5-11-2023

Use of Natural Language Processing Techniques in the Construct and Instrument Development Process

Kai Larsen
University of Colorado Boulder, kai.larsen@colorado.edu

Rajeev Sharma
University of Waikato, rajeev2238@gmail.com

Magno Queiroz
Florida Atlantic University, mqueiroz@fau.edu

Jan Ketil Arnulf
BI Norwegian Business School, jan.k.arnulf@bi.no

Jean-Charles Pillet
Toulouse Business School, jean-charles.pillet@tbs-education.fr

USE OF NATURAL LANGUAGE PROCESSING TECHNIQUES IN THE CONSTRUCT AND INSTRUMENT DEVELOPMENT PROCESS

Research Paper

Kai Larsen, University of Colorado Boulder, United States, kai.larsen@colorado.edu

Rajeev Sharma, University of Waikato, New Zealand, rsharma@waikato.ac.nz

Magno Queiroz, Florida Atlantic University, United States, mqueiroz@fau.edu

Jan Ketil Arnulf, BI Norwegian Business School, Norway, jan.k.arnulf@bi.no

Jean-Charles Pillet, Toulouse Business School, France, jean-charles.pillet@tbs-education.fr

Abstract

The construct and instrument development process relies significantly on human judgment in the initial stages of the process, specifically in developing construct definition statements, and in developing measurement instruments with high content validity. Natural language processing (NLP) techniques can be used to support human judgment and improve the quality of constructs and instruments employed in research. This paper describes the use of NLP techniques in the construct and instrument development process and presents illustrative results from the use of those techniques. We develop an NLP-based algorithm and illustrate its use to assess measurement instruments. The data used to train the algorithm was 37 years of constructs published in premier IS journals during the period of 1980-2016. The empirical illustrations support our premise that the use of NLP techniques can improve the rigor of the process and improve the quality of constructs and instruments employed in research.

Keywords: Natural Language Processing, Textual Similarity, Construct Development, Instrument Development.

1 Introduction

Natural language processing (NLP) is maturing as a technology and a science. Therefore, we think it is time to consider applying this field to the heart of social science theorising: construct validation and measurement scale development. This paper is an attempt at showing how and why this is now possible, arguing that this method could turn into a powerful tool for theory development.

Constructs are the bedrock of theorizing in most social science disciplines, including management, psychology, marketing, and information systems (Whetten, 1989; Suddaby, 2010; MacKenzie et al., 2011; Rivard, 2014; Colquitt et al., 2019; Sumpter et al., 2019). Within the positivist paradigm of research, constructs themselves are considered as conceptual abstractions that cannot be directly observed and, therefore, need to be measured through some observable proxies (Winnie, 1967).

Given the symbiotic relationship between good constructs and good theory, developing high quality constructs and instruments to measure constructs are key elements of the scientific process (Burton-Jones and Lee, 2017). Over time, scholars have evolved a fairly sophisticated multi-step process for developing constructs and measurement instruments that meet the requirements of good science (e.g.,

Churchill Jr, 1979; Nunnally and Bernstein, 1994; MacKenzie et al., 2011). The process fits a design science framework and the concept of a Type V theory of design and action as per Gregor's (2006) taxonomy of theory types, where the objective of the theory is to give explicit prescription for constructing an artifact, in this case a construct and an instrument to measure it (Hevner et al., 2004; Gregor, 2006).

Despite the extensive use of statistical techniques in validating instruments (e.g., Thorndike, 1904; Likert, 1932; Cronbach and Meehl, 1955; MacKenzie et al., 2011), and recent advances in machine learning techniques such as NLP (Lake et al., 2017), many steps in the construct and instrument development process still rely primarily on human judgement. For instance, creating good construct definitions, generating an initial item pool to represent the construct domain, and assessing whether the proposed items capture the domain of the proposed construct are all tasks that rely on human judgment (MacKenzie et al., 2011, Table 1, p. 304).

