frontiers in Psychology

HYPOTHESIS AND THEORY
published: 04 March 2022
doi: 10.3389/fpsyg.2022.819543

# Beyond English: Considering Language and Culture in Psychological Text Analysis

*Dalibor Kučera*[1]* *and Matthias R. Mehl*[2]

[1] Department of Psychology, Faculty of Education, University of South Bohemia in České Budějovice, České Budějovice, Czechia, [2] Department of Psychology, College of Science, University of Arizona, Tucson, AZ, United States

The paper discusses the role of language and culture in the context of quantitative text analysis in psychological research. It reviews current automatic text analysis methods and approaches from the perspective of the unique challenges that can arise when going beyond the default English language. Special attention is paid to closed-vocabulary approaches and related methods (and Linguistic Inquiry and Word Count in particular), both from the perspective of cross-cultural research where the analytic process inherently consists of comparing phenomena across cultures and languages and the perspective of generalizability beyond the language and the cultural focus of the original investigation. We highlight the need for a more universal and flexible theoretical and methodological grounding of current research, which includes the linguistic, cultural, and situational specifics of communication, and we provide suggestions for procedures that can be implemented in future studies and facilitate psychological text analysis across languages and cultures.

**Keywords: natural language processing, cross-language, culture, closed-vocabulary approaches, LIWC**

OPEN ACCESS

***Edited by:***
*Markus Kemmelmeier, University of Nevada, Reno, United States*

***Reviewed by:***
*Marianna Marcella Bolognesi, University of Bologna, Italy*
*Peter Boot, Huygens Institute for the History of the Netherlands (KNAW), Netherlands*

***Correspondence:**
*Dalibor Kučera*
*dkucera@pf.jcu.cz*

***Specialty section:***
*This article was submitted to Cultural Psychology, a section of the journal Frontiers in Psychology*

***Received:*** *21 November 2021*
***Accepted:*** *14 February 2022*
***Published:*** *04 March 2022*

***Citation:***
*Kučera D and Mehl MR (2022) Beyond English: Considering Language and Culture in Psychological Text Analysis. Front. Psychol. 13:819543. doi: 10.3389/fpsyg.2022.819543*

## INTRODUCTION

The use of computerized text analysis as a method for obtaining information about psychological processes is usually dated to the 1960s, when the General Inquirer program was introduced (Stone et al., 1962). Since then, this field has advanced and flourished in ways that were difficult to foresee at the time. The original (word-count) approaches have been enhanced and optimized in terms of the scope and complexity of their dictionaries and methods (Eichstaedt et al., 2020), and the capacity of computers has arrived at processing very large amounts of data in no time. At the same time, extensive digital documentation and sharing, related to the growth of the information society (Duff, 2000; Fuller, 2005), have provided almost unlimited input for text analysis.

Over the last decade, Natural Language Processing (NLP) methods have effectively become an established and attractive go-to method for psychological science (Althoff et al., 2016; Pradhan et al., 2020). At present, they are developed mainly as automated systems that can understand and process texts in natural language, e.g., for conversational agents, sentiment analysis, or machine translation (Amini et al., 2019). The new techniques, employing methods of artificial intelligence, classical machine learning (ML), and deep learning methods (Magnini et al., 2020) are gradually displacing original approaches, with their eventual dominance in the field being a safe prediction (Johannßen and Biemann, 2018; Eichstaedt et al., 2020; Goldberg et al., 2020).

By implication, the field can currently be thought of as being in a transitional phase—although most cited studies in psychology are based on foundations laid with conventional computational techniques (e.g., word counting), their share is gradually decreasing in favor of more complex techniques (e.g., ML processing). This phase is crucial in many ways, not only for the (re)evaluation of existing research backgrounds and evidence but also for the development and optimization of next-generation psychological text analysis methods.

The goal of this article is to provide a critical review of the approaches, methodology, and interpretation of traditional closed-vocabulary text analysis from the specific perspective of multicultural and multilingual research. Attention is paid to three fundamental challenges: (1) the specifics of language and culture, (2) the levels of language analysis in question and the terminology used, and (3) the context of the use of specific tools and methods. The article ends with a discussion of possible adjustments and extensions to methods and outlines further perspectives and desiderata for conducting cross-language research in psychology.

## CHALLENGES IN CROSS-LANGUAGE PSYCHOLOGICAL TEXT ANALYSIS

Over the last two decades, research on psychological aspects of natural word use (Pennebaker et al., 2003; Ramírez-Esparza et al., 2008; Harley, 2013) has provided an impressive bedrock of scientific findings. Most of this research has been carried out using closed-vocabulary approaches, methods based on assigning words within a target text document to categories of a predefined word dictionary (Eichstaedt et al., 2020). Semantic and grammatical features of word use have been identified as psychological markers of personal speaker characteristics, for example, gender and age (Biber, 1991; Mehl and Pennebaker, 2003; Newman et al., 2008), personality characteristics (Tausczik and Pennebaker, 2010; Yarkoni, 2010; Gill and Oberlander, 2019), social characteristics (Berry et al., 1997; Avolio and Gardner, 2005; Dino et al., 2009; Kacewicz et al., 2014), emotions (Brewer and Gardner, 1996; Pennebaker and Lay, 2002; Newman et al., 2008), and health (Ramírez-Esparza et al., 2008; Demjén, 2014). The research has so far mostly been conducted within an explanation framework, but is now also increasingly used for prediction purposes (Yarkoni and Westfall, 2017; Johannßen and Biemann, 2018).

The large number of existing studies speaks to the high relevance of this research, both in terms of establishing consensus between studies and in revealing relationships with other variables as support for concurrent validity with the results of established measures. However, recent studies have also raised important questions about the generalizability of existing findings beyond the original context of investigation, which has highlighted potential constraints on their validity in different languages and cultures (Garimella et al., 2016; Basnight-Brown and Altarriba, 2018; Jackson et al., 2019; Sánchez-Rada and Iglesias, 2019; Chen et al., 2020; Thompson et al., 2020; Dudău and Sava, 2021). The results of the studies also indicate that the comparison and psychological interpretation of linguistic phenomena between different cultures and languages is subject to several fundamental challenges.

### Language and Culture in Question

The first challenge concerns the choice of the language and culture in which the texts are analyzed and interpreted. Currently, the vast majority of psychological language research is based on English, which dominates contemporary science as a lingua franca (Meneghini and Packer, 2007; Seidlhofer, 2011). The preference of research in English is understandable—English is a global language (e.g., the most used language of international communication, information technology, and on the Internet) (Internet Users by Language, 2021), English is the consensual language of academic discourse and, as such, it has a broad research base (Johnson, 2009). Nevertheless, the number of English native speakers (approx. 360–400 million) (König and van der Auwera, 2002), is a small fraction of the world's population. There are approximately 6,900 languages spoken today, of which 347 have more than 1 million speakers (Bender, 2011).

Although it may seem that languages are rather similar to each other, in many cases they exhibit substantial phonological, morphosyntactic, and semantic structural differences. In other words, they operate with different linguistic building blocks, structures, and relations to communicate equivalent ideas (Haspelmath, 2020). As an example, we can describe the variance that exists in even such a basic classification as content (lexical) vs. function (grammatical) words (Corver and van Riemsdijk, 2001). Although most languages allow a relatively clear distinction between these two types, this is not the default for all languages (Asher and van de Cruys, 2018). For example, in indigenous North American languages, the words "sit," "stand," and "lie," considered content words in English, appear as both content and function words (Hieber, 2020). Moreover, many word classes (parts of speech) are not present in some languages (e.g., adjectives are not present in Galela language) (Rijkhoff, 2011). Such differences exist at all levels of language (i.e., language domains, parts of grammar) and further examples will be given below.

In addition to differences between individual languages, differences between cultures using the same language should also be mentioned. As an example, we can use English, which is currently the official language in at least 58 countries (List of Countries Where English Is an Official Language – GLOBED, 2019). Not surprisingly, the use of English shows a number of variations across these cultures. The variations are most often manifested at the level of pragmatics (e.g., accentuated manifestations of egalitarianism in western Anglophone cultures compared to more pronounced patterns of respect in Asian and Polynesian Anglophone cultures) (Thomas and Thomas, 1994), but also at the level of semantics—in understanding the meaning of words (e.g., the word "old" is usually more semantically related to "age" in Australian English and to the "past" in American English) (Garimella et al., 2016). Other aspects also contribute to language variation, such as dialects or the specific use of English by non-native speakers (Wolfram and Friday, 1997; Yano, 2006). Considering that languages show

such variability at both intra-lingual and inter-lingual levels, and function differently in many aspects, this may raise the question of the adequacy of single-language results (or single-culture results) that are often implicitly assumed to be broadly applicable (Wierzbicka, 2013).

## Definition of Levels and Variables of Language Analysis

The second challenge consists of the definition of the level of language (language domain, area of linguistic analysis) we focus on, the terminology used, and the variables in question. In research on the psychology of word use, terminology is often not set in accordance with traditional taxonomy in linguistics and does not adequately reflect interlingual differences. Instead of distinguishing language levels (domains) in dimensions which are more universal and established among linguists, e.g., morphology, syntax, semantics, lexicology, etc. (Hickey, n.d.; Mereu, 1999; Kornfilt, 2020), the focus of the analysis is often described in eclectic ways, based on the specifics of the language in question. For example, English is a language that has a relatively poor morphology compared to other Indo-European languages (Vannest et al., 2002; Milizia, 2020), and the level of morphology is therefore often integrated into a group of diverse variables or is replaced by other concepts. A common example is the sorting of language features into fuzzy categories such as "Linguistic Dimensions" (covering word classes and morphology), "Other Grammar" (covering word classes and both morphology and syntax), and "Psychological Processes" (covering semantics, morphology, syntax, and pragmatics together) in the LIWC2015 program (Pennebaker et al., 2015) (note: this method is described in more detail below). In fact, each of these categories includes strictly linguistic dimensions (variables), only in different configurations.

