological basis of language (Nebeská, 1992; Pradhan, 2016). The study of texts has received a lot of attention, but there have been few concrete findings on the psychological aspects of verbal communication. The causes of these problems are fairly obvious. People who use language and verbal communication have a wide range of preferences regarding tone, vocabulary, and syntax. Because of this, many researchers in the field of psycholinguistics have shifted their attention away from studying the text as a marker of personality and social processes and toward studying language and its connections to brain activity instead (J. W. Pennebaker & Graybeal, 2001).

Given the scope of the problem, selecting an appropriate research strategy is of utmost importance. Researchers often turn to the content analysis method to get an objective and systematic description of overt communication content (Berelson, 1952; Hilao, 2016). Quantitative or qualitative methods can be used to evaluate the

[*]Corresponding author: Dalibor Kucera
[†]Email: dkucera@pf.jcu.cz

content (Ferjenčík, 2000). The focus of content analysis is on the larger context in which a text was created, the communicators' motivations and explicit and implicit communication goals, the text's content, its formal parameters (using which the communication goal is externalized), and the effect the texts have on the recipient. In addition to establishing analytical categories, it is important to isolate lexical units (typically a word, collocation, or sentence) that serve as indicators of semantic units. Analytical classes are based on how often a given lexical unit occurs in the text (Nebeská, 1992). Analytical methods can vary widely depending on the specific discipline. It is possible to divide methods into qualitative and quantitative categories, as well as conceptual and relational ones (describing the existence and frequency of a particular unit vs describing the relationship among the occurrences of a unit, producing the so-called mental models) (Carley, 1993), and representationally-aimed and instrumentally-aimed methods (creating a representation of the sender's original intention of a message vs analyzing a message for occurrences of a set of keywords (extracts information on the conversational meaning of a theme) and what is communicated (the content) as opposed to how it is communicated (the style) (Boonyarattanasoontorn, 2017; Eid & Diener, 2006).

LITERATURE REVIEW

In the internet age, various new opportunities for studying linguistic phenomena have showed up. In this context, Quantitative Text Analysis (QTA) provides a researcher with broad possibilities for descriptions and statistical processing of text samples (J. W. Pennebaker & Stone, 2003). Quantitative text analysis is any systematic reduction of a flow of text (or other symbols) to a standard set of statistically manipulable symbols representing the presence, the intensity, or the frequency of some characteristics relevant to social science (Shapiro & Markoff, 1997). As a scientific method, text analysis was first used during World War II, to analyze the content of Nazi propaganda (Krippendorff, 2018). Since the 1960s, several notable methods in the field of QTA were used. They differed in many aspects and strategies, e.g., coding (judges or computerized word count strategies) and the linguistic parameters they examined. Current approaches to computational linguistic analysis, suitable for psychological use, can be divided into two basic groups closed approaches (i.e., closed vocabulary analysis based on counting the frequency of predefined words that are contained in a corpus (Park et al., 2015) and open approaches (i.e., open vocabulary analysis uncategorized data-driven extraction of linguistic phenomena, such as words, phrases, punctuation, emoticons, or themes; (Schwartz et al., 2013). The application of gained output is manifold - from tracking of consumer behaviour, political opinions, personal preferences, changes in society, to the personalist or psychological applications.

Chung and Pennebaker (2007) many substantial results in application of QTA on psychological topics. His goal was to understand how the words people use in their daily interactions reflect who they are and what they do. Using a special text-mining application, Linguistic Inquiry and Word Count (LIWC); (J. W. Pennebaker, Boyd, Jordan, & Blackburn, 2015), he focused on both content and style characteristics. He used LIWC analysis to process enormous amount of text samples and to compare this data with miscellaneous personal (e.g., social, health, psychological) characteristics of their writers. After several experiments he found impressive relation between personal data and Formal (non-semantic) Parameters of Text (FPT), such as grammatical categories and quantity of certain words - particularly function words, so called particles or junk words (i.e. pronouns, prepositions, articles, conjunctions, and auxiliary verbs) (Chung & Pennebaker, 2007). These words are not directly related to the meaning of the text, compared with content words (such as nouns and regular verbs, which are content heavy), but they have a more social and psychological meaning. Studies focused on verbal communication of people affected by damage to Broca's area show significant changes of their expression of nouns and regular verbs, but not of function words. Damage to Wernicke's area, as distinct from previous, causes an increase in the use of a high number of function words, but decreases the amount of content words. Even at the brain level, then, function words are linked to social skills (Miller, 1995).