A common element among those tasks is that they rely on language processing by humans. For instance, assessing whether a construct has been defined in "unambiguous terms", or whether an item is "representative of an aspect of the content domain of the construct" requires human judges to evaluate the texts of the construct definition statement and item statements and make specific judgments regarding content validity (MacKenzie et al., 2011, p. 304). Such judgements make onerous cognitive demands on humans. They require human judges to successively perform a set of demanding cognitive operations where they sort or rate items regarding the construct definitions that they deem to best represent those items. For example, Hoehle and Venkatesh (2015) attempted to employ Hinkin and Tracey's (1999) variance analysis approach to assess the content validity of the items of their application usability construct. However, their respondents found it challenging to complete the task and warned that other individuals may have difficulties completing the task as well. Hoehle and Venkatesh then opted to discard the variance analysis approach and used the slightly less comprehensive card sorting technique proposed by Anderson and Gerbing (1991). These, and many other tasks in the initial stages of the construct and instrument development process rely almost entirely on the processing of text by human experts. This may help explain why inadequate attention is often paid in practice to rigor in the early stages of the of the process (MacKenzie et al., 2011; Podsakoff et al., 2016; Burton-Jones and Lee, 2017).

Given the core role of language processing in the construct and instrument development process, it is pertinent to ask if recent advances in NLP techniques can be employed to improve the quality of the process. NLP-based techniques are already being employed in multiple domains, including expert-quality language translation (Wu et al., 2016) and answering questions (Ferrucci et al., 2013). Researchers have also started to employ many NLP-based techniques in psychometrics and construct measurement (Arnulf et al., 2018a; Arnulf et al., 2018b). However, how those techniques may be employed to improve the construct and instrument development process is a question that has not yet been addressed in the literature.

This paper contributes to the literature on construct and instrument development process by proposing NLP-based techniques to supplement the human judgements involved in the process. In general, those judgments are employed to assess the language-based correspondence between the texts of construct definition statements and the measurement item statements. We begin by providing a brief synopsis of the role of language-based human judgments in the construct and instrument development process. This is followed by a brief synopsis of the potential utility of NLP-based techniques in the process. We then describe NLP-based techniques to supplement human judgment and assist in the process. This is followed by an empirical illustration of the use of those techniques in the process. We conclude with recommendations for how NLP-based techniques can be employed by future researchers to improve the quality of the construct and instrument development process.

2 Overview of the Construct and Instrument Development Process

The construct and instrument development process described by MacKenzie et al. (2011) reflects the current scholarship in the area. Similar processes are described by Lewis et al. (2005), DeVellis (2016), and Gerbing et al. (1988). The early stages of the process described by MacKenzie et al. (2011) focus on developing high quality constructs and instruments, while the rest of the steps are about assessing the validity of constructs and instruments based on the psychometric assessment of respondent data. Scholars have repeatedly articulated the need to focus more attention on the front-end of the process, i.e., providing a clear conceptual definition and developing indicators that adequately represent the construct, to enhance rigor in the process (Wacker, 2004; MacKenzie et al., 2011; Maul, 2017). This is the focus of our study. Specifically, we propose employing NLP techniques to supplement human judgment in the initial stages of the construct and instrument development process. The front-end of the process is concerned primarily with creating and validating textual artifacts, specifically, construct definition statements and measurement item statements. The later steps in the process are concerned primarily with analyzing respondent data to evaluate the psychometric properties of the constructs and instruments.

The first three steps of MacKenzie et al.'s (2011, p. 297) process constitute the semantic phase of the process. The first step is conceptualization, where researchers "develop a conceptual definition of the construct." This is followed by steps to "Generate items to represent the construct domain" and to "Assess the content validity of the items." Specifically, the first three steps of the process are about designing high quality textual stimuli that can generate responses that can be employed to test theories with a high degree of validity. They are essentially about getting the language of the construct definition statements and item statements 'right.' Assessing the quality of the textual stimuli relies on human experts assessing the text of statements against certain criteria. A key 'expertise' or 'knowledge base' that the human judges draw upon in performing this task is their fluency in language. Essentially, making those semantic judgments is a language processing task. Drawing on the above insights, the following sections identify the semantic assessments performed by human judges in Steps 1-3 of the process.