Another example is the differentiation between 'language content' (content of communication, that is, what is communicated/told, that usually covers lexical and semantic level) and 'language style' (the way the content is conveyed, that is, how the author is communicating, theoretically covering all levels of analysis, including morphology) (Ireland and Pennebaker, 2010). The assumption that language content and style can be unambiguously distinguished at the level of individual variables is questionable, since the definition of words as "content" (e.g., nouns, verbs) or "stylistic" (e.g., pronouns and prepositions) varies considerably between languages (Corver and van Riemsdijk, 2001; Asher and van de Cruys, 2018; Hieber, 2020). Even the most general distinction between function words and content words in one language captures rather a continuum, where prototypical function words and content words appear at opposite ends (Osborne and Gerdes, 2019). In summary, although these conceptual or effectively metaphorical distinctions have proven theoretically generative and practically useful, they can significantly limit the possibilities of cross-language comparison.

The unclear taxonomy and exclusive, domain-specific terminological definition bring with them complications both at the level of interdisciplinary cooperation (e.g., among psychologists and linguists) and at the level of intercultural research (Sonneveld and Loening, 1993). For languages that are relatively close in their structure, the discrepancy in classification may not be pronounced, but when distant languages are studied and compared, substantial differences can arise. The taxonomy of words and their functions is non-trivially language-specific, with different languages providing different classifications of language content and style (Nivre et al., 2016; Kirov et al., 2020). In some languages, the same grammatical relationship is expressed morphologically, in others through function words, while some languages do not mark this information at all (e.g., in grammatical tense or definiteness) (Osborne and Gerdes, 2019; Universal Dependencies: Syntax, 2021).

For example, many locatives are marked by prepositions in English (e.g., "in," "by," "to," "from"), while in Finnish they appear as morphological case-inflections (e.g., "-ssA," "-llA," "-lle," "-stA," "-ltA"). Furthermore, possessives and adverbials can be marked morphologically in Finnish (e.g., "-ni"—"my," "-si"—"your"), but in English they appear as separate words, thus a word form like "auto-i/ssa/ni/kin" ("also in my cars") with stem and four subsequent suffixes would need four separate words in English (Vannest et al., 2002). The Czech language provides another example of the interconnection between language content and style. It also works with a wide range of grammatical suffixes that change paradigmatic and grammatical classification, e.g., the word "uč" ("teach!") with suffixes "-it" ("to teach") "-el" ("teacher") "-ova/á" ("of teacher") "-ní" ("teaching") "-čko" ("little teaching"), where each of the suffixes can changes the inflection and/or semantic nature of the word (Rusínová, 2020). Therefore, a text analysis approach that counts and processes such linguistic units as stand-alone words (Pennebaker and King, 1999; Pennebaker et al., 2014) is inherently limited and potentially biased.

## Approaches and Methods in Question

The third challenge concerns specifics around the commonly employed text analytic approaches and methods. Many methods were primarily designed for the processing of a specific language, or even a specific type of communication (i.e., genre or register), and their use in cross-language research can therefore result in methodological and interpretive difficulties. In this regard, the current approaches to quantitative text analysis, based on lexical and semantic levels of analysis (treating words/tokens as lexical units within a certain semantic field) (Cruse et al., 1986), can be divided into two main groups—closed-vocabulary approaches and open-vocabulary approaches (Schwartz et al., 2013b). Closed-vocabulary approaches operate from "top down" and assign words from a target text to psychologic categories within a specific and fixed dictionary (e.g., a dictionary of emotional words that covers categories of positive and negative emotion categories). This procedure is also referred to as the word-count approach (Schwartz et al., 2013a; Iliev et al., 2015; Kennedy et al., 2021). The result of the analysis is usually the (normalized) frequency within which references to these categories occur in a given text (Eichstaedt et al., 2020).

Compared to that, open-vocabulary approaches operate from "bottom-up" (data-driven), that is, based on language (text) as

such. Algorithms identify related clusters of units (lexical units or elements, for example, punctuation) that naturally occur (and co-occur) within a large set of texts and find lexical and semantic patterns that appear (and appear together) in the data (Park et al., 2015; McAuliffe et al., 2020). Both approaches have their pros and cons; as stated by Eichstaedt et al., "Closed-vocabulary approaches can be rigid, while open-vocabulary approaches can be sensitive to idiosyncrasies of the dataset and the modeler's choices about parameters. Closed-vocabulary approaches are more reproducible but inflexible, where open approaches are more flexible but can vary across datasets" (p. 77) (Eichstaedt et al., 2020). Given the historical dominance of word-count approaches, the following section focuses in detail on closed-vocabulary analysis.

## CLOSED-VOCABULARY APPROACHES IN CROSS-CULTURAL RESEARCH

In terms of the number of published studies, closed-vocabulary approaches still dominate by far the field of psychology of word use. There are many reasons for their preference, for example, their implementation exacts little technical demands (training of the AI, development of algorithms, etc.), they allow relatively uncomplicated interpretation of the results, and they also do not require large datasets to perform the analysis (Eichstaedt et al., 2020; Sharir et al., 2020). Over the last six decades, a number of tools have been developed, e.g., General Inquirer (Stone et al., 1962), DICTION (Hart, 2001), Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), Affective Norms for English Words (ANEW) (Bradley and Lang, 1999), SentiStrength (Thelwall et al., 2010), SentiWordNet (Baccianella et al., 2010), OpinionFinder (Wilson et al., 2005), Regressive Imagery Dictionary (Martindale, 1973), TAS/C (Mergenthaler and Bucci, 1999), Gottschalk-Gleser Scales (Gottschalk et al., 1969), or Psychiatric Content Analysis and Diagnosis (PCAD) (Gottschalk, 2000).

Most of these methods are primarily focused on the level of lexical semantics, that is, on searching for words with specific semantic loading. The analyzed text is usually compared with a predefined dictionary that contains words that represent a concept (e.g., religion words) or a psychological state (e.g., positive emotion words). For example, the concept of 'satisfaction' in DICTION is represented by words such as "cheerful," "smile," or "celebrating" (Hart and Carroll, 2011). Leaving aside the question of the validity of the semantic categories in the dictionary itself (cf. Garten et al., 2018), there are several issues that closed vocabulary analysis has to deal with. A common problem is the interpretation of lexical ambiguity and the meaning of words in different contexts (Hogenraad, 2018). A typical example in English are contronyms or polysemous words such as "fine" (signifying both pleasant and a penalty), "mean" (signifying both bad and average), and "crazy" (signifying both excitement and mental illness). The risk of misinterpretation (misclassification) can be reduced by, e.g., removing or replacing ambiguous words from the dictionary (Schwartz et al., 2013a). However, such a procedure almost necessarily also reduces the sensitivity of the semantic category, and thus the precision of the analysis.

### Level of Lexical Semantics in Cross-Language Adaptation

If we focus on cross-language adaptation of closed-vocabulary methods, it should be emphasized that these tools are naturally based on the specifics of the source (original) language for which they were developed, most often English [see Mehl (2006)]. Therefore, adapting such dictionaries to other languages is often a complicated and time-consuming process that faces a series of additional challenges (Bjekić et al., 2014; Dudău and Sava, 2020; Boot, 2021). First, the methods are most often based on the original cultural and linguistic structure rather than the target culture or language, that is on the imposed-etic approach (Berry et al., 1997). This strategy can lead, among others, to the risk of reductionism or misinterpretation of results, for example, when constructs (variables/categories) do not exist, are not equivalent, or function differently in the original and target language (Church and Katigbak, 1989). Languages often have unique words that are difficult to express in other languages (e.g., words like "toska" in Russian, "jamani" in Swahili or "saudade" in Portuguese). Furthermore, even for words that seem easy to translate, their meaning may shift, e.g., in English, the word "anger" is mainly related to wrath, irateness or rage, while in the Nakh-Daghestanian languages, it is closer to envy and in the Austronesian languages more closely associated with pride (Jackson et al., 2019).

Let us add that semantic changes are not a matter of cross-language comparison only, but they can also occur naturally within one language, such as in different historical stages of a language (Vanhove, 2008; Riemer, 2016; Garten et al., 2018).

Second, the possibility of estimating possible shortcomings of dictionary adaptation can be problematic, since the degree of equivalence varies not only across language features (some words are more cross-linguistically and cross-culturally comparable than others) (Biber, 2014), but also across different communication contexts (Daems et al., 2013; Biber and Conrad, 2019; Dudău and Sava, 2020). For example, the meaning and use of the English word "hump" vary both between English speaking cultures and between situational contexts (e.g., in British English it can refer to an emotional state, in American English it can refer to a vigorous effort, depending on the context in which it can be perceived as vulgar). In some languages, the influence of the context is crucial for the word interpretation and classification, such as in Czech, where sociolinguistic situation (inter-lingual variation) borders on diglossia (Bermel, 2014). Thus, we can assume that dictionaries validated only in a certain communication context (e.g., academic essays) will not be sufficiently effective in another context (e.g., informal conversations).

The topic of comparability of language variables (words, units, features) across languages is discussed in a number of studies. Although many of them have revealed a high degree of similarity in the results of cross-language analysis (Ramírez-Esparza et al., 2008; Windsor et al., 2019; Vivas et al., 2020),

there is increasing evidence pointing to significant differences in lexical and semantic functioning across more distant languages. In the study by Thompson et al. (2020), published in *Nature Human Behavior*, the authors analyzed semantic alignment (neighborhood) for 1,010 meanings in 41 languages using distributed semantic vectors derived from multilingual natural language corpora. While some words within semantic domains with a high internal structure were more closely aligned across languages—especially quantity, time, and kinship (e.g., "four," "day," and "son"), words denoting basic actions, motion, emotions and values (e.g., "blow," "move," and "praise") aligned much less closely. In terms of semantic alignment by parts of speech (word classes), the highest alignment was found in numerals, while other parts of speech were much less aligned (e.g., prepositions were the least aligned). Thus, this study critically questions the idea of widely comparable word meanings across languages, at least from a cross-cultural universalist perspective (Kim et al., 2000).

