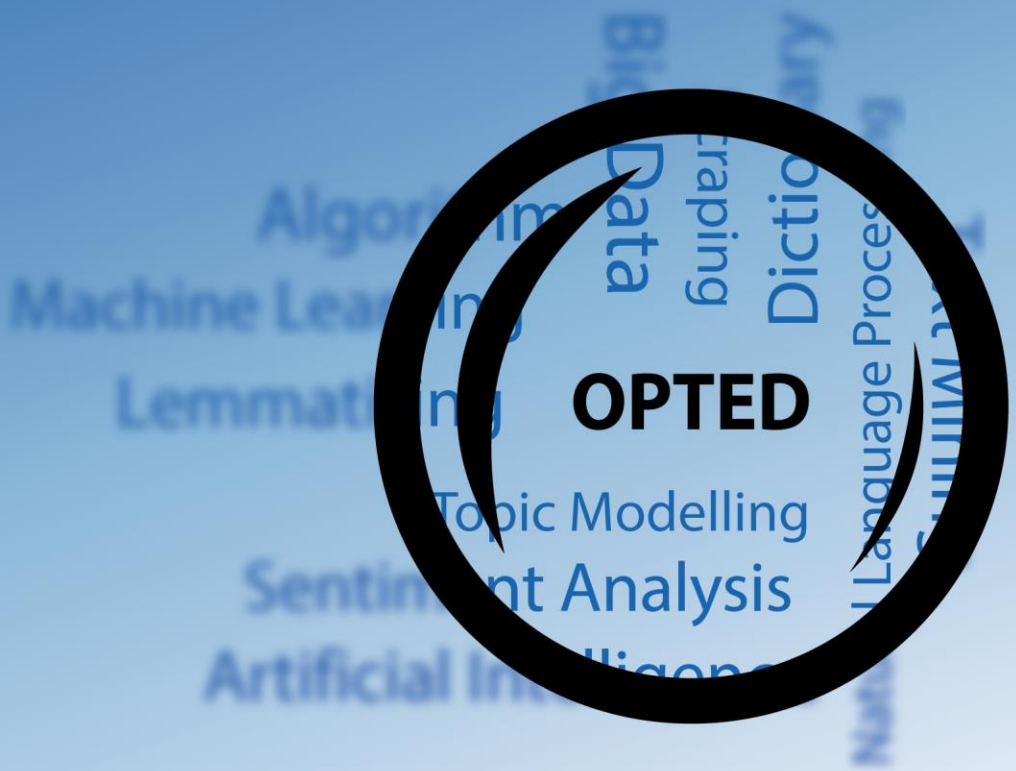
OPTED

How do linguistic differences impact computational text analysis methods? A road map for future validation and integrated strategy development

Guy Shababo, Christian Baden, Alona Dolinsky, Fabienne Lind, Christian Pipal, Martijn Schoonvelde, Anna Smoliarova, Mariken A.C.G. van der Velden, & Avital Zalik



Disclaimer

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report

OPTED
Observatory for Political Texts in European Democracies:
A European research infrastructure

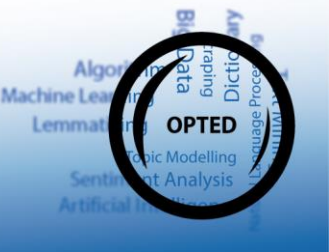
How do linguistic differences impact computational text analysis methods? A road map for future validation and integrated strategy development

Deliverable 6.3

Authors: Guy Shababo[1], Christian Baden[1], Alona Dolinsky[2], Fabienne Lind[3], Christian Pipal[4], Martijn Schoonvelde[5,2], Anna Smoliarova[1], Mariken A.C.G. van der Velden[6], & Avital Zalik[1]

[1] Hebrew University of Jerusalem
[2] University College Dublin
[3] University of Vienna
[4] University of Amsterdam
[5] University of Groningen
[6] Vrije Universiteit Amsterdam

Due date: March 2023



Executive Summary

This Deliverable examines how language differences affect the use of computational text analysis in social scientific research, focusing on potential biases that arise from the intersection between computational algorithms and linguistic differences. Given the monolingual or language-agnostic orientation of most computational methods, as well as the overwhelming dominance of English language applications, we argue that consequential language-dependent challenges to valid multilingual text analysis have not been addressed. We examine some of the main characteristics that differentiate how languages encode information, and examine how these differences interact with computational tools. We show how common modeling assumptions regularly collide with linguistic properties, threatening the validity and comparability of results across languages. We thereby aim to alert the community of computational social scientists to the need to actively consider language as a source of variation and bias that is in dear need of scholarly attention.

1 Introduction

Recognizing multiculturalism and multilingualism as one of its cornerstones, the European Union has long ago decided to confront the challenges raised by its linguistic diversity and turn it into an active field of policy-making, development, and research. There are 24 official languages in the EU. Of these, twenty represent different branches of the Indo-European language family, three are Finno-Ugric languages (Hungarian, Finnish, and Estonian), and one is Semitic (Maltese). The Indo-European languages can be further divided into Germanic languages (Danish, Dutch, English, German and Swedish), Romance languages (French, Italian, Portuguese and Spanish), Slavic languages (Czech, Polish, Slovak and Slovene), Hellenic languages (Greek), Celtic languages (Irish) and Baltic languages (Latvian and Lithuanian; Katzner, 2002). Of these, three - namely English, French and German - act as "procedural languages" in the EU's many international communication activities, and have dominated also in social scientific research (Baden et al., 2022). At the same time, numerous initiatives in both European politics and research aim to strengthen multilingual culture, exchange, and research that integrates the union's official languages, as well as more than 60 other languages that are widely used in the EU, including indigenous regional and minority languages such as Basque, and immigrant languages such as Hebrew, Chinese and Arabic (Eurobarometer, 2012).

In pace with the growing attention to multilingualism within and beyond the European Research Area, the last decade has seen massive growth in the development of computational tools for digital text analysis, dwarfed only by the even faster expansion of digital corpora available for study (van Atteveldt & Peng, 2018). According to a recent review, almost half of all textual research published in leading social scientific journals makes use of some kind of computational tools (Baden et al., 2021). Textual research in the social sciences is experiencing rapid internationalization, both in terms of the increasing inclusion of non-English speaking research sites, and in terms of comparative studies that juxtapose discourses produced in multiple languages. Both developments are of critical importance especially for the European Research Area, which offers countless opportunities for comparative social scientific research, but features numerous linguistic divides. As textual research is undergoing its comparative turn, researchers across and beyond Europe resort to computational tools not only to address the multiplication in required data sizes, but also to achieve measurements that permit a valid comparison across countries, cultures, and languages.

While both developments open up important new avenues for social scientific inquiry, the rapid pace of development comes at the cost of researchers' often only limited understanding of new tools' practical implications for textual measurement and research. As many tools are primarily developed for (and tested on) English language text, researchers regularly face challenges preparing corpora in different languages to meet the requirements of English-thinking computational tools (e.g., Tsarfaty et al., 2013 de Vries et al., 2021; Yarchi et al., 2021); they have reported major losses in tool performance, including pressing validity concerns (e.g., Steimel et al., 2019; Windsor et al., 2019); and especially in the use of computational text analysis methods in cross-linguistically comparative designs, the maintenance of comparative validity has raised important questions (what Chan et al., 2020, called the "Babel problem"; see also e.g., Lind et al, 2019). To date, the dominance of monolingual, English-speaking computational text analysis not only presents a major obstacle for the advancement of social science research and the inclusion of new researchers, research sites and research

questions into the field; In a world that is increasingly penetrated by digital tools, platforms and even decision making algorithms that process digital textual data, this imbalance also presents a tangible threat to equal access to information, equal opportunities and even equal treatment by algorithm-supported corporate and public policy making. It is high time to develop a research agenda focused on how linguistic diversity affects the use of computational text analysis, shaping a truly multilingually inclusive field.

To date, however, there is little methodological knowledge about the challenges raised by the computational analysis of multilingual texts. In computational social science research, most pertinent knowledge derives from practical applications, which permit only little generalization. Other insights stem from efforts at algorithmically bridging linguistic divides, which in the social sciences typically takes place at the level of mapping semantic contents (e.g., using word embeddings, Chan et al., 2020; machine translation, Reber, 2019) without much consideration given to the linguistic encoding of information – what Dundes (1964) called the "set of unique stylistic features or markers" of a text. In the computational sciences, additional bridging strategies tend to rely on implicit, machine-learnt knowledge (e.g., Pires et al., 2019), where linguistic differences are not so much explicated and modeled: rather, tools tend to assume that high-powered machine learning algorithms will compensate automatically for differential textures (e.g., Moon et al., 2019). While there is more systematic attention to linguistic differences in computational linguistics research (Haspelmath et al., 2005), the offered insights primarily address questions located on a level of abstraction well below the study of semantic contents in social scientific research (e.g., Jain et al., 2017). While each discipline offers valuable points of departure for understanding how computational tools process text in different languages, there is no systematic review of how language-specific properties interact with computational tools to create systematic differences in textual measurement. It is the purpose of this Deliverable to provide such a review.

In this Deliverable, we understand multilingual analysis to include any application of computational tools to the study of textual data that bridges between different linguistic codes – either by studying text in multiple languages side by side, or by deploying tools developed in one language to textual materials formulated in different ones.[1] We argue that linguistic differences hold the potential to profoundly affect computational text analysis in three main ways: First, different scripts and morphological conventions can undermine established strategies for mapping unique meanings onto unique tokens, introducing both noise and potential bias into the analysis. Second, languages differ in what additional information is available beyond the lemma (i.e., the word root, which captures the semantic content), potentially resulting in systematic biases in the analysis. Third, differences in word order and morphology can facilitate or undermine the detection of specific forms of grammatical and semantic relations within the text. These linguistic differences are liable to creating systematic biases in algorithmic measurement, which – if undetected – not only erode the performance of tools, but which potentially threaten the validity of computational text analysis research.

To identify key interactions between linguistic properties and textual measurement, we primarily draw knowledge from computational linguistics and language typology, which we contextualize against reported experiences by researchers in the social and computational sciences. We synthesize six types of linguistic differences that appear especially consequential for computational text analysis: 1) the encoding of languages using different scripts; 2) languages' tendency to create polymorphemes (words comprised of multiple smaller words); the morphological encoding of 3) case and grammatical function, 4) tense, aspect and mood, and 5) gender and number; and 6) the implications of different word orders. After a brief discussion of important variations with regard to each topic, we discuss how common computational algorithms should be affected, owing to embedded assumptions about the textual expression of relevant information. Throughout our discussion, we focus on linguistic differences in a narrow sense, excluding variation that may arise from different cultural norms of language use (e.g., Park et al. 2013), jargons or socio-linguistic differences (e.g., Eckert, 2001). For the purpose of this Deliverable, we consider both how different languages express information in different ways, and in particular how these variations collide with common assumptions hard-coded into (English-thinking) computational tools.