2.1 Step 1

The first step in MacKenzie et al.'s (2011) construct and instrument development process involves developing a conceptual definition of the construct. Constructs are not directly observable and therefore need to be conceptualized, defined, and operationalized by researchers (Winnie, 1967; Suddaby, 2010; Weber, 2021). Researchers are advised to create a construct definition statement before they develop instruments that reflect the definition statement. Yet despite its criticality, "this stage of the construct validation process is the one that is often neglected or dealt with in a superficial manner (e.g., by assuming that labeling or naming the construct is equivalent to defining it). This leads to a significant amount of trouble later in the validation process" (MacKenzie et al., 2011, 298).

One of the primary challenges for researchers at this stage is to ensure construct clarity, which involves an assessment of the construct definition statement (Suddaby, 2010). Suddaby argues that "One of the more commonly cited reasons for rejecting a manuscript ... is that reviewers feel the submission lacks "construct clarity"..." and that "Reviewers are quick to reject a manuscript where the core constructs are weakly defined" (Suddaby, 2010, p. 346). Hence, improving the quality of construct definition statements is an important issue for researchers. However, reviews of the process repeatedly report a particular lack of rigor in practice (e.g., MacKenzie, 2003; Podsakoff et al., 2003; Lewis et al., 2005; Suddaby, 2010; Larsen et al., 2013; Rivard, 2014; Larsen and Bong, 2016).

The literature on construct development advises researchers to identify potential attributes of the construct of interest by reviewing a representative set of definitions, identifying necessary, sufficient and shared attributes, developing a preliminary definition and then iteratively refining the initial definition (MacKenzie et al., 2011). Through a process of reflection and feedback from human experts, researchers evolve a construct definition statement that they feel faithfully represents the concept that

they have in mind. The iterative process through which researchers evaluate and refine the definition statement to close the gap between the textual artifact and the conceptualization in the researcher's mind is critical to scientific rigor and the credibility of research findings.

Once the researcher comes up with a satisfactory construct definition statement, an important judgment that needs to be made is whether it faithfully represents the construct. When a researcher develops a construct definition statement, the objective is to capture the phenomenon of interest in accordance to their previous knowledge and theoretical assumptions, i.e., the conceptual domain of the construct. Therefore, assessing whether a construct definition statement faithfully represents the concept of interest is very challenging because the intent informing the definition statement exists only in the researcher's mind. External experts, even those who may have been involved in the process (Podsakoff et al., 2016) may not be able to make that judgment with any degree of reliability.

However, what can still be evaluated externally are certain properties of the text of the construct definition statement that the researcher develops. Here, the objective is not to judge the construct as true or false, or as right or wrong. Rather, the objective is to externally evaluate the text and structure of the definition statement against various criteria that scholars have developed to judge the quality of construct definition statements. What is being evaluated is the construct definition statement, not the construct. The construct definition statement is evaluated in this part of the process as a textual artifact on its own right. Hence, a poor definition statement doesn't imply that the construct is inadequate or invalid. Rather, it implies that the textual artifact, i.e., construct definition statement, can be evaluated and improved to better meet the criteria for good construct definition statements that have been developed in the literature (Suddaby, 2010; MacKenzie et al., 2011).

The literature has identified a number of properties of good construct definition statements that human experts are expected to assess in evaluating the quality of construct definition statements. One set of properties relates to the language employed in the construct definition statement. For instance, MacKenzie et al. (2003, p. 325) proposes that good construct definition statements should specify the construct's conceptual domain in unambiguous terms. Ambiguity can arise from a number of sources, such as the statement being subject to multiple interpretations (Suddaby, 2010; MacKenzie et al., 2011). Similarly, Kahane (1982) suggests that construct definition statements "should not contain vague, ambiguous, obscure, or figurative language", or be "viciously circular, i.e. contain a grammatical variant of the same term" (p. 240). Further, good construct definition statements need to "Describe the necessary and sufficient attributes/characteristics as narrowly as possible" and in a manner that is "consistent with prior research and that clearly distinguishes it from related constructs" (MacKenzie et al., 2011, Table 1). Kahane (1982, p. 240) too argues that "it should not lead to the inclusion of things that one does not want the definition to refer to, or be too narrow by leading to the exclusion of things".