Another study, published in *Science*, examined nearly 2,500 languages to determine the degree of similarity in linguistic networks of 24 emotion terms (Jackson et al., 2019). The study also revealed a large variability in the meaning of emotion words across cultures. For example, some Austronesian languages colexifies the concepts of "pity" and "love," which may index a more positive conceptualization of "pity" compared to other languages. Another example concerns the connotation of "fear," which is more associated with "grief" and "regret" in Tai-Kadai compared to other languages. As the authors show, the similarity of emotion terms could be predicted based on the geographic proximity of the languages, their hedonic valence, and the physiological arousal they evoke. Given the central role of emotion words, and more broadly sentiment analysis, in the field of language analysis, this study has clear implications for cross-language analysis, particularly when comparing distant cultures and languages.

Finally, cultural differences in language use were also documented in a study that focused only on English. Garimella et al. (2016) described the differences between Australia and the United States based on the words they used frequently in their online writings. The results indicated that there are significant differences in the way these words are used in the two cultures, reflecting cultural idiosyncrasies in word use. For example, the adjective "human" is more related to human rights in the Australian context, but more to life and love in the United States context (Garimella et al., 2016). From our point of view, these studies provide important insights: although languages are similar in many ways and they certainly share universal bases, the degree of similarity varies depending on cultural and geographical specifics.

## THE LINGUISTIC INQUIRY AND WORD COUNT PROGRAM AS AN EXAMPLE

So far, we have focused on the analysis on the lexical semantics level only—this level is also common to all closed vocabulary approaches mentioned above. However, one of the methods, the LIWC program, is exceptional in this respect—besides traditional semantic categories (social words, emotion words, etc.), it provides an additional analysis of morphology and syntax features. Therefore, LIWC therefore serves well to illustrate the potentials and pitfalls of cross-linguistic adaptation of the closed vocabulary method in the context of multiple language levels (domains).

Linguistic inquiry and word count (Pennebaker et al., 2015) is currently the most widely used text analysis method in the social sciences. At the time of writing this article, 781 records were available on the Web of Science that contained "LIWC" or "Linguistic Inquiry and Word Count" as the topic, and more than twenty thousand records are listed on Google Scholar. In its current version, LIWC2015, the program offers an intuitive user interface and provides a simple and clear output of the results (Pennebaker et al., 2015), including a range of comparison possibilities (Chen et al., 2020). LIWC dictionaries have been translated and adapted into multiple languages, including Spanish (Ramírez-Esparza et al., 2007), French (Piolat et al., 2011), German (Wolf et al., 2008; Meier et al., 2019), Dutch (Boot et al., 2017; Van Wissen and Boot, 2017), Brazilian-Portuguese (Balage Filho et al., 2013; Carvalho et al., 2019), Chinese (Huang et al., 2012), Serbian (Bjekić et al., 2014), Italian (Agosti and Rellini, 2007), Russian (Kailer and Chung, 2007), Arabic (Hayeri, 2014), Japanese (Shibata et al., 2016), and Romanian (Dudău and Sava, 2020).

English LIWC2015 works with approximately 90 features grouped into 4 domains: "Summary Language Variables" (general text descriptors and lexical variables, including one syntactic variable "words per sentence"), "Linguistic Dimensions" (containing summary variables, word classes variables, and morphological variables, e.g., "total function words ", "articles," "1st person singular," and "negations"), "Other Grammar" (containing word classes variables, and both morphological and syntactic variables, "numbers," "comparisons," and "interrogatives"), and "Psychological Processes" (containing semantic variables and other variables, e.g., "sadness," "non-fluencies," and "causation") (Pennebaker et al., 2015). In terms of the analytic procedure, LIWC operates on relatively simple principles. LIWC uses its own dictionary to simply identify and label the corresponding words in the analyzed text—*via* word-count. Pre-processing in LIWC includes only basic segmentation and requires additional manual tagging (e.g., for specific ambiguous filler words, e.g., "well," "like," or non-fluencies, e.g., "you know"). More advanced NLP procedures, on the other hand, use pre-trained models and perform a sequence of "cleaning" processes in such tasks (e.g., Rayson, 2009; Manning et al., 2014), e.g., part of speech disambiguation and tagging, lemmatization, or parsing (Straka and Straková, 2017).

Several strategies have been used to adapt the LIWC dictionary to other languages (Boot, 2021). These include the supervised translation of the English dictionary word by word (Bjekić et al., 2014; Dudău and Sava, 2020), the use of the existing word corpora and their assignment to corresponding LIWC categories (Andrei, 2014) or as an enrichment of LIWC categories (Gao et al., 2013; Meier et al., 2019), the use of dictionaries in

closely related languages (Massó et al., 2013), the modification of the older version of the dictionary (Zijlstra et al., 2004), or adapting the original dictionary *via* machine translation (Van Wissen and Boot, 2017). The various LIWC languages differ significantly in the number of words contained in the dictionary. For example, the Romanian LIWC dictionary (Ro-LIWC2015) contains 47,825 entries compared to the English LIWC2015 dictionary with 6,549 entries (including words, word stems, and emoticons; cf. LIWC2007 contains 4,500 words, and LIWC2001 contains 2,300 words). The average proportion of words identified (labeled) by LIWC also varies considerably across the different LIWC language dictionaries, for example 87% in English (LIWC2015; cf. 82% in LIWC2007), 88% in German (DE-LIWC2015; cf. 70% in LIWC2001), 70% in Dutch, 54% in French, 66% in Spanish, 70% in Serbian, and 67% in Romanian (Bjekić et al., 2014; Dudău and Sava, 2020), speaking to the fact that the LIWC approach likely yields differential sensitivity across different languages.

## Analysis of Non-semantic Levels of Language

The translation and adaptation process faces most of the issues described above. Here, however, the analysis deals also with additional challenges, connected to level of morphology and syntax of the target languages, for example the pronoun-drop phenomenon (in some languages, users very frequently omit pronouns, particularly in their subject positions; e.g., "tengo hambre" in Spanish dropping the first-person singular pronoun "yo") (Świątek, 2012), grammatical classification (e.g., pronominal adverbs in Dutch, that combine pronouns/adverbs with prepositions—"we doken erin" which replaces "we doken in het"—"we dived into it"), grammatical restrictions (some linguistic features are restricted to particular languages, see below), with case sensitivity problems (LIWC is not case-sensitive which makes it difficult to process certain words, e.g., the German word "Sie" which, if capitalized, serves as formal second person singular or plural pronoun and, when not capitalized, serves as third person plural pronoun), and the above mentioned ambiguity (including, if the capitalized word appears at the beginning of a sentence) (Boot, 2021).

Although some shortcomings of the dictionary translation approach can be partially overcome (e.g., by removing words from the dictionary, adding new words and phrases, or with data pre-processing), they still increase the risk of reduced sensitivity and validity, especially in its reliability and comparability to the original method. As already mentioned, this applies particularly to languages with a grammatical structure more distant from English. For example, due to the grammatical structure of Serbian (a Slavic language), the category of verbs had to be substantially modified, and the category of articles had to be removed completely (Bjekić et al., 2014). Many adjustments were also made in the Romanian adaptation, for example in verb tense, grammatical gender, or diacritics processing (Dudău and Sava, 2020). To sum up, every translation of the LIWC dictionary involves many decisions about which entries (words or categories) should be kept, dropped, or added, and each decision is necessarily a trade-off between computational feasibility and linguistic accuracy (Dudău and Sava, 2021).

## Cross-Language Evaluation of Linguistic Inquiry and Word Count

The extent to which language specifics and LIWC adjustments affect the quality of adaptation is difficult to evaluate, as the studies differ in many aspects. Some studies do not report psychometric validation information for their dictionaries (e.g., Arabic, Turkish, or Russian), while others provide only indirect evidence (Balage Filho et al., 2013). In several studies, equivalence estimates are presented as a general indicator of the quality of adaptation. Equivalence is usually estimated *via* correlation coefficients between the adapted version of LIWC and the English original. If we focus on four major studies, the authors report an average correlation of adapted LIWC and English LIWC as $r = 0.67$ for German based on $N = 5{,}544/6{,}463$ texts in German/English (Europarl corpora and transcriptions of TED Talks transcriptions), $r = 0.65$ for Spanish ($N = 83$ texts in Spanish/English; various Internet sources), $r = 0.65$ for Serbian ($N = 141$ texts in Serbian/English; scientific abstracts, newspapers and movie subtitles), and $r = 0.52$ for Romanian ($N = 35$ books of contemporary literature in Romanian/English) (Ramírez-Esparza et al., 2007; Bjekić et al., 2014; Meier et al., 2019; Dudău and Sava, 2020).

Although the average values of the correlations can be considered satisfactory, upon closer inspection, they vary widely between categories and levels of analysis, especially in morphology and semantics. For example, in the Romanian LIWC, most correlations of non-semantic categories are non-significant (11 of 18 categories). Significant results were found in the category "Pronouns" in the first person (singular 0.93, plural 0.92) and in the third person singular (0.66, plural non-significant), in the category "Other Function Words" in conjunctions (0.37) and negations (0.53) and in the category "Other Grammar" in interrogatives (0.58) and quantifiers (0.66) (Dudău and Sava, 2020). Considering these results and the average proportion of total words identified in the Romanian LIWC (only 67% words were labeled), we must conclude that the Romanian LIWC appears not effective enough for the comparable analysis of non-semantic (grammatic) categories, even though its dictionary is seven times bigger than the English original (Romanian: 47,825 entries; English: 6,549 entries; Dudău and Sava, 2020).

Another issue concerns the specificity of text samples on which validity and equivalence tests were performed. In this sense, the communication context (text type, genre, register) is an important factor that can produce substantial variation both in the frequency of language features and in the associations with other variables, especially psychological ones (Pennebaker et al., 2007; Daems et al., 2013; Haider and Palmer, 2017; Biber and Conrad, 2019; Kučera et al., 2020; Dudău and Sava, 2021). Differences in the sensitivity of LIWC for detecting psychological markers in different types

of text (English only), were shown in the meta-analysis of Chen et al. (2020), in which, for example, the strength of the relationship between extraversion and positive emotion words varied significantly and substantially across communication contexts (e.g., asynchronous/synchronous and public/private communication). Thus, if only one type of communication is used (e.g., only written language), it is difficult to estimate to what extent the translated dictionary has comparable validity for, for example, spoken communication. Moreover, it is possible to assume that the language variation is related to multiple factors, not only to the type of text, but also to, for example, sociodemographic characteristics of speakers (Stuart-Smith and Timmins, 2010), as well as to discourse domain and language itself (Biber, 2014).