[1] We do not consider here the case of multilingual texts machine-translated into a common language, a strategy that often suffices to perform certain (e.g., thematic) analyses (e.g., Reber, 2019) that do not depend on the accurate preservation of grammatical structures or other nuances that tend to get lost in translations especially from less well-resourced languages (Dabre et al., 2020). That said, many of the issues discussed hereunder also appear to affect the accuracy of machine translations. Wang et al. (2021) provide a useful review of the state of the field, outlining the current main challenges.

2 The Monolingual Default of Computational Text Analysis

Computers have long been recognized as valuable tools for studying textual information. Traditionally, much computational text analysis has relied on researcher-created lists of explicit indicators - such as dictionaries measuring references to named entities or broader topics, or complex rule sets aiming to capture the overt grammatical structure of text (e.g., Popping, 2000). Most such uses were initially developed for the study of English language texts, owing to the leading role of anglophone countries in both computing and social science research. Language was rarely a pressing concern: On the one hand, most users were interested in the study of texts formulated in English, or a handful of other languages very similar to English (notably, Dutch; van Cuilenburg et al., 1985). On the other hand, such early, 'rule-based' methods adapted in obvious ways, if somewhat effortfully, to different languages: To measure constructs in different languages, indicator and rule sets would either have to be translated (Lind, 2019) or new tools would need to be developed to match the studied linguistic and cultural context (Grimmer & Stewart, 2013). By today, several dictionaries have been translated, including those dictionaries contained in LIWC (Pennebaker et al., 2015) or the moral foundations dictionary (Graham & Haidt, 2012). In addition, there exists a growing number of cross-lingually comparative studies using rich self-developed, multilingual dictionaries, with researchers laboriously validating the equivalence of recognized contents (e.g. Baden & Tenenboim-Weinblatt, 2018; Lind et al, 2019).

Since these early days, however, numerous new computational tools have been devised (Boumans & Trilling, 2016; van Atteveldt & Peng, 2021). Earlier unsupervised techniques were still lean in their modeling of linguistic properties. For instance, Latent Semantic Analysis (Landauer et al., 1998), a precursor of contemporary word embeddings and word vector models, mostly assumed that words represented unique meanings, which co-occur in documents according to their semantic similarity. More recent unsupervised techniques often include specific target functions that encode assumptions about the linguistic representation of modeled meanings: For instance, topic modeling algorithms (Blei & Lafferty, 2006) assume that topicality is primarily expressed by the consistent co-use of medium-common words within the same documents, with minimal but generally permitted overlaps between those words used to expressed different topics. In supervised machine learning, algorithms regularly make far-reaching assumptions about which elements of the text (e.g., unigrams, bi- or higher n-grams, or even specific character sequences) are indicative of sought-for information, and which kinds of algorithms (e.g., vector-, clustering-, or Bayesian algorithms, neural networks, transformer models) are suitable to separate relevant from irrelevant regularities (Grimmer & Stewart, 2013).

This move from fully researcher-controlled to partly or fully algorithmic text analysis has been productive, to say the least: Especially in the context of proliferating digital media, current research all across the social sciences and humanities relies heavily on algorithmic procedures to model, filter, aggregate, and represent textual patterns of different kinds (Baden et al., 2021). At the same time, the delegation of key analytic decisions to algorithmic, hard-coded modeling assumptions has created a situation in which it is not always apparent how a method processes a given text, and where researchers' limited direct contact with the text erodes their ability to evaluate whether the chosen procedures are suitable for yielding valid and comparable results.

This problem has been aggravated by the progressing internationalization of textual social science research (Kuhn & Weidemann, 2015). While research from anglophone countries and on English-language text remains dominant, there has been a rapid growth of computational text analysis capacities and applications in many other Western and non-Western languages (Baden et al., 2021). At the time of writing, virtually any major language community sustains its own computational text research initiatives (see, for instance, clarin.eu). Drawing chiefly on the existing, English oriented toolbox of computational tools, one major development has focused on translating existing procedures into other languages, resulting in a growing availability of multilingual dictionaries and NLP tools that have been adapted to process different languages (e.g., the *spaCy* and *gensim* python packages). At the same time, especially newer, more machine-reliant tools have been widely applied to different languages without or with only minor changes: Unlike the evident language-boundedness of rule-based tools, most unsupervised and supervised methods are widely perceived as language independent, even if they have been developed using English-language text (Bender, 2011).

In the process of such applications, alas, numerous incompatibilities have been discovered. Identical tools perform notably worse on different languages, and regularly fail to return comparable results (e.g., Steimel et al., 2019; Prettenhofer & Stein, 2010); oftentimes, significant adaptations must be made to different-language texts before these can be processed by existing computational technologies (e.g., de Vries et al., 2021). In this

paper, we agree with Emily Bender's (2011) observation that language-independent text analysis methods cannot rely on unified, putatively language-agnostic algorithmic architectures; rather, they require an explicit awareness of how languages encode the same information and meanings in different ways, necessitating conscious adaptations in algorithmic procedures to achieve equivalent results. In our pursuit of this goal, in the following chapter, we identify and discuss six key variations in how different languages express key information that may be measured using computational tools.

3 Linguistic Concerns

From a semantic point of view, the English word "writer" is a valid translation of the Hebrew word "סופר". Despite this, there are several noteworthy differences. To begin, Hebrew specifies the gender of "סופר" - in this case, male - while English does not. Moreover, the Hebrew word is polysemic: It can also mean "[he] counts", or appear as transliteration of "Super" in the names of markets and retail stores – spelled identically, and disambiguated only by the context of use. Third, in Hebrew, the word only refers to an unspecified writer (a writer), as a specific one would be prefixated with a definite article, "הסופר" [the-writer]. Fourth, in its meaning as verb, "סופר" translates into two words in English ("he counts"), while in Hebrew, the pronoun is implied in the gendered form. Inversely, the English "counts" is also polysemic, and can either be a present-tense third-person singular verb or a plural noun derived from the lemma "(to) count", or the plural form of the legal term or the title "count" – variations that each map onto different words in Hebrew. In German, all meanings of "סופר" constitute separate tokens, none of which is polysemic ("Schriftsteller", "er zählt", "super"), and the same is true for the meanings of "counts" ("zählt", "Zählungen", "Anklagepunkte", "Grafen"). In both English and German, "counts" holds the double meaning of "matters", and "super" can also mean "excellent", while the same is not true in Hebrew. Following the same and other examples through different languages reveals the impressive variety of ways that languages use to encode even simple meanings.

From an analytic point of view, many of these differences are, at first glance, inconsequential. Human coders intuitively process text in any language they are fluent in, and are capable of recognizing semantic equivalencies regardless of the language. Likewise, many algorithms target linguistic patterns that are expressed in more or less similar ways in multiple languages. For instance, named entities tend to appear as separate, contiguous expressions in most languages, and most languages rely on a combination denotative evaluative adjectives and adverbials, as well as connoted nouns and verbs to express evaluative sentiment. To the extent that similar contents are encoded in similar ways in different languages, we plausibly expect computational tools to perform in similar ways.

In other cases, differences exist, but can be easily harmonized by removing morphological features (e.g., lemmatizing to remove case information, redacting the Hebrew definite prefix ה-), or deleting stopwords (e.g., prepositions). However, not all linguistic differences are readily harmonized, and meaningful incommensurabilities remain between languages that may bias algorithmic analyses.

Of course, it is also possible that additional variation is indeed informative, permitting additional analyses in those languages that encode certain information (e.g., gendered languages permit studying what activities are more commonly attributed to which gender). Wherever one language encodes relevant information while the other one doesn't, algorithms may end up creating consequential biases in the analysis – ranging from trivial errors (e.g., when a dictionary fails to capture a verb in all tenses, genders and numbers) to complex and difficult-to-detect biases (e.g., if a supervised machine classifier has access to case information in one language, but not in another). To enable a conscious, valid treatment of linguistic differences, we first of all require an understanding of how languages differ, before we can consider how present differences might impact the computational analysis of text.

Linguists have, of course, been long aware of a wide range of linguistic differences. Existing work in language typology has identified extensive lists of common variations between languages that affect how similar meanings are expressed in different ways, accumulating information far beyond what could be discussed in any methodological paper. The World Atlas of Language Structures (wals.info) lists no less than 192 entries or linguistic features that vary between languages (Haspelmath et al., 2005). While most variations are defined at a low level of abstraction where it remains difficult to identify systematic implications for

computational text analysis, many entries can be synthesized into broader differences in how languages conceptualize important information: For instance, more than a dozen WALS entries jointly concern the way in which a noun's gender, number and case is expressed; almost a dozen entries record variations in the relative positioning of subject, verbs, objects, and qualifiers. In place of trying to capture all the documented variation in linguistic detail, we focus here on six broad patterns concerning languages' use of 1) different scripts, their morphological rules for 2) word formation and the expression of 3) grammatical functions, 4) tense, aspect and mood, 5) gender and number, and finally their 6) rules for sentence formation and word order. While each pattern subsumes a variety of specific differences, each raises a common set of implications for computational analysis that is suitable to guide our analysis. Importantly, the discussed differences are neither binary nor pure, but apply to varying degrees, and permit for important exceptions. In fact, linguists have developed a range of indices for scoring languages on pertinent features (Greenberg, 1960): Languages' preference for separating (so-called "analytic" languages) or concatenating words ("agglutination", in so-called "synthetic" languages) has been captured in the "Index of Fusion" (Fleischacker, 1992). Similarly, an "Index of Synthesis" counts the average number of morphemes per word in a corpus, ranging from 1.06 (Annamite, the Mon-Khmer language spoken in Vietnam; almost all words correspond to exactly one concept) to 2.55 (Swahili; most words are comprised of multiple components that contribute different meanings), with modern English (1.70) located somewhere in between (Greenberg, 1960).

For the purpose of this paper, we do not aim to do justice to each discussed language and linguistic feature, but to flag common aspects of how languages differ in ways capable of affecting computational tools in predictable ways. We are deliberately sidelining the more common distinction of language families, which reflect languages' historical origins but may be misleading when considering the structural similarity of different languages (Haspelmath et al., 2005): While some features are indeed more likely to be shared among related languages (e.g., most Indo-European languages use the Roman alphabet and SVO word order), similar properties are often shared across language families (e.g., German, Russian and Swahili all tend to agglutinate morphemes), while important differences exist also between closely related languages (e.g., most Austronesian languages know no gender, but Tagalog does). In the following, we will discuss some of the most pertinent structural differences between languages in turn.