Another set of properties of good construct definition statements relates to the content of the statements. Specifically, construct definition statements should not include antecedents or outcomes of the focal construct as attributes, i.e., good construct statements should not be causal statements. Causal statements are the domain of theories, which propose relationships between constructs (Whetten, 1989; Weber, 2012). The inclusion of antecedents or outcomes in a construct definition statement could potentially lead to operationalizations that induce spurious correlations between measures of constructs. Specifically, the inclusion of causal terms or phrases, such as 'leads to', 'is influenced by', or 'arises from', could be signals of potentially causal statements.

The above discussion has identified a number of properties of construct definition statements that should be evaluated in Step 1 to assess the definition's suitability as an artifact that can be employed to conduct high quality research. A review of the above properties reveals that they involve semantic judgments to be made by human experts. Importantly, the semantic judgments that need to be evaluated rely on two key skills of the human experts making those judgments, i.e., their language processing ability and domain expertise.

2.2 Step 2

After developing the construct definition statement, researchers need to generate an initial item pool that "fully represents the conceptual domain of the construct" (MacKenzie et al., 2011, p. 304). For this, researchers are encouraged to refer to the theoretical definition of the construct as well as "previous theoretical and empirical research on the focal construct" and related constructs (MacKenzie et al., 2011, p. 304). Researchers are also encouraged to seek human inputs from "experts in the field … [and]... representatives of the population to which the focal construct is expected to generalize" (MacKenzie et al., 2011, p. 304).

The initial item pool thus generated is evaluated against a number of criteria. A key set of criteria against which the item pool is assessed is related to "how the items are written" (MacKenzie et al., 2011, p. 304), for instance "simplicity and precision of text," "avoidance of double-barreled item statements," "inclusion of ambiguous or unfamiliar terms," and "the presence of causal statements." A perusal of the process of evaluating the pool of items against those criteria reveals that this is essentially a language processing task where human experts make semantic judgments.

MacKenzie et al. (2011, p. 304) recommend that assessments of how the items are written follow an iterative cycle. The initial item pool can be assessed for the quality of writing before subjecting it to a content validity assessment (Step 3). Items 'failing' the initial writing quality assessment are reviewed, revised and included in another cycle of assessment. This process is repeated until a satisfactory pool of items is generated.

2.3 Step 3

After developing the construct definition statement (Step 1) and an item pool to represent the construct (Step 2), researchers need to assess the content validity of the items (MacKenzie et al., 2011, p. 304). Step 3 involves assessing content validity based on an independent evaluation of the texts of the items. Specifically, two judgments need to be made in this step, i.e., whether the items are individually "representative of an aspect of the content domain of the construct" and whether they collectively represent "the entire content domain of the construct" (MacKenzie et al., 2011, p. 304).

A number of techniques have been described in the literature to make judgments about content validity. For instance, Davis (1989) employed a card-sort technique in which a panel of judges was provided with multiple construct definition statements and asked to sort a pool of items into the constructs to which they 'belong.' The pool of items provided to the judges included items generated for a number of focal constructs, including perceived usefulness and perceived ease of use. Typically, in these analyses, inter-judge agreement is employed as an indicator of content validity.

Similarly, MacKenzie et al. (2011) recommend a type of content adequacy assessment for assessing content validity. While there are many variants of the process, essentially it provides a set of independent raters with the text of the construct definition statement, the texts of the item statements and asks the raters to provide their assessment of the extent to which each item captures the focal construct (MacKenzie et al., 2011, p. 304). Rater responses are often captured on Likert-type scales. In another variant of the procedure, typically based on Hinkin and Tracey's (1999) content validity approach, raters may also be asked to rate the items on a different but related orbiting construct to evaluate whether the items are capturing unintended constructs too in addition to the focal construct. A perusal of the above construct validity procedures reveals that the key knowledge base involved in making those judgments is the language-related ability of the judges. Essentially, the human judges are making the judgment whether there is semantic correspondence between the text of the construct definition statement and the text of the measurement item statements being evaluated.