The above-mentioned challenges have implications not only for the adaptation of closed-vocabulary methods to other languages, but for the field of psychology of word use more broadly. Due to the predominant interest of research in the English language, psychological language markers are often implicitly presented in studies as relatively universal, generalizable at least to English-speaking cultures (Chung and Pennebaker, 2018). In many classical studies, for example, frequent use of first-person singular pronouns has emerged as a marker of negative emotionality (Pennebaker and King, 1999; Pennebaker et al., 2003; Oberlander and Gill, 2006; Gill et al., 2009; Yarkoni, 2010; Qiu et al., 2012). However, subsequent research in other languages and on other samples relativizes this relationship (Mehl et al., 2012; Bjekić et al., 2014; Holtzman et al., 2019; Kučera et al., 2020, 2021). Given the lack of cross-language and cross-cultural studies, the original assumption of generalizability is understandable. However, considering recent studies, the previous conjectures need to be corrected for regarding the culture, language, and communication contexts and samples in which the relationships emerged. If the different functioning of words in other languages and cultures is not sufficiently described, many generalizations may be biased or misrepresented as a result.

## DEALING WITH CLOSED-VOCABULARY CROSS-LANGUAGE ANALYSIS

Although the issues raised above may raise pessimism regarding the possibilities of closed-vocabulary approaches in cross-language research, we believe that most challenges can (and need to be) overcome, at least to some extent. Closed-vocabulary approaches offer, in contrast to open-vocabulary approaches, several advantages that are important for psychological research. The categories they work with can be intuitively labeled and (and facilitate interpretation, explanation, testing, and accumulation and transfer of results (e.g., into other languages and contexts) (Kennedy et al., 2021). Even if traditional methods are replaced by new technologies (e.g., AI), the demand for interpretations of phenomena based on intuitive categories (e.g., representing variables using established psychological concepts) is bound to survive. In the rest of the article, we therefore focus on suggestions that support the effective use of closed-vocabulary approaches in multilingual and multicultural setting.

### Dealing With Language and Culture

The first challenge we discussed was the language and culture on which the analysis is based and the degree of its similarity to other languages and cultures. To build on the previous arguments, text analysis methods likely provide more different results the further apart studied languages and cultures are, not only because of the methodological differences in analysis, but also because of the specifics of the languages and cultures themselves. As a parallel, we reference the issues concerning the use of Big Five personality questionnaires across cultures (the most widely used method for assessing personality characteristics), which outside of western, educated, industrialized, rich and democratic (WEIRD) populations shows serious limitations and low validity for measuring the domain of basic personality traits (Laajaj et al., 2019). In the same way, striving for better explanations of cross-linguistic variation requires employing the power of cross-cultural comparisons to describe the variation and similarity (Barrett, 2020)—the methodology must be linked to more principled sampling, both at the level of speakers (e.g., representative sample of speakers in a given culture or at least a sample corrected for imbalances) and texts (e.g., to acquire the texts with regard to their ability to be representative for selected communication contexts).

Since the cross-language comparison based on texts from the entire communication spectrum would be difficult to implement, it is necessary to choose specific types of communication (i.e., registers, and genres) to be analyzed. Leaving aside their ease of availability to the researcher, the focus should be on types of text that show a certain degree of cross-language universality. In this regard, existing cross-linguistic studies on register variation can provide important information in this regard. For example, Biber's research finds two language dimensions (i.e., constellations of linguistic features that typically co-occur in texts) that could be considered relatively (although not absolutely) universal: (1) "clausal/oral" discourse vs. "phrasal/literate" discourse, and (2) "narrative" vs. "non-narrative" discourse (Biber, 2014). The first dimension linguistically comprises typical grammatical features (e.g., verb and pronoun classes) and is based functionally on a distinction between a personal/involved focus and informational focus (e.g., private speech vs. academic writing as prototypic genres). The second, narrative dimension, consists of different sets of features (e.g., human nouns and past tense verbs), and typically appears in fictional stories, personal narratives, or folk tales. These general patterns have emerged from different studies of languages other than English, for example, Spanish, Brazilian Portuguese, Nukulaelae Tuvaluan, Korean, Somali, Taiwanese, Czech, and Dagbani (Biber, 2014).

From the point of view of cross-language comparison, it is therefore recommended to choose text types that are at least somewhat comparable on these two dimensions to ensure maximum (in the sense of as much as reasonably possible) comparability. If the selection of texts cannot be made by dimensions defined *ex ante* (e.g., if the texts have already been

collected), it is also possible to subject the texts to *ex post* dimensional analysis *via* multi-dimensional analysis (MDA), an approach that identifies co-occurrence patterns of linguistic features based on the factor analysis (Biber, 1991). Through MDA, it is possible to describe different texts in terms of their similarity in dimensional structure. However, MDA is currently only available for a limited number of languages (in addition to the languages listed above for Scottish Gaelic and written Chinese) (Sardinha and Pinto, 2019).

## Dealing With Levels of Analysis and Language Variables

The second challenge concerns the terminology and language level (domain) that is the subject of the analysis. Since the definition of language variables based on the specifics of one language only is problematic, it is necessary to work with variables that have common characteristics and to categorize them in a more clearly defined system. The issue of universal classification has been addressed in a number of studies, both theoretically and practically (Hasselgård, 2013). If we are to build on newer approaches, two of the available linguistic frameworks can serve as an example to follow, the Universal Dependencies (UD) and the Universal Morphology (UniMorph) projects (Nivre et al., 2016; McCarthy et al., 2020). Both frameworks focus on the annotation of human language and connect many fields of contemporary linguistics (Osborne and Gerdes, 2019; de Marneffe et al., 2021). In both frameworks, morphology (including part of speech) and syntax are considered the most principal non-semantic levels of language analysis in the taxonomies.

Universal Dependencies[1] is a framework for annotation of grammar across different human languages, currently available for 122 languages with 33 more in preparation (Universal Dependencies, 2021). Morphological variables of UD include, for example, the categories of part of speech and lexical and inflectional features (e.g., pronominal type and degree of comparison), and syntactic variables include cover dependency relations between words (relations between a syntactic head and a subordinate element, e.g., multiple determiners attached to the head noun).

The UniMorph project[2] has similar goals as UD and provides normalized morphological paradigms for diverse world languages, especially low-resource languages with inflectional morphology. The schema of UniMorph comprises 23 dimensions of meaning (e.g., person, number, tense, and case) and over 212 features (for the dimension of case, e.g., ablative, absolutive, accusative, etc.) (Sylak-Glassman, 2016; McCarthy et al., 2020).

If we consider Universal Dependencies and the Universal Morphology frameworks from the perspective of cross-language research, i.e., when comparing multiple languages analyses, a comment needs to be added to the number and applicability of linguistic variables. Since the set of linguistic features (categories, dimensions) we can work with is entirely dependent on properties of languages in question, it is necessary to identify features that are shared between these languages—i.e., identically labeled in UD or UniMorph. For example, if we compare the results of UD text analysis in English and Spanish, we can only work with 13 English features, which are shared with Spanish (e.g., degree, gender, person, polarity; see English ParTUT and Spanish AnCora treebanks; Universal Dependencies, 2021). However, UD in Spanish offers more linguistic features (23 features in total), and we can use these "non-English" variables, e.g., in a further comparison with another language.

To sum up, the frameworks provide useful tools, and they can serve as a starting point for better classification and (re-)definition of language variables for the purposes of cross-language psychological analyses. In addition, Universal Dependencies Tools are open-source software, so they are available for free.

## Dealing With Cross-Language Adaptation of Methods

The third challenge is related to current approaches to text analysis and their methods. In terms of cross-language use of semantically based closed-vocabulary approaches, research should focus primarily on identifying and covering the semantic specifics and functioning of words in different languages, not just on translating the text into the language of analysis. Studies that describe the semantic alignment of words across different languages and contexts could help here (Garimella et al., 2016; Jackson et al., 2019; Thompson et al., 2020). For both semantic and morphological analysis, several procedures can be used to increase the comparability of the analyses. For example, it is possible to use statistical adjustments proposed by Dudău and Sava—to employ multilevel analysis with language as the level 2 covariate (especially when text input is available in relatively different languages) or to perform within-language standardization to attenuate the language particularities that could affect the investigation in the multilingual setting.

For example, Brazilian Portuguese probably has linguistic particularities in the use of third-person singular (e.g., in personal pronouns and possessives with a higher degree of inflection), which can cause inconsistencies in cross-language comparisons (Carvalho et al., 2019). To avoid the lack of equivalence between results of analyses in different languages, it is possible to perform within-language standardization, i.e., use the mean and standard deviation of the third person singular variable as the reference parameters for rescaling the values. As the authors state, when comparing the four LIWC language adaptations (English, Dutch, Brazilian Portuguese, and Romanian), the unadjusted calculations show little sign of cross-language equivalence compared to the situation where language specificities are considered, that is, *via* within-language standardization (Dudău and Sava, 2021).

Another way to reduce the difficulties of adapting closed-vocabulary methods and subsequent cross-language comparison is to use machine translation. Two basic approaches are the "translated dictionary" approach, and the "translated text"

[1] universaldependencies.org

[2] unimorph.github.io

approach. The first one consists of automatic translation of entries (usually word by word) from the original dictionary (e.g., English) into the target language. This creates a new dictionary in the target language, which is used to perform analyses in this particular language (e.g., the Danish version of LIWC) (Boot et al., 2017; Van Wissen and Boot, 2017). The second approach consists of translating the analyzed text into the language in which the original method works (e.g., English) and then in performing the analysis with the original method. This approach seems to be effective and straightforward in many ways—it makes the analysis tool accessible to languages for which it has not yet been adapted, and reduces errors associated with the translation process and adaptation of the dictionary into another language. The efficiency of MT systems (e.g., Google Translate) is proving to be very high also in terms of syntax and stylistics and recent studies show that this "translated text" approach outperforms the traditional word-by-word "translated dictionary" approach (Windsor et al., 2019; Araújo et al., 2020; Boot, 2021), for example, in measures of equivalence of Dutch, German, and Spanish language analyses (Boot, 2021).