3.1 Script

A first, obvious difference is the writing system that a language uses. At first glance, script is somewhat orthogonal to language: Many languages have historically changed their script (e.g., Turkish, Mongol, Vietnamese), and in some cases, the same language uses multiple scripts (e.g., Serbian, which can be written in either Cyrillic or Roman letters) – all without significant change in the language itself. However, while different alphabetic scripts are somewhat interchangeable, the same is not true for the comparison between alphabetic, abjadic, syllabic and logographic writing systems. Unlike alphabetic scripts, for instance, abjads (vowel-less scripts) do not specify significant parts of the phonetic qualities of words, with the effect that different words are spelled identically. While the pronunciation – and with it, the identity of the word – can be specified by use of diacritical signs (distinguishing, for instance, "סוֹפֵר" [writer/he counts] from "סוּפַּר" [super]), these are typically omitted in practice, so that different meanings are disambiguated only by their context of use. Of course, homonyms exist occasionally also in alphabetic languages – however, homographs are immensely more common in Hebrew and Arabic, the two main modern languages using abjadic scripts (Tsarfaty et al., 2019). Logographic scripts likewise tend to create a large range of homographs, such as the use of identical characters to express a sound (e.g., in names) or a word. As a consequence, some scripts erode one key assumption in computational text analysis, namely, that each unique character string usually corresponds to exactly one word, so that different words can be separated by their use of different lexical tokens (Farghaly & Shaalan, 2009).

3.2 Morphology: Polymorphemes

Morphology – the rules by which words are formed – comprises a wide range of ways in which languages differ, many of which are highly specific and therefore of lesser interest here. For the systematic study of text

through computational methods, we identify four broad themes that apply widely and result in consequential differences between languages.

A first important difference in the morphology of languages concerns how multiple words can be combined into larger words ("polymorphemes") to express more complex meanings. On the one hand, "analytic" languages tend to keep constituent morphemes separate, expressing their connection by a combination of adjacency and, in some cases, minor inflections. For instance, "health insurance" describes a meaning constituted jointly by both words, but both words remain separate and uninflected. In Hebrew, both words remain separate ("קופת חולים", literally: cash register [for] sick people) but the suffix -ת marks that both are part of a joint construct state. On the other hand, "synthetic" languages tend to concatenate words, sometimes with small inflections interspersed. For instance, German composes "Krankenversicherung" as a new word from the constituent morphemes "Kranke" (sick people) and "Versicherung" (insurance). Other languages fail to neatly fall into this dichotomy: For instance, Chinese constructs polymorphemes through the use of multiple glyphs to represent one construct, which are represented without spaces but each represent separate (and thus separatable) meanings; however, as the entire sentence is encoded without spaces, the language provides no obvious clue as to which characters jointly encode one complex meaning (Yao & Lua, 1998). In this way, 健康 [health] plus 保 [save] and 險 [risk], taken together, mean "health insurance". However, if parsed differently, the middle two characters 康保can also combine to mean "recreational".

At the same time, the constituent constructs tend to remain unaltered in languages using logographic scripts, which modify meanings and construct composite concepts by means of adjacent glyphs (e.g., 病 [sick] plus人 [person] yields 病人 [sick person]). By contrast, most Western languages permit a wide variety of edits to the constituent roots, ranging from pre-, in- and suffixations to truncation that complicate re-identifying the included roots from an agglutinated polymorphemes, especially where derived forms involve changes within the root itself (e.g., in German, "krank" yields "erkranken" [getting sick], "Krankheit" [sickness], but also "kr**ä**nklich" [sickly]).

As a consequence, the use of polymorphemes in a language directly influences the level of semantic abstraction at which textual tokens represent the expressed meaning. Both synthetic and analytic languages are broadly consistent with the idea that each token typically represents one kind of meaning; however, in analytic languages, "one kind of meaning" tends to refer to rather broad domains (e.g., "health"), while more specific meanings arise only from the combination of multiple tokens. By contrast, in synthetic languages, many tokens refer to rather specific concepts that constitute fully independent tokens even for closely related meanings ("Versicherung" [insurance], "Krankenversicherung" [health insurance], "Krankenversicherungsschein" [health insurance certificate]). While the constituent tokens in analytic languages are relatively common and appear in variegated contexts, many synthetic tokens appear infrequently, and in notably more limited contexts. As a consequence, analytic languages represent both semantic relatedness and discursive co-use through the regular association of tokens, whereas semantic relatedness is unavailable as part of the analysis of synthetic languages – unless efforts are made to re-identify, separate and homogenize the constituent roots. Inversely, the relatively specific tokens found in synthetic languages facilitate the distinction of specific uses, while the overlapping use of the same tokens in analytic languages often blurs which associated third expressions pertain to which referenced meaning.

3.3 Morphology: Case/Function

Beyond the formation of words, next, morphological variation also concerns the use of inflections attached to words in order to encode additional information about their grammatical function, tense, aspect and modality, and gender and number. From a computational text analysis point of view, inflections matter because they create variations of words that refer to the same kind of meaning but are used in different ways.

With regard to the marking of case or grammatical functions, the same information may be encoded by the use of prepositions, articles and other separate function words, by use of inflections, or not at all. While some languages (e.g., English) encode most of its grammar into separate words (the exception being genitive case, which is marked morphologically), other languages rely on a mix of both strategies (e.g., Russian morphologically encodes six cases, but uses prepositions for other functional specifications), or encode almost

all grammatical functions into elaborate inflections (e.g., Hungarian, Turkish). In a word-level analysis of English language text, most grammatical information is thus ignored, and a recognition of grammatical functions either requires a separate step of adding grammar tags, or needs to consider at least trigrams (e.g., to capture differences such as "**to** the island", "**on** the island", or "**of** the island"). By contrast, an analysis of Turkish text will by default retain most grammar information, differentiating between these variants ("ada**ya**", "ada**da**", "ada**nın**"). As a consequence, tokens obtained from a Turkish text will generally distinguish between "dog bites man" and "man bites dog", while the same is not true for English unless expressly modeled.

3.4 Morphology: Tense, Aspect and Mood

Tense, aspect and mood[2] can likewise be encoded either by adding separate words, through inflections or variations in the verb itself, or not at all. As an example of the former, the Chinese verb 去 [to go] specifies only the action itself, but not whether and when this action takes place. To identify the time, duration and modality of the event, other words must be added (e.g., 去了 [went; literally: "go, past"]). English, by contrast, marks some of the information in the verb itself (e.g., "go", "went", "gone", "going"), while other parts are marked using auxiliary verbs (e.g., "was going", "will go"). In Hindustani, almost all tense, aspect and modality information are expressed through morphological variation. In addition, languages make different distinctions: In English, "went" is different from "was going", "has been going" or "used to go", however, not all of these variants have exact translations in all other languages. Hebrew encodes transitivity and reflexivity as part of the verb (e.g., "הלך" [went], "התהלך" [went away]), and Russian encodes aspect (e.g., "**ид**ти" [go-recurrently] "**пой**ти" [go-once]), but neither distinguishes between ongoing and completed action (e.g., "going", "go"; Romeo, 2009). In English, present, future and imperative use the same word ("go", "will go", "go!") and the past is distinct ("went"), while in Hebrew, all tenses use different forms, and Arabic even distinguishes two variants of imperative/jussive mood – the latter expressing a lawlike need to do something. Some languages mark subjunctive mood – German does so in multiple shades (Palmer, 2001) – while others (e.g., Hebrew) do not. In each language, therefore, different information is available for analysis, and different variants of meaning appear as equivalent to the machine.

More practically speaking, in some languages, tense, aspect and modality can be redacted by deleting regular affixes (e.g., in Russian, infinitive verbs generally end on -ть, while past-tense verbs on -л [masculine singular], -ла [feminine singular], or -ли [plural]); in English, this strategy works for some (e.g., "go**ing**", "laugh**ed**") but not all inflections (e.g., "went", "gone"); and in Hebrew, tense, aspect and mood are expressed by complex combinations of pre- and infixes that defy straightforward stemming or lemmatization.

3.5 Morphology: Gender/Number

While most languages encode grammar and tense, languages differ widely in whether and which genders are available as linguistic categories at all. In Brazil, for instance, most Arawak languages have two genders, Palikur has three, while Amuesha, Waaurá and Terêna make no gender distinction at all (Payne, 2017). Even where grammatical gender exists, there are vast differences in how gender is manifested in the language. English technically knows three genders, however, its use is limited to pronouns, while neither nouns nor adjectives nor verbs encode gender. German has three genders as well, but applies it also to nouns and adjectives, using gendered articles, as well as different endings for nouns and adjectives (e.g., "ein**e** gut**e** Student**in**" [a-female good-female student-female]). Dutch has three genders for pronouns and possessive forms, but otherwise distinguishes only living (either male or female) from neuter. French and Hebrew both have two genders only, however, also verbs can respond to the gender of the noun (e.g., in the French imperfect: "un**e** bon**ne** étudiant**e** est allé**e**" [a-female good-female student-female go-past-female]).

Most languages mark number, but again, marking may be largely limited to nouns or include adjectives and verbs, as well. Gender and number may interact, and some languages have different plural (and sometimes

[2] While tense locates an action in time, aspect describes the state of this action at that time, and mood distinguishes whether the action is in fact taking place, or remains a possibility or demand.

dual; for instance, in Arabic, where it serves also as polite register) forms for different genders, often defaulting to the male plural form for heterogeneous groups. In fact, several languages are presently undergoing transformation of their gendering practices, establishing new neutral forms or using female morphological forms as part of a political effort (Prewitt-Freilino et al., 2012). Even in Chinese, modern writers have started combining the Latin letter *X* with the character 也 (literally: "also", but in this case it serves as a graphic element of the glyphs, rather than a stand-alone glyph) to produce a gender-neutral case. Such distinctions often fail to translate into languages that lack such distinction, as was neatly illustrated when the *Times of Israel*'s English edition spent a full paragraph to explain to its readers the meaning of an Israeli minister's unconventional use of the female plural form for addressing the public.[3] For the analysis, gender and number create important differences in what will be distinguished: In Arabic, one, two or many women or men doing the same thing will each create different morphological forms of nouns, adjectives, and verbs; in German, we get one female, one male and one plural version for each noun and adjective, while in English, only plural nouns and the third person singular form of the verb are distinguished.

3.6 Word Order

The final type of linguistic difference that we consider here concerns the ordering of words in a sentence. To illustrate the variability of word orders, consider one of the most widely translated sentences in the world, the opening statement of the Book of Genesis (Gen 1:1). In its English rendering, the sentence reads "In the beginning, God created the heavens and the earth", placing the subject before the verb and object (SVO), as is the case also in Modern Hebrew and a majority of Western languages. The original in Biblical Hebrew, which (like Irish or Arabic) followed a VSO word order, is subtly different: "In the beginning created God the heavens and the earth." Korean and Japanese both follow a SOV word order that places the verb at the end of the sentence: "God, the heaven and the earth created." Perhaps the most distinct group of languages is exemplified by Latin, which follows a free word order, a phenomenon also known as "scrambling"[4] (Sabel & Saito, 2005). While fully free word order has become rare in living languages, modern languages such as Hungarian, Polish and Japanese share some of Latin's flexible word order; and even language with generally strict word order sometimes permit freer sequences (e.g., in poetry or for emphasis) or use different orderings for emphasis.