Text and Personality Characteristics

The idea that specific wording is to some extent linked to the personality of the communicator, has appeared in professional literature for many decades (Sanford, 1942; Scherer & Giles, 1979). Other psychologists have followed-up on this observation (Robinson & Giles, 1990; Weintraub, 1989). Comparably to non-verbal forms of

social behavior, verbal forms, such as speech acts, are a means of achieving goals, and therefore comply with the definition of a psychological trait (Cheng, 2011). Thus, several researches focused on a link between word use and certain personality characteristics.

RESULTS AND DISCUSSION

Results of this studies relate to many psychological variables, e.g., extraversion, which is associated with more frequent use of the words positively emotionally colored e.g., "happy, wonderful, amazing" (J. W. Pennebaker & King, 1999). This is also confirmed by Schwartz et al. (2013) or Yarkoni (2010) research, which documented the use of longer words and acronyms (Holtgraves, 2011) and lack of words expressing distinctions e.g., "Besides, in contrast, etc."; (J. W. Pennebaker & King, 1999). Another relevant characteristic is neuroticism communicators with higher scores use more often nominative singular (e.g., "I, my, me") and fewer number of positively emotional words (ibid). Another trait, openness to experience, is characterized by a higher incidence of higher frequency of citations and references to social processes (Sumner, Byers, & Shearing, 2011). Outside the dimension of the Big Five model traits there has been described many other personality characteristics, e.g., social desirability, which is related to avoidance of appropriation (e.g., a lower number of expressions "my, your") (Knapp, Hart, & Dennis, 1974), distress and 1st person words (J. Pennebaker, 2003), or social and interpersonal orientation, which is characterized by a lower number of pronouns in the 1st person and a higher number of identifying pronouns session (e.g., "anyone who") for people who are involved on social interaction (Cegala, 1989).

Current Research in Czech Language

Any research of psychological aspects of language use is, of course, highly dependent on the target language. English language undoubtedly dominates the field, judging by the number of researches and research studies, as well as speakers. Nevertheless, it is believed that focusing psycholinguistic research solely on this language, as the world language and lingua franca, is not an ideal approach due to fact that the morphological, lexical, and stylistic structure of other languages may be rather different and/or may display features not pre-sent in the English language. The Czech language, a member of the West-Slavic language group, differs from English in a number of aspects, such as in terms of inflection (Czech is highly inflected while English only weakly inflected, e.g., affixes cumulate grammatical functions), lexicology (e.g., lack of diminutives in English as opposed to its abundance in Czech), or syntax (e.g., fixed word order in English) (Hornova, 2003). The given over-view of characteristics clearly proves that difference between the two languages are significant, albeit on the stylistic, syntactical, or morphological level, and supports the authors conjecture that Czech texts are more variable and can thus provide more information on the author and their characteristics (Kucera, Hemmerová, & Haviger, 2016).

Drawing on the research results of relevant foreign studies, the CPACT research has been designed. Computational Psycholinguistic Analysis of Czech Text (CPACT) is a three-year research project devoted to the study of verbal communication. The CPACT project is carried out at the University of South Bohemia from 2016 and is funded by a grant from the Czech Science Foundation (GACR, grant no. 16-19087S). The project team connects experts from five academic institutions, including Czech Academy of Sciences, Charles University, Masaryk University, the University of Hradec Králové and the University of South Bohemia. The research aims to understand the relations between a person's personality and the words they use, as already mentioned in the previous section following foreign researches (Kucera, 2017). The research employs both psychological testing methods and computational linguistics methods. We work with these two sets of data and by means of explorative statistic methods we seek particularly a significant correspondence among the tested items and the individual textual parameters. Subsequently, all results are interpreted in detail and compared with relevant foreign researches to detect any possible similarities or contrasts. The research is also meant to initiate a follow-up survey working with substantially lower number of variables (meaning with shortlisted tests and a shortlisted group of textual parameters, see below), yet with higher accuracy rate due to the parameters that seem to be promising for other studies. The project is currently in the data processing phase psychological and psycholinguistic outputs will be therefore published in follow-up studies.