## Dealing With Methods Based on Machine Learning

Finally, acknowledging where the field is heading, we would like to comment on questions around new technologies in psychological text analysis more generally. The use of artificial intelligence (AI), machine learning (ML), and machine translation (MT) is already closely related to many aspects of text analysis, for example, within open-vocabulary approaches (Eichstaedt et al., 2020). Undoubtedly, modern technologies offer enormous potential based on the performance and sophistication of up-to-date computational systems, but also raise fundamental questions about methods of data processing, their supervision, and interpretation of results (Mønsted et al., 2018; Stachl et al., 2020).

The ML and MT methods allow us to expand the spectrum of observed variables and at the same time effectively predict their relationships. However, from the perspective of our paper, their disadvantage is the problematic interpretation of the analytical processes itself, i.e., the so-called black box problem (Castelvecchi, 2016). For example, it is possible to train AI on a large number of texts to effectively recognize the specific characteristics of speakers (and then, e.g., allow the AI to predict them), but it is difficult to get clearer information on what procedures and variables (features) are involved in the process (Zednik, 2019). AI is thus more of a promising method for predicting relationships, rather than a method that provides their explanation and deeper insight (Yarkoni and Westfall, 2017).

It is not within the scope of this article to discuss all aspects of ML/MT utilization; however, we would like to focus on one issue that we consider particularly important in relation to cross-language research and the use of closed-vocabulary analysis in psychology. These are the quality and complexity of the training data, especially in the context of different languages and different types of communication.

Successful use of ML depends to a large extent on the data on which the system is trained, both in terms of quantity and quality (Ehrlinger et al., 2019). Regarding the number of training texts, a general rule of thumb is that more data usually means higher effectiveness of the system (Baeza-Yates and Liaghat, 2017). In terms of data quality, the situation is much less clear. In addition to routine data quality controls (e.g., cleaning dataset from irrelevant texts), the nature of texts should also be considered, especially at the level of the type of communication that is the subject of the ML training (Smith et al., 2013; Modaresi et al., 2016; Medvedeva et al., 2017; Ott et al., 2018). For example, several studies have shown that current electronic communication is dominated by the so-called "electronic/internet discourse" (e-discourse), which takes the form of semi-speech (between speaking and writing) (Abusa'aleek, 2015). This e-discourse has its own features such as unconventional spelling and combinations of visual and textual elements (Lyddy et al., 2014; Pam, 2020).

Following this concept, we can assume that if ML is, say, trained primarily on parallel corpora of formal written communication (e.g., press releases or parliament transcripts in two or more languages), its effectiveness for processing (translating) the e-discourse or other more specific communication might be noticeably reduced, and vice versa (Koehn and Knowles, 2017; Søgaard et al., 2018). Increased error rates for certain types of text (styles, genres, registers) have been described for systems as complex as Google Translate (Putri and Havid, 2015; Afshin and Alaeddini, 2016; Prates et al., 2018). These errors mainly concern lexical/discourse errors and style errors (note: lexical errors occur when MT translates words wrongly or does not translate them, discourse errors occur when MT could not recognize the meaning of the word in its context, and style errors occur when the word is inappropriate in a given context). In the 2016 research, error rates (based on comparison with human translation) were quantified at 5.9% for lexical/discourse errors and 8% for style errors (Afshin and Alaeddini, 2016). Higher sensitivity to errors was found in the translation of function words, especially adjectives and adverbs (Putri and Havid, 2015). In addition to these errors, problems referred to as "machine-bias" can arise. A classic example is the case of gender preference in Google Translate, that is, when Google MT exhibited a strong tendency toward male defaults (Prates et al., 2018). Although the issue was quickly handled by Google through (forced) equal representation of gender categories in translation, the underlying problem itself is not resolved that easily, since MT was probably trained on (historical) data in which the male gender is more common, which resulted in the preferred in translation. In these situations, it is therefore necessary to apply methods such as "post-editing," i.e., the process of making corrections or amendments to automatically generated text (machine translation output) (Temizöz, 2016; Gutiérrez-Artacho et al., 2019).

The quality of MT is constantly changing with the ever-increasing training data and the participation of new technologies (e.g., automatic transcription of oral communication). At the same time, the accumulation of data facilitates the representation

of more diverse types of communication and language varieties (dialects, sociolects, etc.), which contributes to solving number of problems of traditional closed-vocabulary approaches (MT is based on authentic varieties of language, not on *a priori* assumptions about their functioning). However, the increase in the amount of training data is not proportional between languages—languages that are used more often in electronic communication (especially English) provide automated systems with much more data than the so-called "low-resource/resource-poor languages" (Thuy et al., 2018). Although it is possible to apply procedures that link datasets of resource-sufficient and resource-poor languages (Impana and Kallimani, 2017), the issue of reduced comparability cannot be overlooked (Seki, 2021). The described situation is a parallel to the previously mentioned problem of disproportionate representation of certain types of communication in the ML dataset. In the application of MT in psychological research, it is therefore necessary to emphasize the need for control and documentation of the ML training process, especially when working with languages that generate fewer texts compared to the world's most used languages, and when working with types of text that are more distant to original training data.

## CONCLUSION

At the beginning of our article, we stated that we are currently in a "transitional phase of research" within the field of text analysis. After more than 60 years of research on psychological aspects of word use, new technologies and methods are entering this discipline at a rapid rate. Original programs based on simple word counting are being challenged by automated machine learning systems and large-scale "big data" analyses (Gandomi and Haider, 2015) that allow for extensive cross-cultural comparisons. New technologies offer great potential, but the question is when (or whether) they will completely replace traditional techniques. It will also be important to consider to what extent the original methods can support more advanced analyses in terms of their focus, interpretation, and explanation of linguistic phenomena. In this regard, current research raises a number of questions related to the relevance of older studies, considering different language structures in different cultures and contexts of human communication (Kim et al., 2000; Jackson et al., 2019; Thompson et al., 2020).

In our critical analysis here, we focused on closed-vocabulary approaches, a relatively old method of text analysis. Nevertheless, even today, its contribution needs to be appreciated and its strengths highlighted. We would like to celebrate the ground-breaking research and many quality papers that have been published in this field over the last two decades (for all, see, e.g., Pennebaker et al., 2003). Research in Anglophone cultures has provided many excellent tools for text analysis in English, but it has also amplified universalist tendencies to adapt target languages to default methods, instead of adapting these methods to target languages and their specifics (e.g., Bjekić et al., 2014; Dudău and Sava, 2020). Given the richness and variety among different languages, many relationships between language and psychological variables are undoubtedly reduced this way (Kim et al., 2000; Wierzbicka, 2013; Kučera, 2020).

In summary, we can state three basic considerations: (1) To further the science of the psychology of word use, it is necessary to promote close interdisciplinary cooperation, especially with the fields of linguistics, computer science, and cultural psychology. Within that, linguistics can provide a clear taxonomy of language, a background in cross-linguistic research, and useful analytic tools (e.g., MDA for dimensional text description or UD for their morpho-syntactic annotation) (Biber, 2014; de Marneffe et al., 2021). (2) If we are looking for relationships between mind, behavior, and language use, it is not possible to overlook the specifics of different languages and cultures. Although studies conducted in English are usually more accessible to both researchers and the public (e.g., given the tools available and the amount of data), it is critical to compare the results with studies in other languages and cultures in order to evaluate the generalizability of relationships and to understand their meaning more deeply (Kim et al., 2000; Wierzbicka, 2013). (3) In cross-language psychological research, all present-day methods can be used. However, it is necessary to consider their functionality in different contexts (e.g., define more universal variables and comprehend situational/cultural aspects of communication) (Biber and Conrad, 2019; Cvrček et al., 2020), and critically assess their development and use. This consideration also applies to current machine learning systems, in which the possibility of methodological supervision is usually limited (in terms of control of the analysis process) and in which the fundamental condition for their effectiveness is the quality of training data (Koehn and Knowles, 2017; Ott et al., 2018). These three points can be related to both new studies and studies already conducted, for which a review of their results could be expected.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## AUTHOR CONTRIBUTIONS

DK: conceptualization, investigation, writing—original draft, and writing—review and editing. MM: conceptualization, supervision, writing—original draft, and writing—review and editing. Both authors contributed to the article and approved the submitted version.

## FUNDING

This study was supported by the Fulbright Scholarship Program, Fulbright-Masaryk Scholarship no. 2020-28-11.

## REFERENCES


Abusa'aleek, A. (2015). Internet linguistics: a linguistic analysis of electronic discourse as a new variety of language. *Int. J. Engl. Linguist.* 5. doi: 10.5539/ijel.v5n1p135

Afshin, H., and Alaeddini, M. (2016). A Contrastive Analysis of Machine Translation (Google Translate) and Human Translation: efficacy in Translating Verb Tense from English to Persian. *Mediterr. J. Soc. Sci.* 7:40. doi: 10.5901/mjss.2016.v7n4S2p40

Agosti, A., and Rellini, A. (2007). *The Italian LIWC Dictionary: Technical Report*. Austin: LIWC.Net.

Althoff, T., Clark, K., and Leskovec, J. (2016). Large-scale Analysis of Counseling Conversations: an Application of Natural Language Processing to Mental Health. *Trans. Assoc. Comput. Linguist.* 4, 463–476. doi: 10.1162/tacl_a_00111

Amini, H., Farahnak, F., and Kosseim, L. (2019). "Natural Language Processing: An Overview," in *Frontiers in Pattern Recognition and Artificial Intelligence*, eds M. Blom, N. Nobile, and C. Y. Suen (Singapore: World Scientific), 35–55. doi: 10.1142/9789811203527_0003

Andrei, A. L. (2014). *Development and evaluation of Tagalog linguistic inquiry and word count (LIWC) dictionaries for negative and positive emotion*. Mclean: Mitre Corp Mclean.