From a linguistic point of view, word order affords a range of subtle variations in meaning to be encoded by the positioning of words. In many languages, word order governs which out of multiple named entities acts as subject or object – as in the famous case of "man bites dog". However, word order often conveys much richer information. Consider, for example, the sentence "He gave her the rose." While the word "only" can be placed anywhere within this sentence, each word position conveys different meaning. Especially for qualifiers, word order is frequently critical not only for determining which quality refers to which entity or activity, but also for managing more nuanced meanings (Koktova, 2011). At the same time, word order typically encodes only part of the relational structure of the sentence, and is complemented – or in case of free-order languages, replaced – by morphological markers, such as object case inflections or case, number and gender congruency. Depending on how a language divides the labor of marking relatedness and grammatical roles between word order and morphology, different strategies may be required to recuperate the relevant information.

For computational analysis, differential word orders create complications mostly in three ways. Where algorithms disregard word order (most notably, in bag-of-words [BOW] algorithms), subject-object differentiation and relational information is largely removed to the extent that a language uses word order to encode these; however, where the same information is encoded via morphology, it remains present. Next, many algorithms approximate relatedness as adjacency, an assumption that holds much better in some languages than in others and may foreground either subject-verb associations (in SVO languages), subject-object associations (in SOV languages), or both (in VSO languages). Third, any algorithms that aim to reconstruct parts of the

[3] https://www.timesofisrael.com/labors-1-woman-opposition-no-to-a-religious-occupying-non-democratic-state/

[4] Consider, for instance, the famous opening lines of Ovid's metamorphoses: "In nova fert animus mutatas dicere formas / corpora…", which translates literally as "into new pushes mind changed to-speak forms entities", and requires a reader to first reconstruct the sentence's grammatical structure, which is encoded solely via congruent case/number/gender, before a translation can be attempted: "I intend to speak of forms changed into new entities".

grammatical or relational structure of expressed meaning need to take into account that equivalent information is found in systematically different places within the sentence (e.g., Fogel-Dror et al., 2019).

In practice, none of the linguistic differences discussed above are necessarily pure or dichotomous, and there exist numerous additional variations that may be less common, but nevertheless hold important implications for computational analysis. Still further challenges arise where texts are in themselves heterogeneous – e.g., where English language phrases are embedded within different-language text, possibly in transliterated form, or when a text switches between languages or character systems for certain contents (e.g., Jurgens et al., 2014). While we cannot possibly hope to offer a complete mapping of key issues, we argue that the discussed differences are both common and consequential, and therefore worthy of attention. In the next section we discuss how these presented sources of linguistic variation interact with algorithmic procedures commonly applied in computational text analysis, and what implications may arise from such collisions.

4 When algorithms meet different languages

Algorithmic analyses of text come in countless variants and serve numerous purposes, each of which makes different assumptions about the nature of meaning encoded in the text, and the manner of its encoding. Algorithmic assumptions are not always apparent, however, and may be buried deep within the code of computational tools. In the following, we will review a range of classic algorithms that are commonly used in computational text analysis, with a specific interest in what assumptions these algorithms make to achieve their analytic objectives, and how these assumptions fare as they are applied to different languages with different qualities. In doing so, we recognize that algorithmic models necessarily disregard many linguistic properties of text and purposefully simplify other aspects, which are out of focus for an intended analysis (Smaldino, 2017). However, for researchers to be able to trust the performance of algorithmic procedures and arrive at reasonably equivalent analyses in different languages, it is necessary to examine possible mismatches between the way that specific meaning is encoded in a language, and those ways in which an algorithm attempts to decode it.

In our examination, we distinguish broadly between (1) preprocessing algorithms, which serve to focus subsequent analysis on meaningful variation in the data, harmonizing what is equivalent (Denny & Spirling, 2018); and (2) analytic algorithms, which operate upon preprocessed or un-preprocessed data to answer tangible research questions (Boumans & Trilling, 2016). This distinction may appear anachronistic, given that many more recent language models – notably, BERTs – include built-in strategies for statistically representing textual data and are commonly believed to obviate the need for preprocessing; however, against this contention, we will argue that preprocessing still plays a key role both for focusing subsequent algorithms on relevant information and ensuring equivalence in multilingual analysis.

4.1 Preprocessing

Preprocessing can be broadly defined as any changes that are applied to textual data with the intention to remove all variation that is deemed uninformative toward a given research question, retaining only variation, and all variation, that is needed to perform the analysis (Denny & Spirling, 2018). In doing so, preprocessing can strike different trade-offs between inclusiveness and focus, removing, retaining or sometimes adding variation in the textual data. The primary evaluative standard for preprocessing is its capacity to validly separate relevant information from irrelevant variation, resulting in different best practices depending on the specific analysis in question (Maier et al., 2021). In the context of applying computational text analysis in a multilingual world, however, preprocessing can additionally be evaluated against the desiderate that the variation retained for analysis reflects reasonably *equivalent* information from different language texts (Baden et al., 2022).

Within preprocessing, we can broadly distinguish between four key interventions in the data: a) Tokenization or segmentation concerns the process by which text is parsed into smaller constituent units – which may represent words, but also groups of words or character sequences below the level of words (e.g.,

Goldberg & Elhadad, 2013); inversely, concatenation can be used to merge multi-word expressions; b) Lemmatization and stemming are used to remove variation within tokens that is considered uninformative, ideally creating a situation where identical tokens carry identical information (Khyani et al., 2021); c) Cleaning, filtering, pruning and stopword removal serve to redact entire tokens whose occurrence is deemed unrelated to the analysis in question (Maier et al., 2021; Scott & Matwin, 1999); beyond these, finally, additional preprocessing steps can be used to d) augment the information offered by the tokens alone, for instance, by disambiguating homographs, or adding grammatical information (Somayajula et al., 2022). While tokenization, lemmatization and augmentation primarily govern which variations of identical roots will be distinguished or treated as identical, lemmatization and cleaning primarily concern the removal of any information about case, gender, number, tense, and other details that are deemed irrelevant for a given analysis.

4.1.1 Tokenization

Tokenization is a procedure by which raw text is split into separate tokens that represent unique meanings. In most English-language text analysis, tokenization is a non-issue: Since English is, by and large, an analytic language, most space-delimited sequences of characters correspond to separate words and can be used as tokens without much need for further processing. However, this is not true for many other languages. In languages using syllable-based logographic script, such as Chinese, text is typically encoded without spaces between characters, rendering the process of tokenization ambiguous. While algorithms exist that try to infer which characters belong to the same word (e.g., Gao et al., 2005), Yao and Lua demonstrate how easily available methods fail with even simple sentences (Yao & Lua, 1998) – chiefly because characters with valid joint meaning may still occur as separate words. For the Japanese and Korean usage of Chinese characters (*Kanji* and *Hanja* respectively) the process is even more complex, as logographic symbols are incorporated into other scripts.

Beyond Eastern languages, tokenization is often useful to focus the analysis on unique meaning-carrying tokens. One issue concerns the case of polymorphemes that concatenate multiple words, as is common in languages such as Russian or German: While there is nothing intrinsically wrong about treating "Bundestagsuntersuchungsausschuss" [inquiry committee of the federal parliament] and not its constituent morphemes as the token, the high specificity – and thereby, low frequency – of many of the most informative tokens in synthetic languages severely limits their capacity to inform statistical analyses (Schick & Schütze, 2020). Whenever languages differ with regard to their tendency to form polymorphemes, tokens obtained as space-delimited words tend to be structurally dissimilar from those obtained from more analytic languages, such that a comparative analysis may need to obtain tokens at a comparable level of abstraction. Next to the splitting of polymorphemes into their constituent roots, also concatenating multi-word construct states into single tokens (e.g., "בית_חולים" [hospital]) can be useful to obtain equivalent representations of the text (Otani et al., 2020). Concatenation can also help distinguish names such as the "White House" from a "white house", which can be recognized as common token sequences (and in some languages, by relying on capitalization as a cue; Attia, 2007) by using named entity recognition tools.

In languages that express grammatical functions and roles by means of affixing prepositions, conjunctions, articles, and other function words, tokenization is often inevitable to enable the identification of semantically relevant expressions. Without tokenization, the Hebrew "בית" [house], "הבית" [the-house], "והבית" [and-the-house], "ושבביתך" [and-that-in-your-house] and many other expressions each appear as independent tokens. Only after tokenization can a computer recognize the recurring reference to the same word and separate it from the affixated articles, prepositions, and conjunctions (Goldberg & Elhadad, 2013).

In comparative analyses, omitting required tokenization steps not only erodes statistical power (by creating many semantically redundant, rare tokens), but also threatens comparative validity, as analyses in one language will treat as independent what is recognized as identical in another. The frequent unavailability and often low accuracy of tokenization algorithms (Escartín, 2014) thus presents a major obstacle to computational text analysis especially in languages with pronounced synthetic tendencies or logographic writing systems.[5]

[5] One recent strategy to overcome present tokenization challenges is to use algorithms that use more complex tokenization strategies that operate on a sub-word level (e.g., BiRNN-CRF, back-propagating character-level models such as LSTM;

In practice, tokenization issues are often downplayed. Studying word embeddings in 25 different languages, for instance, van Paridon and Thompson (2021) argue tokenization is unnecessary, since all included languages are space delimited. However, their languages include several that affix grammatical function words (notably, Hebrew), effectively constituting each distinct use of the same word as an independent token. Similar issues are reported by Öztürk & Ayvaz, 2018, who perform a sentiment analysis of English and Turkish text. As a good practice example, Schuler (2020)'s study of Vietnamese public discourse recognizes the need to concatenate numerous multi-word expressions into meaningful tokens. Similarly, Yarchi et al. (2021) go to considerable lengths to tokenize Hebrew social media text, segmenting words to separate meaningful lemmas from affixed conjunctions, prepositions and articles, while simultaneously concatenating common construct states, and harmonizing acronyms. In the same vein, Ruzza et al. (2000) constitute named entities as single tokens so as to aid their study of referenced information sources in Italian.