However, a number of issues impact researchers' ability to employ human judges for making semantic assessments. For example, prior research finds that the cost of scaling up Hinkin and Tracey's (1999) content validity approach to assess multiple constructs simultaneously can become prohibitive (Hoehle and Venkatesh, 2015). Further, Colquitt et al (2019) employed human raters recruited through Mechanical Turk to perform a content validity assessment task and report that the exercise involved a

significant dollar cost. Such funding is not routinely available to most researchers. Scholars have argued that the 'cost' of applying rigor in these early stages of the construct and instrument development process is very high due to their dependence on human judgements (Podsakoff et al., 2003; MacKenzie et al., 2011; Podsakoff et al., 2016). For example, subject matter experts might be difficult to identify, and even if they are available, the cost of recruiting them for content validity assessments can be prohibitive. This may explain why researchers often recruit students, who are not necessarily subject matter experts, to perform content validity assessments (Rai and Tang, 2010; Hoehle and Venkatesh, 2015; Colquitt et al., 2019).

Summarizing the above discussion, Steps 1-3 of the process are about designing high quality stimuli that can generate findings and theories with a high degree of validity. They are essentially about getting the language of the construct definition statements and item statements 'right.' Assessing the quality of the stimulus relies on human judges assessing the text of statements against certain criteria. A key 'expertise' or 'knowledge base' that human judges draw upon in performing this task is their facility with language. Essentially, it is a language processing task. In contrast, the rest of the steps in MacKenzie et al.'s (2011) process (with the exception of Step 4, Specifying the measurement model), are about assessing various psychometric properties based on an analysis of respondent data. Furthermore, implicit in MacKenzie et al.'s (2011, p. 310-311) Steps 2 and 3 are two other key assessments that need to be made, i.e., assessments of the convergent and discriminant validities of the instrument. MacKenzie et al. formally include this assessment in Step 5 of their process, which involves making those assessments based on data collected from respondents in a pre-test of the instrument. However, in Steps 2 and 3 they also refer to making judgments on whether items 'belong together' and 'do not belong to unintended constructs', which capture the essence of convergent and discriminant validities.

3 Does Language Matter?

This paper rests on an important premise that semantic properties of measurement instruments can materially affect research findings. Recent literature investigates this issue and provides evidence that the language employed in the texts of measurement items can significantly influence research findings. For instance, Sharma et al. (2014) conducted multiple meta-analyses to investigate how the similarity of texts of items for predictor and criterion variables influences the observed correlations between those constructs. Sharma et al. (2014) found that the greater the lexical overlap (i.e., the number of shared words) between items for the predictor and items for the criterion, the greater is the observed correlation between the two constructs.

Similarly, Gefen and Larsen (2017) measured the lexical similarity between instruments employed to measure key constructs in the technology acceptance model (TAM), e.g., perceived usefulness, perceived ease of use, behavioral intentions, and use. Gefen and Larsen (2017) found that a correlation matrix based on the lexical similarity of the texts of the instruments employed to measure the TAM constructs was very similar to those reported in studies empirically testing the TAM model. They also found that results of structural equation models to test the TAM model using empirical data were similar to those obtained by using textual similarity scores. Their findings suggest that the empirical findings in the domain investigated could be replicated by a semantic analysis without the need to collect any empirical data. Gefen and Larsen (2017) conclude that "part of the reason for the phenomenal statistical validity of TAM across contexts may be related to the lexical closeness among the keywords in its measurement items" (p. 727). They further conclude that "the results raise the possibility that a significant portion of variance in survey-based research results from word co-occurrences in the language itself" (p. 727).