Araújo, M., Pereira, A., and Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Inf. Sci.* 512, 1078–1102.

Asher, N., and van de Cruys, T. (2018). Content vs. function words: the view from distributional semantics. *Proc. Sinn Und Bedeutung* 22, 1–21.

Avolio, B. J., and Gardner, W. L. (2005). Authentic leadership development: getting to the root of positive forms of leadership. *Leadersh. Q.* 16, 315–338. doi: 10.1016/j.leaqua.2005.03.001

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10),* (France: European Language Resources Association (ELRA)), 2200–2204.

Baeza-Yates, R., and Liaghat, Z. (2017). "Quality-efficiency trade-offs in machine learning for text processing," in *2017 IEEE International Conference on Big Data (Big Data)*, (Boston: IEEE), 897–904.

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). "An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis," in *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, (Porto Alegre: SBC).

Barrett, H. C. (2020). Towards a Cognitive Science of the Human: cross-Cultural Approaches and Their Urgency. *Trends Cogn. Sci.* 24, 620–638. doi: 10.1016/j.tics.2020.05.007

Basnight-Brown, D. M., and Altarriba, J. (2018). "The influence of emotion and culture on language representation and processing," in *Advances in culturally-aware intelligent systems and in cross-cultural psychological studies*, ed. C. Faucher (Berlin: Springer), 415–432.

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguist. Issues Lang. Technol.* 6, 1–26.

Bermel, N. (2014). "Czech diglossia: Dismantling or dissolution?," in *Divided Languages?*, eds J. Arokay, J. Gvozdanovic, and D. Miyajima (Berlin: Springer), 21–37.

Berry, D. S., Pennebaker, J. W., Mueller, J. S., and Hiller, W. S. (1997). Linguistic bases of social perception. *Pers. Soc. Psychol. Bull.* 23, 526–537.

Biber, D. (1991). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Lang. Contrast* 14, 7–34.

Biber, D., and Conrad, S. (2019). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Bjekić, J., Lazareviæ, L. B., Živanoviæ, M., and Kneževiæ, G. (2014). Psychometric evaluation of the Serbian dictionary for automatic text analysis—LIWCser. *Psihologija* 47, 5–32. doi: 10.2298/psi1401005b

Boot, P. (2021). Machine-translated texts as an alternative to translated dictionaries for LIWC. *Open Science Framework* [Preprint]. doi: 10.31219/osf.io/tsc36

Boot, P., Zijlstra, H., and Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch J. Appl. Linguist.* 6, 65–76. doi: 10.1075/dujal.6.1.04boo

Bradley, M. M., and Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report C-1*. Gainesville: University of Florida, Center for research in psychophysiology.

Brewer, M. B., and Gardner, W. (1996). Who is this "We"? Levels of collective identity and self representations. *J. Pers. Soc. Psychol.* 71:83. doi: 10.1037/0022-3514.71.1.83

Carvalho, F., Rodrigues, R. G., Santos, G., Cruz, P., Ferrari, L., and Guedes, G. P. (2019). "Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks," in *Anais Do VIII Brazilian Workshop on Social Network Analysis and Mining*, (Porto Alegre: SBC), 24–34.

Castelvecchi, D. (2016). Can we open the black box of AI?. *Nat. News* 538:20. doi: 10.1038/538020a

Chen, J., Qiu, L., and Ho, M.-H. R. (2020). A meta-analysis of linguistic markers of extraversion: Positive emotion and social process words. *J. Res. Pers.* 89:104035. doi: 10.1016/j.jrp.2020.104035

Chung, C. K., and Pennebaker, J. W. (2018). "What do we know when we LIWC a person? Text analysis as an assessment tool for traits, personal concerns and life stories," in *The SAGE Handbook of Personality and Individual Differences: The Science of Personality and Individual Differences*, eds V. Zeigler-Hill and T. K. Shackelford (Thousand Oaks: Sage), 341–360.

Church, A. T., and Katigbak, M. S. (1989). Internal, external, and self-report structure of personality in a non-western culture: an investigation of cross-language and cross-cultural generalizability. *J. Pers. Soc. Psychol.* 57:857.

Corver, N., and van Riemsdijk, H. (2001). *Semi-lexical categories: The function of content words and the content of function words*. Berlin: Walter de Gruyter.

Cruse, D. A., Cruse, D. A., Cruse, D. A., and Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Cvrček, V., Laubeová, Z., Lukeš, D., Poukarová, P., Řehořková, A., and Zasina, A. J. (2020). Author and register as sources of variation: a corpus-based study using elicited texts. *Int. J. Corpus Linguist.* 25, 461–488.

Daems, J., Speelman, D., and Ruette, T. (2013). Register analysis in blogs: correlation between professional sector and functional dimensions. *Leuven Work. Papers Linguist.* 2, 1–27.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Comput. Linguist.* 47, 255–308.

Demjén, Z. (2014). Drowning in negativism, self-hate, doubt, madness: linguistic insights into Sylvia Plath's experience of depression'. *Commun. Med.* 11, 41–54. doi: 10.1558/cam.v11i1.18478

Dino, A., Reysen, S., and Branscombe, N. R. (2009). Online Interactions Between Group Members Who Differ in Status. *J. Lang. Soc. Psychol.* 28, 85–93. doi: 10.1177/0261927X08325916

Dudău, D. P., and Sava, F. A. (2020). The development and validation of the Romanian version of Linguistic Inquiry and Word Count 2015 (Ro-LIWC2015). *Curr. Psychol.* doi: 10.1007/s12144-020-00872-4

Dudău, D. P., and Sava, F. A. (2021). Performing multilingual analysis with Linguistic Inquiry and Word Count 2015 (LIWC2015). An equivalence study of four languages. *Front. Psychol.* 12:570568. doi: 10.3389/fpsyg.2021.570568

Duff, A. S. (2000). *Information Society Studies (Vol. 3)*. East Sussex: Psychology Press.

Ehrlinger, L., Haunschmid, V., Palazzini, D., and Lettner, C. (2019). "A DaQL to monitor data quality in machine learning applications," in *International Conference on Database and Expert Systems Applications*, eds S. Hartmann, J. Küng, S. Chakravarthy, G. Anderst-Kotsis, A. Tjoa, and I. Khalil (Cham: Springer), 227–237.

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., et al. (2020). Closed and open vocabulary approaches to text analysis: a review, quantitative comparison, and recommendations. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/t52c6

Fuller, S. (2005). Another sense of the information age. *Inf. Commun. Soc.* 8, 459–463. doi: 10.1080/13691180500418246

Gandomi, A., and Haider, M. (2015). Beyond the hype: big data concepts, methods, and analytics. *Int. J. Inf. Manag.* 35, 137–144. doi: 10.1016/j.ijinfomgt.2014.10.007

Gao, R., Hao, B., Li, H., Gao, Y., and Zhu, T. (2013). "Developing simplified Chinese psychological linguistic analysis dictionary for microblog," in *International*

*Conference on Brain and Health Informatics*, (Berlin: Springer International Publishing), 359–368. doi: 10.1007/978-3-319-02753-1_36

Garimella, A., Mihalcea, R., and Pennebaker, J. (2016). "). Identifying Cross-Cultural Differences in Word Usage," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (Japan: The COLING 2016 Organizing Committee), 674–683. https://www.aclweb.org/anthology/C16-1065

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., and Dehghani, M. (2018). Dictionaries and distributions: combining expert knowledge and large scale textual data content analysis. *Behav. Res. Methods* 50, 344–361. doi: 10.3758/s13428-017-0875-9

Gill, A. J., Nowson, S., and Oberlander, J. (2009). "). What are they blogging about? Personality, topic and motivation in blogs," in *Third International AAAI Conference on Weblogs and Social Media*, eds E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng (California: The AAAI Press), 18–25.

Gill, A. J., and Oberlander, J. (2019). "Taking care of the linguistic features of extraversion," in *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, eds W. D. Gray and C. D. Schunn (Mahwah: Lawrence Erlbaum Associates), 363–368. doi: 10.4324/9781315782379-99

Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M. J., Kuo, P. B., Pace, B. T., et al. (2020). Machine learning and natural language processing in psychotherapy research: alliance as example use case. *J. Couns. Psychol.* 67, 438–448. doi: 10.1037/cou0000382

Gottschalk, L. A. (2000). The Application of Computerized Content Analysis of Natural Language in Psychotherapy Research Now and in the Future. *Am. J. Psychother.* 54, 305–311. doi: 10.1176/appi.psychotherapy.2000.54.3.305

Gottschalk, L. A., Winget, C. N., and Gleser, G. C. (1969). *Manual of Instructions for Using the Gottschalk-Gleser Content Analysis Scales: Anxiety, Hostility, and Social Alienation–personal Disorganization*. California: University of California Press.

Gutiérrez-Artacho, J., Olvera-Lobo, M.-D., and Rivera-Trigueros, I. (2019). "Hybrid machine translation oriented to cross-language information retrieval: English-Spanish error analysis," in *World Conference on Information Systems and Technologies*, eds Á Rocha, H. Adeli, L. Reis, and S. Costanzo (Cham: Springer), 185–194.

Haider, T., and Palmer, A. (2017). "Modeling communicative purpose with functional style: Corpus and features for German genre and register analysis," in *Proceedings of the Workshop on Stylistic Variation*, (Stroudsburg: Association for Computational Linguistics), 74–84.

Harley, T. A. (2013). *The Psychology of Language: From Data to Theory*. East Sussex: Psychology press.

Hart, R. P. (2001). "Redeveloping DICTION: Theoretical considerations," in *Progress in Communication Sciences*, ed. M. West (New York: Springer), 43–60.

Hart, R. P., and Carroll, C. (2011). *DICTION: The text-analysis program*. Thousand Oaks: Sage.

Haspelmath, M. (2020). The structural uniqueness of languages and the value of comparison for language description. *Asian Lang. Linguist.* 1, 346–366. doi: 10.3389/fneur.2019.01207

Hasselgård, H. (2013). "Crosslinguistic Differences in Grammar," in *The Encyclopedia of Applied Linguistics*, ed. C. A. Chapplle (Hoboken: Blackwell Publishing Ltd). doi: 10.1002/9781405198431.wbeal0290

Hayeri, N. (2014). *Does gender affect translation?: Analysis of English talks translated to Arabic*. Ph.D. thesis. Austin: The University of Texas.