METHODOLOGY

Methods of Text Analysis in the CPACT Research

In the paper, we deal with technical aspects of computational linguistic analysis. The formal text analysis, i.e. closed vocabulary analysis, depends on the appropriate definition of the morphological (lexical) category of the individual unit (e.g., word, punctuation or emoticon). For this purpose, the process of lemmatisation (i.e. identification of lexical unit as a dictionary entry) and unambiguity (i.e., elimination of ambiguity in interpretation of the particular word or in other words homonymy) is employed. The outcome of this process is the allocation of morphological signs (tags) to every lexical unit of the text (Petkevič, 2006). For examples of automatic language tagging process see Figure 1.

"Pojďte s námi do ZOO."

Pojďte (jít Vi-P---2--A---1) **s** (s RR--7----------) **námi** (my PP-P7--1-------)
do (do RR--2----------) **ZOO** (ZOO NNFS2-----A----) . (. Z:----------)

Example 1: Analysis transcription

No.	Category (Subtag Name)	Description in English	Description in Czech
1	POS	Part of Speech	Slovní druh
2	SUBPOS	Detailed Part of Speech	Slovní poddruh
3	GENDER	Agreement Gender	Rod
4	NUMBER	Agreement Number	Číslo
5	CASE	Case	Pád
6	POSSGENDER	Possessor's Gender	Rod vlastníka
7	POSSNUMBER	Possessor's Number	Číslo vlastníka
8	PERSON	Person	Osoba
9	TENSE	Tense	Čas
10	GRADE	Degree of Comparison	Stupeň
11	NEGATION	Negation (by prefix)	Negace
12	VOICE	Voice	Slovesný rod
13	UNUSED	Reserved for future use	Volná pozice
14	UNUSED	Reserved for future use	Volná pozice
15	VAR	Variant, Style, Register	Varianta, styl

Figure 1. Description of values and categories

POS	Detailed part-of-speech used (SUBPOS)
p (pronoun)	0 1 4 5 6 7 8 9 D E H J K L P Q S W Y Z

POS & SUBPOS	possible form(s)	lit. translation (description)
PO	naň	on-him (compound with -n)
PI	jehož	whose (in relative clause)
P4	jaký	what
P4	který	which
P5	něj	him (he, after prep, only)
P6	sebe	himself (long form)
P7	se, si	refl. pronouns
P8	svůj	his (poss. refl. pronoun)
P9	něhož	who, in rel. clause, after prep.
PD	tento	this (demonstrative)
PE	což	which (in rel. clause)
PH	mě	me (pers. pron. clitic)
PJ	jenž	who, in rel. clause
PK	kdo	who (rel./interrogative)
PL	všechen	all
PP	ty	you (personal)
PQ	co	what (rel./interrogative)
PS	můj	my (possessive)
PW	nic	nothing (negative)
PY	oč	about-what (compound with -c)
PZ	nějaký	some
PZ	něco	something

Figure 2. Tagging categorise of pronouns

Results of the automatized analyses provide data in four different categories: Definition of the general morphological tags, Definition of the lexical category of the word, Numbers of specific features and configurations and Emotional load of the word (Sentiment analysis) (see Table 1).

Table 1: Linguistic categories in the CPACT analysis

Category	Description
Definition of the general morphological tags	Part of speech (noun, adjective, pronoun, number, verb, adverb, preposition, conjunction, particle, interjection), detailed part of speech, agreement gender, agreement number, case, possessor's gender, possessor's number, person, tense, degree of comparison, negation (by prefix), voice, variant
Definition of the lexical category of the word	Diminutives, vulgarisms, words of common Czech (colloquial words), phrasemes
Numbers of specific features and configurations	Number of words, sentences and complex sentences, number of words and punctuation marks, number of words and number of sentences, number of different lemmas (basic forms) in relation to the number of words, number of finite verbs in relation to the number of sentences (sentence complexity), number of punctuation marks in relation to the number of sentences, number of exclamation marks in relation to the number of sentences, adjective-noun sequence ratio, number of sentences starting with a conjunction, number of vulgar words in relation to the number of sentences, number of colloquial words in relation to the number of sentences, the presence of emotionally charged words
Emotional load of the word (Sentiment analysis)	Application of two dictionaries observing positive, negative and neutral emotional characteristics of the words (Veselovská, Hajic, & Sindlerová, 2014)

The total number of linguistic variables in the CPACT research reaches relatively high values - 212 variables in total. The output is therefore significantly larger than the comparable Linguistic Processes module in the LIWC 2015 program (J. W. Pennebaker et al., 2015) where 29 grammatical/summary variables is processed (83 variables in total, see Table 2). While LIWC processes only some linguistic variables, the CPACT research is working with a significantly larger range of grammatical categories and subcategories and this range allows much more extensive exploratory research. On the other hand, the CPACT research doesn't operate with further semantic/lexical information analysis (except sentiment analysis), while the LIWC processes this information in many categories.