4.1.2 Lemmatization

While also lemmatization serves to enable an algorithm to recognize equivalent or related meanings despite the presence of agglutinations or inflections, lemmatization invariably does so at the expense of throwing away differentiation – be that grammatical case, tense, gender *et cetera*, or in more far-reaching lemmatization, the distinction between different words derived from the same root. To reduce all variants of a word to its dictionary form, lemmatization removes any inflections added to a word or word root to encode its grammatical role, case, number, gender, tense, modality, or other qualities required by a language. In its simplest form, it involves the removal of attached prefixes (e.g., the German "ge-" for past participle), suffixes (e.g., Russian case markers, such as "-ого" or "-у") or infixes (e.g., the Arabic reflexive marker "-ت-", or the Spanish diminutive infix "-it-"). In English, which hardly uses morphological encoding, the effect of most lemmatization is limited: verbs lose tense information, as 'going', 'gone' and 'went' are reduced to 'go'; third-person '-s' suffices and genitive ''s' are removed, but most tokens remain unaffected by stemming and lemmatization. However, this is not true for many other languages. Where in English, "dog" and "man" remain unchanged regardless of who bites whom, case inflections differentiate both possibilities in Ukrainian ("людин**а** кусає собак**у**" ["man bites dog"]; "собак**а** кусає людин**у**" ["dog bites man"]), constituting independent tokens for the analysis unless lemmatization is applied to inform the computer of their identity. In languages such as French or Hebrew, lemmatization is not limited to verbs, which involves removing information about tense, gender, number and modality; but also, nouns and adjectives contain number and gender information that can be stripped, to the point that even numerals change if gender information is removed. Korean additionally encodes important information about the difference in social status into the verb (and infrequently, the noun) – information that is lost by grouping 가요 ["go"-informal] and 가세요 ["go" -polite] into 가다 ["go"-neutral]. In Hungarian, lemmatization removes almost the entire grammatical structure of the sentence, which is encoded via inflections. By contrast, in Chinese, lemmatization does not even remove tense, as tense is expressed by separate words.

Depending on the trade-off between the need for harmonizing text to its lemmas and the need to retain encoded information, different lemmatization procedures may be appropriate – especially if commensurable information is available but merely encoded differently in different languages (Bergmanis & Goldwater, 2018). Unfortunately, lemmatization tools remain unreliable in many languages, and different approaches may yield varying results even within one language (Denny & Spirling, 2018).

Where adequate language resources are available, lemmatization can play a key role for reducing the dimensions of the text and improving performance. For instance, Garcia and Gerton (2021) used lemmatization to gain additional traction for their analysis of COVID-19-related sentiments on Twitter. Similarly, Reber (2019) employed lemmatization to boost the robustness of machine translation, redacting morphological variation to focus his subsequent topic model analysis on thematic variation only. By contrast, Nielsen (2019) did not lemmatize, with the effect that some of his topic models of Arabic text are defined by redundant tokens that add little semantic information. In the same vein, van Paridon and Thompson's (2021) analysis of word

Al-Rfou et al., 2019; Shao et al., 2018). However, Nayak et al. (2020) demonstrate that also transformer models such as BERT are not resistant to tokenization issues (see below).

embeddings in 25 languages shows major performance losses especially for morphologically richer languages (e.g., Finnish).

4.1.3 Filtering

In most computational text analyses, there are significant parts of the text that are known to be uninformative toward the intended analysis and can thus be safely discarded (e.g., stylistic and grammatical information tends to be irrelevant for thematic analyses; Hotho et. Al, 2005). While classic filtering steps such as the removal of boilerplate content or topically irrelevant material are relatively language-independent, this is not true for relative pruning (the removal of the most and least common tokens in a corpus, as neither words that occur everywhere, nor words that hardly occur at all, tend to add much to a statistical analysis; Maier et al., 2021) and the removal of stopwords (specific common words that can usually be ignored; Saif et al., 2014). For many languages, ready-to-use stopword lists are available, typically including articles, pronouns, prepositions, conjunctions, and a range of language-specific function words (e.g., *spaCy*; Honnibal & Montani, 2017). That said, what words are uninformative toward an intended analysis may differ from case to case (Nothman et al., 2018), raising questions about the viability of generalized stopword lists.

Both stopword lists and pruning react strongly to linguistic differences: In languages like English, which express most grammatical functions using separate function words, stopword lists tend to be long, while many of the included tokens do not even exist as separate words in other languages. Considering the most common words in each language, the same percentage of top words that is needed to capture most function words in English reaches deep into the realm of substantively informative words in more synthetic or morphologically rich languages. In the same vein, redacting the bottom few percent of least frequent terms removes much more information in English than it would, for instance, in German or Russian text, which use numerous, highly specific, agglutinated expressions with very low frequencies. Inversely, setting a fixed minimum frequency threshold removes many more words in agglutinative languages than in analytic ones (e.g., "Krankenversicherungsschein" has a presence of only three tokens per million in the DeReKo Corpus, while "health", "insurance" and "certificate" account for 246, 71, and 29 tokens per million respectively in the British National Corpus). The same biases also pertain to any algorithms that weight tokens based on their relative frequency (e.g., tf-idf).

More consequential still, much of the grammatical information removed by stopword lists and pruning in English cannot be removed in this way in many other languages, as it is encoded by morphology, which requires lemmatization.[6] To redact equivalent information from text in different languages, it may be appropriate to remove numerous stopwords in one language, while applying lemmatization to another one. Especially if adequate lemmatization tools are absent for some studied languages, removing stopwords may in fact increase differences between the textual representations in different languages.

A neat demonstration of such concerns is offered by Greene et al. (2018), whose study revealed major differences in the effects of seemingly equivalent preprocessing steps in different languages. Similarly, Ruedin (2013) showed how important carefully adapted stopword lists are for achieving equivalent results in his automatic classification of manifesto texts in German and French. By contrast, relative pruning approaches such as removing the N most frequent words are liable to redact substantively different information in different languages (Bruinsma & Gemenis, 2019).

4.1.4 Augmentation

Unlike the above preprocessing steps, augmentation does not remove information from the text, but rather uses different strategies to disambiguate and add distinctions that are necessary to ensure that identical tokens carry identical information for the analysis. While homographs are comparatively rare in English, this is far from true for Semitic and many East Asian languages – especially those using abjad or logographic scripts, which

[6] In languages where also function words are conjugated, many more tokens may need to be removed, some of which occur with only intermediate frequency (Alajmi et al., 2012).

require considerable efforts at disambiguation.[7] To disambiguate homographs, algorithms typically rely on available context and an analysis of grammatical roles (which, in turn, requires reliable part-of-speech recognition tools that are unavailable for many languages; Torres, 2019).

Other augmentation strategies may involve adding grammar or part-of-speech tags to help focus an intended analysis (e.g., to distinguish between uses of the same word in different grammatical roles; Maimaiti et al., 2021). Inversely to the removal of grammatical variation by means of lemmatization, it is often possible to retroactively enter such information into an analysis by augmenting morphologically poorer languages to match the detail encoded in other languages (e.g., "hand[noun][plural][ablative-instrumental]" encodes the same information as the Latin "manibus"); however, such augmentation strategies fail when required information is not available at all (e.g., we can only know whether an "insurance agent" is female if some nearby pronoun, name or context reveals this information).

Taken together, preprocessing strategies not only serve to redact any information deemed uninformative for an intended analysis, they also play an important role for ensuring equivalence in the representation of textual information. Beyond informing analytic algorithms which contents are to be treated as identical or independent, preprocessing may also mitigate biases in performance that arise from relevant additional information (e.g., gender, case) that is encoded only in some languages, and ensure that textual meaning is represented in similarly-sized chunks, focusing subsequent analyses on a similar level of abstraction.

4.2 Analytic Algorithms

Analytic algorithms comprise any computational procedures aimed to model, recognize and classify intended textual meanings or extract meaningful patterns for the purpose of conducting some theoretically informed analysis. Such algorithms may be applied to either raw or preprocessed textual data, but generally assume that presented textual data includes all the information needed to arrive at a meaningful classification, in a form that will be recognized by the computer. Within analytic algorithms, we can broadly distinguish between rule-based, unsupervised, and supervised algorithms. Importantly, while most rule-based algorithms operate directly on the (preprocessed or raw) textual data, both unsupervised and supervised algorithms rely on statistical procedures that require a transformation of the textual into a numerical representation of the data. In the following, we will review these algorithms in turn, focusing on their being affected by common linguistic differences, and summarize key points in table 1 below.

4.2.1 Rule-based algorithms

Generally speaking, rule-based algorithms are profoundly affected by linguistic differences, but without this dependency causing much major concern. This is because for role-based algorithms, classification rules are fully researcher-controlled and fairly transparent: They demand manual tailoring to different linguistic contexts, and facilitate evaluating whether textual meanings are modeled plausibly (e.g., Lind et al., 2019). In their most common variant – dictionaries – explicit qualification rules specify those expressions used to express an intended meaning in a given language: If Russian uses two different words for "blue" ("синий" for dark and "голубой" for light blue), it is fairly plain what adjustments are required to ensure an equivalent classification. Problems still arise for the disambiguation of homographs, which can result in the erroneous detection of irrelevant expressions – a common problem in Semitic and East Asian languages – and the segmentation of multi-word expressions (e.g., consider "white trash", "white trash can", "white trash can be", only two of which would be correctly classified as hate speech); however, most validity issues caused by linguistic differences are relatively easy to spot, even if they may sometimes require complex disambiguation efforts to address (Baden & Stalpouskaya, 2015). Generally, morphologically rich languages require more complex dictionaries: While truncations often help to recognize different inflections of the same word, they

[7] For instance, the Arabic word "قدم" may mean "he introduced", "arrive from", "foot" or other meanings (Farghaly & Shaalan, 2009). In Japanese, the Hanja 明 may denote the noun "vision" or "wisdom" – particularly in the Buddhist sense, but also "brightness"; it can act as name indicator to denote the name of the Chinese Ming dynasty, but also the male name "Akira"; and it can be read in over a dozen pronunciations, including "myō", "mei", "min" and even "yoshi".

are also likely to invite accidental false hits (e.g., "insur*" will find "insured" and "insurance", but also "insurrection"), and it may be necessary to spell out the relevant morphological forms to ensure that all relevant variants are detected. As table 1 indicates, dictionaries are relatively unaffected by word-order.

Things are still more plain (although by no means easier) for rule-based linguistic tools, such as part-of-speech taggers, shallow-, deep- and dependency parsers, which inevitably need to be developed for each language anew (Bender, 2011). For rule-based methods, therefore, the main challenge raised by linguistic differences concerns the comparative validation of captured meanings (Lind et al, 2019; Baden et al., 2022): Simple translations of dictionaries and rule sets are almost certain to yield non-comparable results, owing to omitted alternative spellings or expressions, unresolved homographs, and many other incommensurabilities. While some unsupervised tools can be leveraged to discover additional indicators and rules (e.g., dictionary expansion, word embeddings; e.g., Rapp & Zock, 2009), there is no way around extensive comparative validation.