Hickey, R. (n.d.). *English Linguistics. In English Linguistics in Essen*. Duisburg: University of Duisburg and Essen. https://www.uni-due.de/ELE/

Hieber, D. W. (2020). "The languages and linguistics of indigenous North America: Word Classes," in *The languages and linguistics of indigenous North America: A comprehensive guide (The World of Linguistics 13)*, eds C. Jany, K. Rice, and M. Mithun (Berlin: Mouton de Gruyter).

Hogenraad, R. (2018). Smoke and mirrors: Tracing ambiguity in texts. *Digit. Scholarsh. Humanit.* 33, 297–315. doi: 10.1093/llc/fqx044

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C., Deters, F. G., et al. (2019). Linguistic markers of grandiose narcissism: a LIWC analysis of 15 samples. *J. Lang. Soc. Psychol.* 38, 773–786.

Huang, C.-L., Chung, C. K., Hui, N., Lin, Y.-C., Seih, Y.-T., Lam, B. C., et al. (2012). The development of the Chinese linguistic inquiry and word count dictionary. *Chin. J. Psychol.* 54, 185–201. doi: 10.3389/fpsyg.2021.648677

Iliev, R., Dehghani, M., and Sagi, E. (2015). Automated text analysis in psychology: methods, applications, and future developments. *Lang. Cogn.* 7, 265–290. doi: 10.1186/s13063-015-0931-7

Impana, P., and Kallimani, J. S. (2017). "Cross-lingual sentiment analysis for Indian regional languages," in *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, (New Jersey: IEEE), 1–6.

Internet Users by Language (2021). *Internet World Stats*. Available online at: https://www.internetworldstats.com/stats7.htm (accessed September 24, 2021).

Ireland, M. E., and Pennebaker, J. W. (2010). Language style matching in writing: synchrony in essays, correspondence, and poetry. *J. Pers. Soc. Psychol.* 99:549. doi: 10.1037/a0020386

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., et al. (2019). Emotion semantics show both cultural variation and universal structure. *Science* 366, 1517–1522. doi: 10.1126/science.aaw8160

Johannßen, D., and Biemann, C. (2018). "Between the Lines: Machine Learning for Prediction of Psychological Traits - A Survey," in *Machine Learning and Knowledge Extraction*, eds A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl (Berlin: Springer International Publishing), 192–211. doi: 10.1007/978-3-319-99740-7_13

Johnson, A. (2009). The Rise of English: the Language of Globalization in China and the European Union. *Macalester Int.* 22:39. doi: 10.1089/omi.2017.0192

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., and Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *J. Lang. Soc. Psychol.* 33, 125–143. doi: 10.1177/0261927x13502654

Kailer, A., and Chung, C. K. (2007). *The Russian LIWC2007 dictionary*. Austin: LIWC.Net.

Kennedy, B., Ashokkumar, A., Boyd, R. L., and Dehghani, M. (2021). Text analysis for psychology: methods, principles, and practices. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/h2b8t

Kim, U., Park, Y.-S., and Park, D. (2000). The challenge of cross-cultural psychology: the role of the indigenous psychologies. *J. Cross Cult. Psychol.* 31, 63–75.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., et al. (2020). UniMorph 2.0: universal Morphology. *ArXiv* [Preprint]. Available online at: http://arxiv.org/abs/1810.11101 (accessed September 24, 2021).

Koehn, P., and Knowles, R. (2017). "Six Challenges for Neural Machine Translation," in *Proceedings of the First Workshop on Neural Machine Translation*, (Pennsylvania: Association for Computational Linguistics), 28–39. doi: 10.18653/v1/W17-3204

König, E., and van der Auwera, J. (eds) (2002). *The Germanic Languages*. Oxfordshire: Routledge.

Kornfilt, J. (2020). *Parts of Speech, Lexical Categories, and Word Classes in Morphology. In Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press. doi: 10.1093/acrefore/9780199384655.013.606

Kučera, D. (2020). *Osobnostní markery v textu: Aplikace kvantitativní psychologicko-lingvistické analýzy písemného projevu při popisu osobnosti [Personality markers in text: Application of quantitative psychological-linguistic analysis of written text in personality description]*. Czechia: Jihočeská univerzita v českých Budìjovicích.

Kučera, D., Haviger, J., and Havigerová, J. M. (2020). Personality and Text: quantitative Psycholinguistic Analysis of a Stylistically Differentiated Czech Text. *Psychol. Stud.* 65, 336–348. doi: 10.1007/s12646-020-00553-z

Kučera, D., Haviger, J., and Havigerová, J. M. (2021). *Personality and Word Use: Study on Czech Language and the BigFive*. Available online at: https://osf.io/vdb34 (accessed September 24, 2021).

Laajaj, R., Macours, K., Hernandez, D. A. P., Arias, O., Gosling, S. D., Potter, J., et al. (2019). Challenges to capture the big five personality traits in non-WEIRD populations. *Sci. Adv.* 5:eaaw5226. doi: 10.1126/sciadv.aaw5226

List of Countries Where English Is an Official Language – GLOBED (2019). *Education Policies for Global Development*. Available online at: http://www.globed.eu/wp-content/uploads/2019/11/English_official_language.pdf (accessed September 24, 2021).

Lyddy, F., Farina, F., Hanney, J., Farrell, L., Kelly, O., and Neill, N. (2014). An Analysis of Language in University Students' Text Messages: language In University Students' Text Messages. *J. Comput. Mediat. Commun.* 19, 546–561. doi: 10.1111/jcc4.12045

Magnini, B., Lavelli, A., and Magnolini, S. (2020). "Comparing Machine Learning and Deep Learning Approaches on NLP Tasks for the Italian Language," in *Proceedings of The 12th Language Resources and Evaluation Conference*, (Marseille: European Language Resources Association), 2110–2119.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Pennsylvania: Association for Computational Linguistics), 55–60.

Martindale, C. (1973). An experimental simulation of literary change. *J. Pers. Soc. Psychol.* 25:319. doi: 10.1007/s10936-020-09741-4

Massó, G., Lambert, P., Penagos, C. R., and Saurí, R. (2013). "Generating new LIWC dictionaries by triangulation," in *Asia Information Retrieval Symposium*, (Berlin: Springer), 263–271.

McAuliffe, W. H. B., Moshontz, H., McCauley, T. G., and McCullough, M. E. (2020). Searching for Prosociality in Qualitative Data: comparing Manual, Closed–Vocabulary, and Open–Vocabulary Methods. *Eur. J. Pers.* 34, 903–916. doi: 10.1002/per.2240

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., et al. (2020). "UniMorph 3.0: Universal Morphology," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (France: European Language Resources Association), 3922–3931.

Medvedeva, M., Haagsma, H., and Nissim, M. (2017). "An analysis of cross-genre and in-genre performance for author profiling in social media," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, (Cham: Springer), 211–223. doi: 10.1007/978-3-319-65813-1_21

Mehl, M. R. (2006). "Quantitative Text Analysis," in *Handbook of Multimethod Measurement in Psychology*, eds M. Eid and E. Diener (Washington: American Psychological Association), 141–156.

Mehl, M. R., and Pennebaker, J. W. (2003). The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. *J. Pers. Soc. Psychol.* 84:857. doi: 10.1037/0022-3514.84.4.857

Mehl, M. R., Robbins, M. L., and Holleran, S. E. (2012). How taking a word for a word can be problematic: context-dependent linguistic markers of extraversion and neuroticism. *J. Methods Meas. Soc. Sci.* 3, 30–50.

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., et al. (2019). "LIWC auf Deutsch": the Development, Psychometrics, and Introduction of DE-LIWC2015. *PsyArXiv* [Preprint]. doi: 10.17605/OSF.IO/TFQZC

Meneghini, R., and Packer, A. L. (2007). Is there science beyond English?: initiatives to increase the quality and visibility of non−English publications might help to break down language barriers in scientific communication. *EMBO Rep.* 8, 112–116. doi: 10.1038/sj.embor.7400906

Mereu, L. (1999). *Boundaries of Morphology and Syntax*. Amsterdam: John Benjamins Publishing.

Mergenthaler, E., and Bucci, W. (1999). Linking verbal and non-verbal representations: computer analysis of referential activity. *Br. J. Med. Psychol.* 72, 339–354. doi: 10.1348/000711299160040

Milizia, P. (2020). "Morphology in Indo-European languages," in *Oxford Research Encyclopedia of Linguistics*. Available online at: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-634 (accessed June 30, 2020).

Modaresi, P., Liebeck, M., and Conrad, S. (2016). *Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016*. Verona: CLEF. 970–977.

Mønsted, B., Mollgaard, A., and Mathiesen, J. (2018). Phone-based metric as a predictor for basic personality traits. *J. Res. Pers.* 74, 16–22. doi: 10.1016/j.jrp.2017.12.004

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender differences in language use: an analysis of 14,000 text samples. *Discourse Process.* 45, 211–236. doi: 10.1080/01638530802073712

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., et al. (2016). "Universal Dependencies v1: A Multilingual Treebank Collection," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, (France: European Language Resources Association (ELRA)), 1659–1666.

Oberlander, J., and Gill, A. J. (2006). Language with character: a stratified corpus comparison of individual differences in e-mail communication. *Discourse Process.* 42, 239–270.

Osborne, T., and Gerdes, K. (2019). The status of function words in dependency grammar: a critique of Universal Dependencies (UD). *Glossa* 4:17.

Ott, M., Auli, M., Grangier, D., and Ranzato, M. (2018). Analyzing uncertainty in neural machine translation. *Int. Conf. Mach. Learn.* 80, 3956–3965.