Table 2: Linguistic categories in the LIWC analysis-linguistic processes module

Category	Description
Summary language variables	Word count, analytical thinking, clout, authentic, emotional tone, words/sentence, words > 6 letters, dictionary words
Linguistic dimensions	Total function words, total pronouns, personal pronouns, 1st pers singular, 1st pers plural, 2nd person, 3rd pers singular, 3rd pers plural, impersonal pronouns, articles, prepositions, auxiliary verbs, common adverbs, conjunctions, negations
Other grammar	Common verbs, common adjectives, comparisons, interrogatives, numbers, quantifiers
Psychological processes (Lexical categories)	Affective processes, social processes, cognitive processes, perceptual processes, biological processes, drives, time orientations, relativity, personal concerns, informal language

Research Sample, Texts Sources and Psychological Data

The research sample comprises 200 assessed subjects selected according to the criteria of age (15-24, 25-34, 35-55 and 55 + years), gender (male and female) and education (elementary school, high school and university), i.e. by means of the proportionate stratified sampling, with regards to the identical percentage of representation of

the target groups that is in the whole Czech population (pursuant to the accessible data from the Czech Statistical Office from 2014).

The assessed subjects are supposed to create textual resources intended for linguistic analysis pursuant to the criteria defined beforehand. These are four types of below-mentioned texts with overall length of 180-200 words. The text is written on the computer (in a pre-defined electronic interface) on the same day. The factual content of the text with the particular message might be entirely fabricated. It just has to follow the orientation of so-called scenarios: a Cover letter (TXT1), a Letter from Vacation (TXT2), a Complaint (TXT3) and a Letter of Apology (TXT4).

- Cover Letter (TXT1): "You have found a job offer that captivated your interest and you really aspire to be hired for this particular position. Therefore, you are going to write a letter to the company's director as a response to his/her offer trying to persuade the director that it is you who is the right candidate for this position."
- Letter from Vacation (TXT2): "You are enjoying your time on amazing vacation. Everything is going well as expected and you fully indulge in your popular activities. Therefore, you have decided to write a letter to your friend and convince him/her to come over and enjoy such perfect time with you."
- Complaint (TXT3): "Until recently you were living with satisfaction in your apartment (or your house), not missing any single thing. Nevertheless, recently issues that made a hell out of a pleasant living appeared. Although you originally strived to sort out the issues in a gentle way, it did not help. Therefore, you decided to write an official complaint to the respective authorities."
- Letter of Apology (TXT4): "You have done something that substantially harmed your relationship with a person you were very close to for a long time. You promised something you did not fulfil. You feel sorry and you know that you made a mistake. Since you do not want to lose such person, you have decided to write a letter of apology to him/her." Sequence of the texts is selected randomly during a day. As other two resources of verbal data we have selected two rewritten semi-structured personal (oral) interviews 5-15 minutes long that follow a certain pre-defined scenario as well. Sequence of the interviews is selected always randomly during a day. It is a Job Interview (TXT5) and Narration about a Pleasant Experience (TXT6).
- Job Interview (TXT5): "In a separate room there is a director of a company behind a desk (an elderly authoritative man), a video camera is placed on a tripod in front of the assessed subject. There is strong studio lighting and the director conducts a job interview with the assessed subject not in a very friendly and communicative way)."
- Narration about a Pleasant Experience (TXT6): "In a separate room (different from the previous one) there is a nice elderly lady offering refreshment and cheerfully welcoming the assessed subject. She asks him/her to narrate a story about a nice experience he/she can recall. She does not intervene the narration, just supports a relaxed communication and motivates the subject."