The conclusions from Gefen and Larsen's findings and Sharma et al.'s findings are very similar, i.e., that the semantic properties of measurement instruments could significantly influence the observed correlations. Similar conclusions can be drawn from the findings reported by Arnulf et al., (2014), Nimon et al. (2016), McGregor et al. (2017), and Larsen and Bong (2016). One possible implication of the above stream of research is that the empirical findings reported from the use of those instruments were to some extent embedded within the texts of those measurement instruments. Importantly, from

the perspective of the validity of scientific research they could imply that the observed correlations may simply be a textual artifact and may not fully correspond to the empirical reality.

An alternative explanation for findings reported in the above studies could be that the constructs and instruments examined in those studies had not followed the rigorous development process recommended in the literature. That speculation does not hold as the constructs and measurement instruments examined in the above studies are some of the most rigorously developed in their respective literatures. For instance, Gefen and Larsen (2017) examined the TAM instruments from MIS research, Sharma et al. (2014) examined attitude and behavior instruments from the research on the theory of planned behavior/theory of reasoned action, and Arnulf et al. (2014) examined the Multi-factor Leadership Questionnaire (MLQ) instruments from leadership research. It is unlikely that the effect of semantic properties of measurement instruments on empirical findings reported cumulatively in the above literature is a chance finding.

Scholars have argued that the initial stages of the construct and instrument development process are resource intensive and involve a lot of back-and-forth interaction and involvement of expert raters over multiple iterations of the process (MacKenzie et al., 2011). At the same time, current descriptions of the process do not offer any other source of expertise beyond human judgment that can be called upon to improve the quality of constructs and instruments. Taken together, the above findings lead us to conclude that supplementing human judgment to conduct semantic assessments in the initial stages of the construct development process can pay rich dividends in the form of scientifically rigorous construct definitions and measurement instruments. More importantly, they suggest that there is a need to pay greater attention to assessing the semantic properties of measurement instruments so that the findings generated from the use of those instruments meet the demands of scientific rigor.

4 A Proposal for Use of NLP-based Techniques in the Construct and Instrument Development Process

In this paper we propose employing NLP-based techniques to supplement human judgment in the initial stages of MacKenzie et al.'s (2011) construct and instrument development process. NLP refers to a family of computer-based techniques that process and analyze natural language data. These techniques have been successfully employed to perform various complex language processing tasks such as automatic translation from one language to another (e.g., Wu et al., 2016), producing summaries of a large chunk of text (e.g., Hannah et al., 2011), speech recognition to convert speech into text and vice versa (Koehn et al., 2007), sentiment analysis (e.g., Sajja and Akerkar, 2012), opinion mining (e.g., Pang and Lee, 2008), proofreading (e.g., Kontostathis and Pottenger, 2006), copy editing, grammar checking (e.g., Zhang et al., 2008) and automated essay scoring (e.g., Kakkonen et al., 2008). Many of these applications are based on algorithms that get close to 'understanding the meaning' of the text (Landauer, 2007), with recent developments in generative AI performing at expert level, through, for example, passing bar exams (Katz et al., 2023) and the United States Medical Licensing Examination for physicians (Nori et al., 2023).

NLP techniques typically involve the application of various machine learning and statistical inference techniques to a large corpora of text from which they learn and improve the rules that they then employ in performing various tasks. NLP techniques mimic human judgments on various language processing tasks and often perform at levels comparable to or exceeding that of human experts. These techniques have become extremely sophisticated over the past few years and are even embedded in ubiquitously available applications on devices such as smartphones.

Assessments of content validity in Step 3 of the construct and instrument development process involve making a key judgment, i.e., whether an item belongs with a construct or not. Cosine similarity, a simple measure of the similarity of two text segments could be employed to provide a quantitative assessment of the extent to which an item statement overlaps with the semantic space of the construct definition statement. We go beyond cosine similarity and introduce the 'semantic overlap probability' (SOP), designed to be interpretable in much the same way as the Pearson correlation coefficient, which is also

a measure of the closeness or relatedness of two concepts. Importantly, while computing the Pearson correlation requires respondent data, SOP can be estimated based on the texts of the statements only.