Pam, P. (2020). A stylistic investigation of selected internet discourses as tools for national development. *Res. J. Mod. Lang. Lit.* 1, 18–39.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., et al. (2015). Automatic personality assessment through social media language. *J. Pers. Soc. Psychol.* 108:934. doi: 10.1037/pspp0000020

Pennebaker, J., Chung, C., Frazee, J., Lavergne, G., and Beaver, D. (2014). When Small Words Foretell Academic Success: the Case of College Admissions Essays. *PLoS One* 9:e115844. doi: 10.1371/journal.pone.0115844

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin: University of Texas at Austin.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). *The Development and Psychometric Properties of LIWC2007*. Austin: The University of Texas at Austin.

Pennebaker, J. W., and King, L. A. (1999). Linguistic styles: language use as an individual difference. *J. Pers. Soc. Psychol.* 77:1296. doi: 10.1037//0022-3514.77.6.1296

Pennebaker, J. W., and Lay, T. C. (2002). Language use and personality during crises: analyses of Mayor Rudolph Giuliani's press conferences. *J. Res. Pers.* 36, 271–282.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: our Words, Our Selves. *Annu. Rev. Psychol.* 54, 547–577. doi: 10.1146/annurev.psych.54.101601.145041

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., and Pennebaker, J. W. (2011). La version française du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation. *Psychol. Française* 56, 145–159. doi: 10.1016/j.psfr.2011.07.002

Pradhan, T., Bhansali, R., Chandnani, D., and Pangaonkar, A. (2020). "Analysis of Personality Traits using Natural Language Processing and Deep Learning," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, (Piscataway: IEEE), 457–461. doi: 10.1109/ICIRCA48905.2020.9183090

Prates, M. O., Avelar, P. H., and Lamb, L. (2018). Assessing gender bias in machine translation–a case study with Google translate. *ArXiv* [Preprint]. Available online at: https://arxiv.org/abs/1809.02208 (accessed September 24, 2021).

Putri, G. D., and Havid, A. (2015). Types of errors found in Google Translation: a model of MT evaluation. *Proc. ISELT FBS Univ. Negeri Padang* 3, 183–188.

Qiu, L., Lin, H., Ramsay, J., and Yang, F. (2012). You are what you tweet: personality expression and perception on Twitter. *J. Res. Pers.* 46, 710–718. doi: 10.1016/j.jrp.2012.08.008

Ramírez-Esparza, N., Chung, C. K., Kacewicz, E., and Pennebaker, J. W. (2008). "The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches," in *Proceedings of the 2008 International Conference on Weblogs and Social Media*, (California: association for the Advancement of Artificial Intelligence (AAAI)), 102–108.

Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., and Suriá, R. (2007). La psicología del uso de las palabras: un programa de computadora que analiza textos en español. *Rev. Mex. Psicol.* 24, 85–99.

Rayson, P. (2009). *Wmatrix: A web-based corpus processing environment*. Lancaster: Lancaster University.

Riemer, N. (ed.) (2016). *The Routledge Handbook of Semantics*. Oxfordshire: Routledge.

Rijkhoff, J. (2011). *When can a language have adjectives? An implicational universal*. Berlin: De Gruyter Mouton.

Rusínová, Z. (2020). "Sufix (přípona)," in *Nový encyklopedický slovník češtiny online*. eds P. Karlík, M. Nekula and J. Pleskalová (Brno: Masarykova univerzita).

Sánchez-Rada, J. F., and Iglesias, C. A. (2019). Social context in sentiment analysis: formal definition, overview of current trends and framework for comparison. *Inf. Fusion* 52, 344–356. doi: 10.1016/j.inffus.2019.05.003

Sardinha, T. B., and Pinto, M. V. (2019). *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Publishing.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013b). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One* 8:e73791. doi: 10.1371/journal.pone.0073791

Schwartz, H. A., Eichstaedt, J., Blanco, E., Dziurzynski, L., Kern, M. L., Ramones, S., et al. (2013a). "Choosing the Right Words: Characterizing and Reducing Error of the Word Count Approach," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1. Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, (Pennsylvania: Association for Computational Linguistics), 296–305.

Seidlhofer, B. (2011). *Understanding English as a lingua franca*. Oxford: Oxford University Press.

Seki, K. (2021). Cross-lingual text similarity exploiting neural machine translation models. *J. Inf. Sci.* 47, 404–418. doi: 10.1177/0165551520912676

Sharir, O., Peleg, B., and Shoham, Y. (2020). The cost of training nlp models: a concise overview. *ArXiv* [Preprint]. Available online at: https://arxiv.org/abs/2004.08900 (accessed September 24, 2021).

Shibata, D., Wakamiya, S., Kinoshita, A., and Aramaki, E. (2016). "Detecting Japanese patients with Alzheimer's disease based on word category frequencies," in *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, (Japan: The COLING 2016 Organizing Committee), 78–85.

Smith, J., Saint-Amand, H., Plamadă, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). "Dirt cheap web-scale parallel text from the common crawl," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Stroudsburg: Association for Computational Linguistics), 1374–1383.

Søgaard, A., Ruder, S., and Vuliæ, I. (2018). On the limitations of unsupervised bilingual dictionary induction. *ArXiv* [Preprint]. Available online at: https://arxiv.org/abs/1805.03620 (accessed September 24, 2021).

Sonneveld, H. B., and Loening, K. L. (1993). *Terminology: Applications in interdisciplinary communication*. Amsterdam: John Benjamins Publishing.

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., et al. (2020). Personality research and assessment in the era of machine learning. *Eur. J. Pers.* 34, 613–631. doi: 10.1002/per.2257

Stone, P. J., Bales, R. F., Namenwirth, J. Z., and Ogilvie, D. M. (1962). The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. *Behav. Sci.* 7:484. doi: 10.1002/bs.3830070412

Straka, M., and Straková, J. (2017). "Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe," in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, (Stroudsburg: Association for Computational Linguistics), 88–99.

Stuart-Smith, J., and Timmins, C. (2010). "The role of the individual in language variation and change," in *Language and Identities*, eds C. Lamas and D. Watt (Edinburgh: Edinburgh University Press), 39–54. doi: 10.3389/frai.2020.00046

Świątek, A. (2012). *Pro-drop phenomenon across miscellaneous languages*. Poland: Pedagogical University of Cracow.

Sylak-Glassman, J. (2016). *The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema*. Maryland: Center for Language and Speech Processing Johns Hopkins University.

Tausczik, Y. R., and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 24–54.

Temizöz, Ö (2016). Postediting machine translation output: subject-matter experts versus professional translators. *Perspectives* 24, 646–665. doi: 10.1080/0907676X.2015.1119862

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* 61, 2544–2558. doi: 10.1186/s12888-015-0659-7

Thomas, D. R., and Thomas, Y. L. (1994). "Same language, different culture: understanding inter-cultural communication difficulties among English speakers," in *Proceedings of the International English Language Education Conference: National and International Challenges and Responses* (Kuala Lumpur: Language Centre, Universiti Kebangsaan Malaysia) 211–219.

Thompson, B., Roberts, S. G., and Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nat. Hum. Behav.* 4, 1029–1038. doi: 10.1038/s41562-020-0924-8

Thuy, N. T. T., Bach, N. X., and Phuong, T. M. (2018). "Cross-language aspect extraction for opinion mining," in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, (Piscataway: IEEE), 67–72.

Universal Dependencies (2021). *Universal Dependencies*. Available online at: https://universaldependencies.org/ (accessed September 24, 2021).

Universal Dependencies: Syntax (2021). *Syntax: General Principles*. Available online at: https://universaldependencies.org/u/overview/syntax.html (accessed September 24, 2021).

Van Wissen, L., and Boot, P. (2017). "An electronic translation of the LIWC Dictionary into Dutch," in *Electronic Lexicography in the 21st Century: Proceedings of ELex 2017 Conference*, (Leiden: Lexical Computing Ltd), 703–715.

Vanhove, M. (2008). *From Polysemy to Semantic Change: Towards a typology of lexical semantic associations*. Amsterdam: John Benjamins Publishing.

Vannest, J., Bertram, R., Järvikivi, J., and Niemi, J. (2002). Counterintuitive Cross-Linguistic Differences: more Morphological Computation in English Than in Finnish. *J. Psycholinguist. Res.* 31, 83–106. doi: 10.1023/A:1014934915952

Vivas, J., Kogan, B., Romanelli, S., Lizarralde, F., and Corda, L. (2020). A cross-linguistic comparison of Spanish and English semantic norms: looking at core features. *Appl. Psycholinguist.* 41, 285–297.

Wierzbicka, A. (2013). *Imprisoned in English: The Hazards of English as a Default Language*. Oxford: Oxford University Press.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., et al. (2005). "OpinionFinder: A system for subjectivity analysis," in *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, (Stroudsburg: Association for Computational Linguistics), 34–35.

Windsor, L. C., Cupit, J. G., and Windsor, A. J. (2019). Automated content analysis across six languages. *PLoS One* 14:e0224425. doi: 10.1371/journal.pone.0224425

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., and Kordy, H. (2008). Computergestützte quantitative textanalyse: Äquivalenz und robustheit der deutschen version des linguistic inquiry and word count. *Diagnostica* 54, 85–98. doi: 10.1026/0012-1924.54.2.85

Wolfram, W., and Friday, W. C. (1997). The role of dialect differences in cross-cultural communication: proactive dialect awareness. *Bull. Suisse de Linguistique Appl.* 65, 143–154.

Yano, Y. (2006). Cross-cultural Communication and English as an international language. *Intercult. Commun. Stud.* 15:172.

Yarkoni, T. (2010). Personality in 100,000 Words: a large-scale analysis of personality and word use among bloggers. *J. Res. Pers.* 44, 363–373. doi: 10.1016/j.jrp.2010.04.001

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Zednik, C. (2019). Solving the Black Box Problem: a Normative Framework for Explainable Artificial Intelligence. *ArXiv* [Preprint]. Available online at: http://arxiv.org/abs/1903.04361 (accessed September 24, 2021).

Zijlstra, H., Van Meerveld, T., Van Middendorp, H., Pennebaker, J. W., and Geenen, R. (2004). De Nederlandse versie van de 'linguistic inquiry and word count'(LIWC). *Gedrag Gezond* 32, 271–281.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.