In this, valuable guidance is offered by Lind et al. (2019), who compared different methods of dictionary construction across seven European languages. For instance, Tenenboim-Weinblatt and Baden (2018) relied on extensive back-and-forth translations and qualitative discourse analysis to validate the equivalent performance of their multilingual dictionary of conflict-related discourse; Maurer & Diehl (2020) used crowd-coding to validate their dictionary for studying Twitter sentiments toward presidential candidates (Maurer & Diehl, 2020). For their sentiment analysis of English and Turkish language tweets, Öztürk and Ayvaz (2018) carefully edited their Turkish dictionary to include all relevant inflections as well as common mistakes. In many cases, additional disambiguation efforts are required to ensure that dictionaries validly distinguish between meanings that are expressed in different ways, or involving homographs, across different languages (Baden & Tenenboim-Weinblatt, 2017). By contrast, simple lists of keywords and names regularly run into problems, as was noted by Segev (2019), whose dictionary of country names excludes references that are ambiguous: Especially if different languages create different ambiguities (e.g., Turkey is a country or a bird in English; 美国 means U.S. or beautiful country in Chinese; Чили refers to Chile or chili peppers in Russian; Ecuador is a country or the equator in Spanish), the dictionary rapidly accumulates consequential biases.

4.2.2 Unsupervised and supervised algorithms

For both unsupervised and supervised algorithms, linguistic differences impact performance and comparative validity at two separate stages: During the modeling of the corpus, and during the stage of pattern extraction (see also Baden et al., 2022).

As both unsupervised and supervised algorithms build upon a statistical representation of textual contents, one key decision within the modeling of available data concerns the transformation of textual into numerical data. One common way to do this is to create a so-called "bag of words" (BOW), wherein a term document matrix is generated to count the frequencies of any unique tokens found in the data. This procedure thus analyzes documents at the level of single tokens as they emerge after preprocessing (Denny & Spirling, 2018), ignoring syntax, word order, and document structure (Kim et al., 2012).

This procedure, however, is considerably more plausible for languages wherein most meanings are indeed encoded within a single token, and word order is relatively uninformative. Analytic languages, which are rich in construct states and other multi-word expressions, suffer notably greater losses in information through the BOW representation than synthetic languages: While the German token "Krankenversicherung" preserves the specific object under consideration, counting the English tokens "health" and "insurance" once each among a large number of words loses this information. As a consequence, BOW representation raises the level of abstraction in the analysis of analytic languages to much broader concepts than in synthetic ones.

To address this limitation, one strategy can be to recognize relevant multi-word expressions (e.g., Tsvetkov & Wintner, 2011) or more generally to include also higher-order n-grams as tokens – i.e., sequences of two, three or more successive words, enabling subsequent analyses to recognize multi-word expressions ("health insurance" will be included as one token; Nakov & Hearst, 2005). At the same time, this strategy often fails for languages with looser word ordering, and generates numerous extremely infrequent tokens in synthetic languages.

Furthermore, the grammatical information retained by higher n-grams differs between languages with different word order: In English (a SVO language), n-grams are likely to capture subject-verb and verb-object relations, but not subject-object relations. In Korean (a SOV language), the same procedure should mostly capture subject-object and object-verb relations, or neither in Latin (a scrambled word order language; see table 1). Another concern that arises in cross-lingually comparative analyses is that sequences of multiple words encode considerably more information in synthetic as opposed to analytic languages.[8]

A recently developed alternative strategy used by some higher-powered language models (BERTs etc.) is to rely on short character sequences, not words, as the primary tokens entered into analysis (e.g., Al-Rfou et. al, 2019; Chen et. al, 2018), entering counts of recurring character sequences (e.g., "heal", "ealt", "alth"), syllables or subwords (e.g., "in", "sur", "ance") into the analysis. While this strategy is capable of recognizing constituent morphemes regardless of whether these appear as separate words or in in agglutinated form, it still faces challenges recognizing variations that rely on infixes (e.g., "סופר", "kr**ä**nklich", see above) and creates numerous artifactual homographs that suggest related roots where none exist (e.g., "**ins**urance", "f**ins**", "ra**ins**uit"). Moreover, comparative validity issues remain, as the information encoded by an alphabetic character sequence is much less rich than the information encoded by a sequence of abjad or logographic characters, again focusing analyses onto very different levels of semantic abstraction.

Unless a suitable tokenization strategy is adopted, especially synthetic languages contain numerous unique tokens, most of which appear very infrequently, resulting in much larger and sparser count matrices than analytic languages. This problem is multiplied where morphologically rich languages are not effectively tokenized or lemmatized, as the information captured in a single column for English can easily spread over dozens of rare, unique tokens generated by different cases, genders, affixed prepositions, and many other variations (Zalmout & Habash, 2017). Similarly, alphabetic and abjad scripts permit much fewer character combinations than logographic ones, which result in larger and sparser matrices. As a consequence, meanings that are easily identified in an alphabetic, morphologically poor, analytic language such as English may in other languages generate data matrices far too sparse and distributed to support confident detection.

Illustrating the challenges that arise from morphological richness even for high-powered classifiers, Otmakhova et al. (2022) showed that BERT requires considerably more layers to reach acceptable performance when processing text in Russian, a morphologically rich language, compared to Korean and English. In the same vein, Mitts (2019) understands to use at least bigrams when analyzing Hebrew text to avoid missing common construct states. Inversely, Chang & Masterson (2019) suggest that LSTM models that completely disregard word order might be useful for bridging between languages with very different word order schemes. Additional insights may be gleaned also from turning De Vries' et al.'s (2018) conclusion (that machine-translated textual data yields "similar" topic models) on its head and instead focus on those systematic differences - with mean correlations as low as .75 - generated by the putatively neutral intervention of translation.

Turning finally toward the supervised or unsupervised classification of textual meanings, what can be found depends quite directly on the composition of the token document matrix constructed before. If tokens were counted at the level of lemmas, unsupervised analyses will reveal persistent patterns in the co-occurrence of broad themes, and supervised analyses will rely primarily on thematic patterns for classification. By contrast, a reliance on more specific, possibly inflected tokens may yield patterns at a much higher level of specificity and enable supervised classification to access much more detailed information, but at the cost of losing any information about which tokens arise from the same lemma and thus express related meanings. In this context, BERTs' reliance on sub-word tokens preserves most relevant variation while also recognizing most shared roots, and may indeed enable classifiers to rely on whatever information appears most relevant; however, this strategy may still be derailed by correlated, artifactually identical tokens, common homographs and infixes.

Both the unsupervised extraction of meaningful patterns and the supervised classification of textual meanings depend on the extent to which information that is indicative of relevant meanings is a) linguistically available in the textual data, b) retained or added during preprocessing, and c) rendered recognizable for the machine through the numerical representation of the textual data. If in an analysis of Arabic text, verbs' gender

[8] Consider, for instance, the German sentence "Krankenversicherungen verweigern Kostenübernahme", which is fully captured in one trigram, or two bigrams – while its English translation, "health insurances refuse to take over costs" requires six bigrams, five trigrams, or one seven-gram.

inflections were retained, unsupervised analyses can discriminate between activities attributed to men and women (e.g., creating gendered topics in an LDA). Likewise, algorithmic classifiers should have a much easier time recognizing valid classes of a gendered practice than if this information is missing. In English, by contrast, gender is unmarked beyond pronouns, barring the same opportunities unless associated pronouns are preserved or verbs are correspondingly augmented.

In machine classification, relevant, retained morphological variation should tend to improve performance, while both irrelevant variation and unmarked differentiation (e.g., homographs, removed meaningful variation) introduces noise, erodes statistical power and may derail intended analyses (see table 1). As a consequence, morphologically rich languages may hold valuable opportunities for supervised classification, as relatively many potentially relevant distinctions can be encoded in retained tokens; yet, they require considerable efforts at removing uninformative variation. In addition, the study of languages using non-alphabetic scripts tends to introduce considerable amounts of ambiguity to the process, for which there is no easy remedy.

Similar arguments apply to analyses that rely on the recognition of grammatical relations and associations between expressions. If an analyzed phenomenon depends on expressed subject-verb relations, low-level n-gram representations of SVO languages should enable confident detection, while the same is likely not true for SOV languages or simple bags of words or short character sequences. Retained case/number/gender inflections may help identify related expressions in morphologically rich languages, while the same capacity may depend on the retention of selected stopwords and word order in English.

At the same time, differences in the availability of relevant information inevitably create comparative validity issues: Depending on what information is available, the same statistical algorithm will identify different patterns, or rely on different sets of indicators for textual classification – with unknown implications for classification performance. While it is generally possible to reduce the information encoded by different languages to the point of achieving roughly equivalent representations, such may not be desirable where non-equivalent information is in fact relevant for an intended analysis. Accordingly, some analyses may work considerably better, or be feasible only in languages that encode the requisite information.

Table 1 below summarizes key points of the above discussions.

Table 1. Linguistic differences as sources of bias in computational text analysis.

Linguistic Difference	**Script** Some scripts have a tendency to create ambiguous character sequences	**Polymorphemes** Some languages concatenate words into long, highly specific words	**Morphology** Some languages encode grammar, gender etc. using inflections	**Word Order** Some languages encode grammar and relatedness using word order
Key issue	**Unique meanings map unto words in different ways:** **Tokens express multiple meanings**	**Multiple tokens express equivalent meaning**	**Different additional relevant information may be available**	**Different information is expressed by proximity**
Possible danger	independent instances treated as related; spurious findings	related instances treated as independent; sparsity issues reduce power; detection failures	removal decreases performances; detection biases	detection biases
Expected implication: ***Rule-based algorithms***	**irrelevant expressions are detected** e.g., a dictionary may over-measure references to "atom" in Arabic (ذرة), which means also "corn"	**relevant variants are missed** e.g., a dictionary may under-measure references to distortion in Russian by failing to consider the inflected variants of искажен(-ие, -ный, -ная, ...)	**enables separate classification** e.g., a dictionary may distinguish between Finland as an actor, an object or a location in Finnish (Suomi, Suomeen, Suomessa), but not in Swedish	no obvious major effect
Expected implications: ***Unsuperv. algorithms***	**unrelated patterns are merged** e.g., a hybrid topic confounding terms related to censoring or examining (both 审查 in Chinese)	**related patterns are separated** e.g., a topic model may separate president from presidente in French to create separate topics for female and male contenders for president	**enables investigation of subtle differences** e.g., word embeddings can show whether references to competence are close or distant from one another for men and women in Czech (kompetentny, kompetentna), but not in German	**obstructs extraction of relatedness** **may focus analysis on different relations** e.g., low-order n-grams may suffice to recognize verb-object relations in Italian (a SVO language), but higher-order n-grams may be needed in Irish (a VSO language), where low-order n-grams should focus on verb-subject relations instead. e.g., low-order n-grams can recognize some grammatical functions in English, relying on adjacent conjunctions, prepositions, etc., while the same information is encoded via morphology in Hungarian
Expected implications: ***Supervised algorithms***	**classifiers may confound valid and irrelevant associations; ambiguous patterns weaken classification** e.g., miss references to peace in Hebrew, due to dominant use of the same word as "hello" (שלום)	**valid indicators may be missed; sparsity depresses performance** e.g., ignore the comparatively rare passive form of feel (felt) in English as indicator for affect	**relevant differences boost performance** e.g., classification of polite vs. incivil discourse is boosted by the availability of a linguistic politeness marker 요 in Korean or the Dutch formal/informal address form "U/Je", but not in English	
Key question	**Does everything that looks the same to the computer mean the same?**	**Does everything that means the same (for my study) look the same to the computer?**	**What linguistic distinctions bear on the analysis & should be retained or added?**	**What relational information is included at what level of n-gram representation?**
Available strategies	augmentation; disambiguation	lemmatization; tokenization	Augmentation; lemmatization; stopword lists	n-gram representation; stopword lists
	Additionally for comparative analyses:			
	Are the ways in which unique tokens capture unique meaning; in which relevant differences are marked; and in which relational information is represented equivalent across studied languages?			