In order to identify personal and social characteristics of the authors of the text, a set of psychological tests is used and assessed subjects are supposed to sit for these tests during one day in a particular computer room. The tests were selected due to their diagnostical targeting to the three following areas: (1) pathologic characteristics and states (in particular anxiety, depression and tension/stress), (2) general personality dimensions (personality traits) and motivational tendencies and (3) interpersonal and social skills. The following tests have been employed: Big Five Inventory (BFI), Persönlichkeits- Stil- und Störungs-Inventar (PSSI), State-Trait Anxiety Inventory (STAIX2), Multi-Motive Grid (MMG); Depression Anxiety Stress Scales(DASS21), Interpersonal Adjective Scales revised (IAS-R), Sense of Humor Questionnaire (SHQ), Social Phobia Safety Behaviors Scale (SPSBS); Self-Monitoring Scale (SMS) and Basic Olomouc Body Rating (BOBR).

We used two variants of tests, namely self-report variant (self-description of the author of the texts) and other-report variant (the author is described by a related person). The reason for using the other-report variant is primarily bias, which influences the results of self-report questionnaires (Vazire, 2010); e.g., results can be affected by self-serving bias, misinterpretation, specifics and limitations of introspection, (Furnham, 1986). Persons who describe the author of the text (within the other-report test variant) declared a mutual close relationship and above-standard familiarity, so we can consider the information obtained from this variant of the test as relevant.

Discussion

As mentioned, the project is currently in the data processing phase. Descriptive statistics focused on the texts and questionnaires as well as their mutual relation are therefore conducted with all the data. Regarding the description of the texts, the essential information is what verbal means the assessed subjects use, how they are related to their age, gender and education and particularly how the texts differentiate from each other. We seek particularly the information related to the presumption of different vocabulary of various texts and their different grammatical structure. Subsequently, all results are supposed to be interpreted in detail and laid into correspondence with comparable foreign researches in order to detect any possible concord or contrasts. The research is also supposed to initiate another consequent survey working with substantially lower number of variables (meaning with shortlisted tests and a shortlisted group of textual parameters), yet with higher accuracy rate due to the parameters that seem to be perspective for other studies.

CONCLUSION, RECOMMENDATIONS AND IMPLICATIONS

As has been noted earlier, the possibility of psychological assessment in terms of diagnostics of interpersonal and intrapersonal characteristics through computational text analysis proves to be an interesting and useful tool not only for psychologists, but also for educational specialists and professionals and researchers in other areas. The major benefits are not only its unobtrusive character (because it does not require the presence of the of the person examined nor the need for other diagnostic tools) but also an evident possibility of a deeper understanding of the author's identity, his personality characteristics, opinions and attitudes. If a database were created to show which textual characteristics relate to the particular personal trait of a writer, it would be possible to fully process a huge amount of text entirely automatically. Such a mechanism would also allow the researcher to predict the likelihood of certain personality traits of a given writer. In this text, we introduced the possibility of extensive linguistic analysis, which is available in the Czech language. At present, it is significantly outstripping the possibilities of processing English using the LIWC program. After the explorative phase of the CPACT project, it will be possible to focus more precisely on those characteristics that are important from a psychological perspective and to make a subsequent reduction in the number of linguistic variables (formal parameters of text) for the analysis.

To conclude, experimental use of quantitative psycholinguistic analysis of formal parameters is a truly interesting method which brings about many promising ideas and research suggestions. Should it be explored by researchers, be it linguists or psychologists, it could generate new, potentially unexpected information on human communication, its nature and characteristics.

ACKNOWLEDGMENT

The research Computational Psycholinguistic Analysis of Czech Text and this study is funded by a grant from the Czech Science Foundation (GA R, grant no. 16-19087S).

REFERENCES

Berelson, B. (1952). *Content analysis in communication research.* New Jersy, NJ: Free press.

Boonyarattanasoontorn, P. (2017). An investigation of Thai students English language writing difficulties and their use of writing strategies. *Journal of Advanced Research in Social Sciences and Humanities*, *2*(2), 111-118. doi:https://doi.org/10.26500/jarssh-02-2017-0205

Carley, K. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology*, 75-126. doi:https://doi.org/10.2307/271007

Cegala, D. J. (1989). A study of selected linguistic components of involvement in interaction. *Western Journal of Communication*, *53*(3), 311-326. doi:https://doi.org/10.1080/10570318909374309

Cheng, K. H. (2011). Further linguistic markers of personality: The way we say things matters. *International Journal of Psychological Studies*, *3*(1), 2-10. doi:https://doi.org/10.5539/ijps.v3n1p2

Chung, C., & Pennebaker, J. (2007). *Social communication: Frontiers of social psychology: The psychological functions of function words*. New York, NY: Psychology Press.