5 Conclusion

In present-day computational text analysis, English language is not only the dominant use case, but also the technological default (Bender, 2011). Almost all development, and a significant part of gathered experiences, focuses on English-language text, imbuing many available tools and algorithms with a deep-seated preconception of how languages encode text (Baden et al., 2021). However, most languages around the world are unlike English in consequential ways, raising important questions about the validity and comparability of findings. In the present article, we have tried to map some of the most pertinent ways in which languages differ, and examine how these differences collide with commonly used algorithms in computational text analysis.

Throughout our review, we have argued that there are at least three ways in which linguistic differences interact with computational tools: Different languages facilitate or complicate the recognition of meaningful, unique units in text; they use varying strategies encode different amounts of additional information (notably, case, gender, number, and tense); and they express relationships between discrete words in different ways. In each respect, English represents the "easy" case, where most lemmas correspond to space-delimited words, little additional information is encoded, and relationships are marked by word order and separate function words. For the study of other languages, and especially of multilingual text, neither of these assumptions can be taken for granted. Rather, researchers are well-advised to consider and expressly model each aspect:

- At what level of the text are relevant meanings expressed, and how can these be recognized as unique tokens?
- What differences in the expression of relevant meanings are valuable for the analysis, and how can these be retained?
- How are relevant relationships marked in the text, and how can this information be rendered available for the analysis?

And additionally, for multilingual analyses:

- What differences exist between languages' expression of relevant information, and how can equivalent representations be derived?

Our inquiry reveals the central role that available preprocessing routines play not only in the focusing of subsequent analyses on relevant textual variation, but also for the creation of suitable working conditions for statistical analyses, and the attainment of cross-lingually comparative performance and validity. Decisions made during the preprocessing stage have important and immediate implications for the modeling and analysis of textual contents (Denny & Spirling, 2018). Given the very scant recognition and often absent methodological justification of applied preprocessing steps in existing work, there appears to be an urgent need for systematic attention to this consequential stage in the preparation of textual data (Baden et al., 2021; Tsarfaty et al., 2013).

As researchers mount sophisticated algorithms upon preprocessed textual data, we argue that linguistic differences are readily recognized in their impact on rule-based algorithms, while their profound influence on unsupervised and supervised tools too often remains unappreciated and woefully undertheorized. Given the considerable variation in information offered by text in different languages, the common assumption that powerful computational tools will somehow compensate for linguistic differences (Devlin et al., 2019) requires a bold, and largely unsubstantiated leap of faith. In light of our discussion, it appears implausible that purportedly language-independent computational tools can dependably attain acceptable levels of comparative validity in multilingual textual analysis. We thus agree with Bender (2011) that there can be no one preprocessing routine, textual representation or analytic algorithm that equally suits all languages – rather, different languages raise different needs for data harmonization, modeling and analysis, for different intended purposes. Depending on the intended analysis, different information may need to be retained and modeled, and this information may be expressed differently in different languages, necessitating the development of carefully justified, language-sensitive strategies. In fact, many languages offer rich information far beyond what is commonly available in English text, and computational text analysts may do well to consider how to render such information useful for analysis, possibly by augmenting information in other languages that is not linguistically encoded (especially considering the much superior availability of NLP tools for augmentation in English). Inversely, blindly relying on identical algorithms to process different languages may result in profoundly different information being retained, even among closely related languages, and cause

consequential biases during analysis (Grimmer & Stewart, 2013). By beginning to chart important sources of bias and systematic error, we hope to illuminate an urgent research agenda for the field of computational social sciences, whose implications reach well into the adjacent disciplines of computational linguistics, digital humanities, and computer science.

In practice, advancing our understanding of linguistic differences and their implications for automated text analysis is especially pressing in the context of Europe's multilingual and densely interconnected public spheres. Both for academic research and applied uses, which range from data journalism to social media analytics and algorithmic decision making, undetected linguistic biases introduce consequential errors that undermine our ability to study, comprehend and shape social, economic and political processes in a setting where limiting oneself to one language sphere alone is often not a viable option (Grin, 1996). Importantly, multilingualism in textual analysis cannot content itself with focusing on only a few, dominant languages, but needs to consider the whole range of linguistic variations that shape reality within and beyond the European Research Area. That said, it is exactly its multilingual diversity that also uniquely positions the European community of researchers to advance beyond the monolingual, English-focused state of the field to effectively address the various challenges that arise from multilingual communication.

As a European research infrastructure project that spans many countries and languages, OPTED is uniquely placed to kick-start the necessary research agenda. Beyond serving as a hub where knowledge and experiences from the study of text across linguistic boundaries can be collected, exchanged, and accumulated (see OPTED Deliverables D6.1 and D6.2), a first concrete step will be to define and empirically describe key linguistic biases in textual analysis. In this work package's final deliverable, D6.4, we will experimentally establish the impact of key differences in linguistic structures upon social scientific computational text analysis, and examine the viability of suitable strategies for countering the detected biases. Drawing upon this knowledge, and much more future work that remains to be done in this vein, OPTED holds the potential to become a key point of departure for the future development of computational tools and methodologies.

That said, this paper is only a first step in this direction. Existing linguistic variation by far exceeds the few broad topics that we could review, and there are entire families of algorithms that we have not even mentioned here. In the present Deliverable, we have focused on common issues in widely studied languages that affect the most popular algorithms and analyses. As we have illuminated issues that are barely considered, let alone studied in present research, there is a pressing need for further research to substantiate our arguments, gauge the real-world impacts of identified issues, develop and test viable solutions to enable valid, comparative analyses. Beyond rendering users of computational tools aware of the potential biases involved in the analysis of (especially non-English) textual materials, we hope to inspire a much-needed methodological discourse, paving the way toward more language-aware computational development.

References

Alajmi, A., Saad, E. M., & Darwish, R. (2012). Toward an Arabic stop-words list generation. *International Journal of Computer Applications*, *46*(8), 8–13. https://doi.org/10.5120/6926-9341

Al-Rfou, R., Choe, D., Constant, N., Guo, M., & Jones, L. (2019). Character-level language modeling with deeper self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 3159-3166. https://doi.org/10.1609/aaai.v33i01.33013159

Attia, M. (2007). Arabic tokenization system. In *Proceedings of the 2007 workshop on "Computational Approaches to Semitic Languages: Common Issues and Resources,* 65–72.

Baden, C., Dolinsky, A., Lind, F., Pipal, C., Schoonvelde, M., Shababo, G., & van der Velden, M. A. C. G. (2022). *Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis.* OPTED Working Paper. https://www.opted.eu/results/

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2021). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods & Measures.* https://doi.org/10.1080/19312458.2021.2015574

Baden, C., & Stalpouskaya, K. (2015). *Common methodological framework: Content Analysis. A mixed-methods strategy for comparatively, diachronically analyzing conflict discourse.* INFOCORE Working Paper 2015/10. https://www.infocore.eu/results/

Baden, C., & Tenenboim-Weinblatt, K. (2017). Convergent news? A longitudinal study of similarity and dissimilarity in the domestic and global coverage of the Israeli-Palestinian conflict. *Journal of Communication, 67*(1), 1-25. https://doi.org/10.1111/jcom.12272

Baden, C., & Tenenboim-Weinblatt, K. (2018). The search for common ground in conflict news research: Comparing the coverage of six current conflicts in domestic and international media over time. *Media, War & Conflict, 11*(1), 22-45. https://doi.org/10.1177/1750635217702071

Bender, E. M. (2011). On achieving and evaluating language-Independence in NLP. *Linguistic Issues in Language Technology, 6*(3), 1–26. https://doi.org/10.33011/lilt.v6i.1239

Bergmanis, T., & Goldwater, S. (2018). Context sensitive neural lemmatization with lematus. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 1391–1400.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning,* 113–120. https://doi.org/10.1145/1143844.1143859

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit. *Digital Journalism, 4*(1), 8-23. https://doi.org/10.1080/21670811.2015.1096598

Bruinsma, B., & Gemenis, K. (2019). Validating Wordscores: The promises and pitfalls of computational text scaling. *Communication Methods and Measures*, 13(3), 212-227. https://doi.org/10.1080/19312458.2019.1594741

Carneiro, H. C., França, F. M., & Lima, P. M. (2015). Multilingual part-of-speech tagging with weightless neural networks. *Neural Networks*, 66, 11–21. https://doi.org/10.1016/j.neunet.2015.02.012

Chan, C. H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., & Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures*, 14(4), 285-305. https://doi.org/10.1080/19312458.2020.1812555

Chang, C., & Masterson, M. (2020). Using Word Order in Political Text Classification with Long Short-term Memory Models. *Political Analysis*, 28(3), 395-411. https://doi.org/10.1017/pan.2019.46

Chen, H., Huang, S., Chiang, D., Dai, X., & Chen, J. (2018). Combining character and word information in neural machine translation using a multi-level attention. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 1284-1293. https://doi.org/10.18653/v1/N18-1116

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. Political Analysis, 26(4), 417-430. https://doi.org/10.1017/pan.2018.26

Dabre, R., Chu, C., & Kunchukuttan, A. (2020). *A survey of multilingual neural machine translation*. ACM Computing Surveys (CSUR), 53(5), 1-38. https://doi.org/10.1145/3406095

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis, 26*(2), 168189. https://doi.org/10.1017/pan.2017.44

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

de Vries, W., Bartelds, M., Nissim, M., & Wieling, M. (2021). Adapting monolingual models: Data can be scarce when language similarity is high. *arXiv:*2105.02855. https://doi.org/10.48550/arXiv.2105.02855

Dundes, A. (1964). Texture, text, and context. *Southern Folklore Quarterly*, 28(4), 251-265.

Eckert, P. (2001). Linguistic variation as social practice. Oxford: Blackwell.

Escartín, C. P. (2014). Chasing the perfect splitter: A comparison of different compound splitting tools. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC14),* 3340–3347.

Farghaly, A. & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, *8*(4), 1–22. https://doi.org/10.1145/1644879.1644881

Fleischacker, S. (1992). *Integrity and moral relativism*. E. J. Brill.

Fogel-Dror, Y., Shenhav, S. R., Sheafer, T., & van Atteveldt, W. (2019). Role-based association of verbs, actions, and sentiments with entities in political discourse. *Communication Methods and Measures*, 13(2), 69-82.