Eid, M. E., & Diener, E. E. (2006). *Handbook of multimethod measurement in psychology.* New York, NY: American Psychological Association.

Ferjenčík, J. (2000). *Introduction to the methodology of psychological research in research: How to examine the human soul*. Novato, CA: Portál.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, *7*(3), 385-400. doi:https://doi.org/10.1016/0191-8869(86)90014-0

Hilao, M. P. (2016). Creative teaching as perceived by English language teachers in private universities. *Journal of Advances in Humanities and Social Sciences*, *2*(5), 278-286. doi:https://doi.org/10.20474/jahss-2.5.4

Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality*, *45*(1), 92-99. doi:https://doi.org/10.1016/j.jrp.2010.11.015

Hornova, L. (2003). *Reference dictionary of grammatical terms*. Olomouc, Czechia: Palacky University Olomouc Publisher.

Knapp, M. L., Hart, R. P., & Dennis, H. S. (1974). An exploration of deception as a communication construct. *Human Communication Research*, *1*(1), 15-29. doi:https://doi.org/10.1111/j.1468-2958.1974.tb00250.x

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. New York, NY: Sage publications.

Kucera, D. (2017). Computational psycholinguistic analysis of Czech text and the CPACT research. In *4th International Multidisciplinary Scientific Conference on Social Sciences and Arts,* Albena, Bulgaria.

Kucera, D., Hemmerová, E., & Haviger, J. (2016). Quantitative psycholinguistic analysis of formal parameters of Czech text. In *Proceedings of International Scientific Council of SGEM,* Sofia, Bulgaria.

Miller, G. (1995). *The science of words*. New York, NY: Library.

Nebeská, I. (1992). *Introduction to psycholinguistics*. New York, NY: H&H.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934-952. doi:https://doi.org/10.1037/pspp0000020

Pennebaker, J. (2003). The social, linguistic, and health consequences of emotional disclosure. In, Suls, J., (Ed.), *Social psychological foundations of health and illness*. Malden, MA: Blackwell Publication.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc 2015* (Technical report). University of Texas, Austin, TX.

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, *10*(3), 90-93. doi:https://doi.org/10.1111/1467-8721.00123

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296-1300. doi:https://doi.org/10.1037//0022-3514.77.6.1296

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, *85*(2), 291-300. doi:https://doi.org/10.1037/0022-3514.85.2.291

Petkevič, V. (2006). *Reliable morphological disambiguation of czech: Rule-based approach is necessary*. Slovakia, Bratislava: Slovak Academy of Sciences.

Pradhan, S. (2016). English language teaching: A next gate to social awareness. *International Journal of Humanities, Arts and Social Sciences*, *2*(4), 156-158. doi:https://doi.org/10.20469/ijhss.2.20005-4

Robinson, W. P., & Giles, H. (1990). *Handbook of language and social psychology*. New York, NY: Wiley.

Sanford, F. H. (1942). Speech and personality. *Psychological Bulletin*, *39*(10), 811-845.

Scherer, K. R., & Giles, H. (1979). *Social markers in speech.* Cambridge, UK: Cambridge University Press.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... others (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, *8*(9), 737-791. doi:https://doi.org/10.1371/journal.pone.0073791

Shapiro, G., & Markoff, J. (1997). *A matter of definition: Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah, NJ: Erlbaum.

Sumner, C., Byers, A., & Shearing, M. (2011). Determining personality traits & privacy concerns from facebook activity. *Black Hat Briefings*, *11*(7), 197-221.

Vazire, S. (2010). Who knows what about a person? The Self Other Knowledge Asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281-300.

Veselovská, K., Hajic, J., & Sindlerová, J. (2014). Subjectivity lexicon for Czech: Implementation and improvements. *Journal for Language Technology and Computational Linguistics*, *29*(1), 47-61.

Weintraub, W. (1989). *Verbal behavior in everyday life*. New York, NY: Springer Publishing Co.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363-373. doi:https://doi.org/10.1016/j.jrp.2010.04.001