Gao, J., Li, M., Huang, C.-N., & Wu, A. (2005). Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics*, *31*(4), 531–574. https://doi.org/10.1162/089120105775299177

Goldberg, Y., & Elhadad, M. (2013). Word segmentation, unknown-word resolution, and morphological agreement in a Hebrew parsing system. *Computational Linguistics, 39*(1), 121-160. https://doi.org/10.1162/COLI_a_00137

Graham, J., & Haidt, J. (2012). *The moral foundations dictionary.* http://moralfoundations.org

Greene, Z., Ceron, A., Schumacher, G., & Fazekas, Z. (2016, November 1). The nuts and bolts of automated text analysis. Comparing different document pre-processing techniques in four countries. https://doi.org/10.31219/osf.io/ghxj8

Greenberg, Joseph H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*. 26(3), 178-194.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Haspelmath, M., Dryer, M., Gil, D., & Comrie, B. (Eds., 2005) *The World Atlas of Language Structures.* Oxford University Press.

Honnibal, M. & Montani, I. (2017). spaCy 2: Natural language understandings with Bloom embeddings, convolutionary neural networks and incremental parsing. https://sentometrics-research.com/publication/72/

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20*(1): 19-62.

Jain, V. K., Kumar, S., & Fernandes, S. L. (2017). Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. *Journal of Computational Science*, *21*, 316–326. https://doi.org/10.1016/j.jocs.2017.01.010

Jurgens, D., Dimitrov, S. & Ruths, D. (2014). Twitter users #CodeSwitch hashtags! #MoltoImportante #wow. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 51-61.

Khyani, D., Siddhartha, B. S., Niveditha, N. M., & Divya, B. M. (2021). An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Sc/ience and Technology,* 22 (10), 350 - 357.

Kim, H. D., Park, D. H., Lu, Y., & Zhai, C. (2012). Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–10. https://doi.org/10.1002/meet.14504901209

Koktova, E. (2011). *Word-order based grammar*. de Gruyter. https://doi.org/10.1515/9783110803396

Kuhn, M., & Weidemann, D. (2015, Eds.). *Internationalization of the social sciences*. Transcript Verlag. . https://doi.org/10.14361/9783839413074

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2&3), 211-257. https://doi.org/10.1080/01638539809545028

Lind, F., Eberl, J. M., Heidenreich, T., & Boomgaarden, H. G. (2019). When the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, 13, 21.

Maimaiti, M., Liu, Y., Luan, H., Pan, Z., & Sun, M. (2021). Improving data augmentation for low-resource NMT guided by POS-tagging and paraphrase embedding. *Transactions on Asian and Low-Resource Language Information Processing*, 20(6), 1-21. https://doi.org/10.1145/3464427

Maier, D., Niekler, A., Wiedemann, G., & Stoltenberg, D. (2021). How document sampling and vocabulary pruning affect the results of topic models. *Computational Communication Research*, 2(2), 139-152. https://doi.org/10.5117/CCR2020.2.001.MAIE

Maurer, P., & Diehl, T. (2020). What kind of populism? Tone and targets in the Twitter discourse of French and American presidential candidates. European Journal of Communication, 35(5), 453-468. https://doi.org/10.1177/0267323120909288

Mitts, T. (2019). Terrorism and the Rise of Right-Wing Content in Israeli Books. *International Organization*, 73(1), 203-224. https://doi.org/10.1017/S0020818318000383

Moon, T., Awasthy, P., Ni, J., & Florian, R. (2019). Towards lingua franca named entity recognition with BERT. arXiv:1912.01389. https://doi.org/10.48550/arXiv.1912.01389

Nakov, P., & Hearst, M. A. (2005). Search engine statistics beyond the n-gram: Application to noun compound bracketing. *Proceedings of the 9th Conference on Computational Natural Language Learning* (CoNLL-2005), 17-24.

Nayak, A., Timmapathini, H., Ponnalagu, K., & Venkoparao, V. G. (2020, November). Domain adaptation challenges of BERT in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP* (pp. 1-5). https://doi.org/10.18653/v1/2020.insights-1.1

Nielsen, R. A. (2020). Women's authority in patriarchal social movements: the case of female Salafi preachers. American Journal of Political Science, 64(1), 52-66. https://doi.org/10.1111/ajps.12459

Nothman, J., Qin, H., & Yurchak, R. (2018). Stop word lists in free open-source software packages. *Proceedings of Workshop for NLP Open Source Software* (NLP-OSS), 7–12. https://doi.org/10.18653/v1/W18-2502

Öztürk, N., & Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, 35(1), 136-147. https://doi.org/10.1016/j.tele.2017.10.006

Otani, N., Ozaki, S., Zhao, X., Li, Y., St Johns, M., & Levin, L. (2020). Pre-tokenization of multi-word expressions in cross-lingual word embeddings. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 4451–4464. https://doi.org/10.18653/v1/2020.emnlp-main.360

Palmer, F. R. (2001). *Mood and modality* (2nd Ed.). Cambridge University Press.

Park, J., Barash, V., Fink, C., & Cha, M. (2013). Emoticon style: Interpreting differences in emoticons across cultures. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1), 466-475. https://doi.org/10.1609/icwsm.v7i1.14373

Payne, T. E. (2017). Morphological typology. In A. Y. Aikhenvald & R. M. W. Dixon (Eds.), *The cambridge handbook of linguistic typology* (p. 7894). Cambridge University Press. https://doi.org/10.1017/9781316135716.003

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). *Linguistic Inquiry and Word Count: LIWC2015.* Pennebaker Conglomerates.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? Proceedings of *the 57th Annual Meeting of the Association for Computational Linguistics, 4996-5001*. https://doi.org/10.18653/v1/P19-1493

Popping, R. (2000). *Computer-assisted text analysis*. London: Sage Publications. https://doi.org/10.4135/9781849208741

Prettenhofer, P., & Stein, B. (2010). Cross-language text classification using structural correspondence learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* 1118-1127.

Prewitt-Freilino, J. L., Caswell, T. A., & Laakso, E. K. (2012). The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex Roles*, 66(3), 268-281. https://doi.org/10.1007/s11199-011-0083-5

Rapp, R., & Zock, M. (2009). Automatic dictionary expansion using non-parallel corpora. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 317-325). Berlin: Springer.

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods & Measures, 13*(2), 102-125. https://doi.org/10.1080/19312458.2018.155798

Romeo, N. (2009). The grammaticalised use of the Burmese verbs la 'come' and θwà 'go'. *Cross-linguistic Semantics of Tense, Aspect, and Modality*, 148, 131-154.

Ruedin, D. (2013). The role of language in the automatic coding of political texts. Swiss Political Science Review, 19(4), 539-545. https://doi.org/10.1111/spsr.12050

Ruzza, M., Tiozzo, B., Rizzoli, V., Giaretta, M., D'Este, L., & Ravarotto, L. (2020). Food risks on the web: Analysis of the 2017 fipronil alert in the Italian online information sources. Risk Analysis, 40(10), 2071-2092. https://doi.org/10.1111/risa.13533

Sabel, J. & Saito, M. (2005, Eds.). *The free word order phenomenon: Its syntactic sources and diversity*. de Gruyter.

Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC14)*, 810-817.

Schick, T., & Schütze, H. (2020). Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5), 8766-8774.

Schuler, P. (2020). Position Taking or Position Ducking? A Theory of Public Debate in Single-Party Legislatures. *Comparative Political Studies*, 53(9), 1493–1524. https://doi.org/10.1177/0010414018758765

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. *Proceedings of the 16th International Conference on Machine Learning (ICML '99)*, 379–388.

Shao, Y., Hardmeier, C., & Nivre, J. (2018). Universal word segmentation: Implementation and interpretation. *Transactions of the Association for Computational Linguistics*, 6, 421-435.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (pp. 311-331). Routledge.

Somayajula, S. A., Song, L., & Xie, P. (2022). A multi-level optimization framework for end-to-end text augmentation. *Transactions of the Association for Computational Linguistics*, 10, 343-358. https://doi.org/10.1162/tacl_a_00464

Steimel, K., Dakota, D., Chen, Y., & Kübler, S. (2019). Investigating multilingual abusive language detection: A cautionary tale. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP),* 1151–1160. https://doi.org/10.26615/978-954-452-056-4_132

Tenenboim-Weinblatt, K., & Baden, C. (2021). Gendered communication styles in the news: An algorithmic comparative study of conflict coverage. *Communication Research*, 48(2), 233–256. https://doi.org/10.1177/0093650218815383

Torres, F. N. (2019). An experimental review of a supervised method for word sense disambiguation. In B. Nolan & E. Diedrichsen (Eds.), *Linguistic Perspectives on the Construction of Meaning and Knowledge* (pp. 372-386). Cambridge Scholars.

Tsarfaty, R., Seddah, D., Kübler, S., & Nivre, J. (2013). Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, 3(1), 15-22. https://doi.org/10.1162/COLI_a_00133

Tsarfaty, R., Seker, A., Sadde, S., & Klein, S. (2019). What's wrong with Hebrew NLP? And how to make it right. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* https://doi.org/10.18653/v1/D19-3044

Tsvetkov, Y. & Wintner, S. (2011). Identification of multi-word expressions by combining multiple linguistic information sources. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* 836-645.

Segev, E. (2019). From where does the world look flatter? A comparative analysis of foreign coverage in world news. Journalism, 20(7), 924–942. https://doi.org/10.1177/1464884916688292

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures, 12*(2-3), 81-92. https://doi.org/10.1080/19312458.2018.1458084
van Atteveldt, W. & Peng, T.-Q. (2021). *Computational methods for communication science.* Routledge.
van Cuilenburg, J. J., Kleinnijenhuis, J. & de Ridder, A. (1985). Een theorie over evaluative betogen: Naar netwerkanalyse van journalisteke teksten. *Acta Politica*, 20(3), 291-330.
van Paridon, J., & Thompson, B. (2021). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, 53, 629-655.
Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2021). Progress in machine translation. *Engineering*, 18, 143-153. https://doi.org/10.1016/j.eng.2021.03.023
Windsor, L. C., Cupit, J. G., & Windsor, A. J. (2019). Automated content analysis across six languages. *PloS one*, 14(11), e0224425. https://doi.org/10.1371/journal.pone.0224425
Yao, Y., & Lua, K. T. (1998). Splitting-merging model of Chinese word tokenization and segmentation. *Natural Language Engineering*, *4*(4), 309324. https://doi.org/10.1017/S1351324998002058
Yarchi, M., Baden, C. & Kligler-Vilenchik, N.(2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication, 38*(1-2), 98-139. https://doi.org/10.1080/10584609.2020.1785067
Zalmout, N., & Habash, N. (2017). Optimizing tokenization choice for machine translation across multiple target languages. *The Prague Bulletin of Mathematical Linguistics, 108*(1), 257. https://doi.org/10.1515/pralin-2017-0025