The SOP measure can also be employed to assess convergent and discriminant validity of instruments in Step 3. Specifically, convergent validity assesses whether items in an instrument "belong together". High SOP scores between items of an instrument indicates that the items "belong together", and possess high convergent validity. Similarly, SOP can be employed to assess the extent of discriminant validity between instruments of constructs that are purported to be distinct. Low SOP scores between items of instruments representing two distinct constructs indicate that the items "do not belong together", suggesting a high degree of discriminant validity. More details about our use of NLP to create SOP scores are provided below.

5 Method and Results

We employ a Long Short-term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997; Goldberg, 2017, p. 179) based on Keras high-level neural networks API (Chollet, 2015) to implement a recurrent neural network to assess the quality of constructs and instruments. Neural networks are neuron-like networks of processing units that collectively perform complex computations. They are often organized into layers, including an input layer for data imput (e.g., text segments), hidden layers that transform the data into intermediate representations, and an output layer that produces a response (e.g., a semantic similarity score) (Lake et al., 2017). A recurrent neural network is a a class of neural networks typically used to process sequential data in applications such as language translation, speech recognition, and image captioning. The data used to train the algorithm was 37 years of constructs published in *MIS Quarterly* (MISQ)*, Journal of Management Information Systems* (JMIS)*,* and *Information Systems Research* (ISR) during the period of 1980-2016. It was provided to the research team by the Human Behavior Project at the University of Colorado. The set consisted of 4,476 constructs, of which 3,387 constructs had originally been reported with a definition, and a total of 17,838 items. To the best of our knowledge, this is the largest set of definitions and items used for natural language processing in the IS field. For constructs in this period, a dataset consisting of all possible pairs of items was created. For each, a target was constructed representing a 'one' for items from the intended construct and 'zero' for items in different constructs. For definition-item pairs, those pairs that came from the same construct were given a target value of 'one' whereas those from different constructs were labelled as 'zeros'. In both cases, the 'zero' targets were downsampled to create a balanced training sample.

An NLP model was built from this training dataset to integrate and capture the language components that drive the relationship between a construct definition and its individual item statements. This development proceeded through a number of steps including data preparation, feature selection, model building, testing, and prediction stages. Our NLP model returns an estimate of the probability that an item 'belongs' to a construct, which we refer to as the SOP score. The higher the SOP score, the higher the probability that an item is semantically congruent with the construct definition, and *vice versa*.

Table 1 reports the assessment of the ease of use construct employed by Venkatesh et al. (2003). The assessment shows that the SOP between the construct definition statement and the four item statements are 0.56, 0.57, 0.61 and 0.68. The reasonably narrow range of scores suggests that all item statements share a similar level of overlap with the construct definition statement. However, whether this is a 'good' or 'acceptable' range is something that awaits further empirical examination. Similarly, the inter-item SOP scores range from 0.71 to 0.97, suggesting that the items likely "belong together" and indicative of high convergent validity. Further empirical examination is needed to interpret these scores, for instance, whether too high scores (e.g., 0.97 in the above example) are indicative of inadequate/narrow domain coverage and what are lower bounds of the score that would be considered 'acceptable'. The SOP scores can be employed for construct refinement in a manner similar to how Cronbach's alpha is employed in conventional psychometric analysis, i.e., items with 'unacceptable' SOP inter-item scores and/or 'unacceptable' definition-item OP scores are reviewed in order to improve the quality of the instrument.

Construct: Ease of use (Venkatesh et al., 2003) **Definition:** The degree to which using an innovation is perceived as being difficult to use.	**Definition semantic overlap**	**Item sematic overlap**		
		I1	I2	I3
I1: My interaction with the system is clear and understandable.	0.68			
I2: I believe that it is easy to get the system to do what I want it to do.	0.57	0.78		
I3: Overall, I believe that the system is easy to use.	0.56	0.81	0.71	
I4: Learning to operate the system is easy for me.	0.61	0.85	0.79	0.97

Table 1. Semantic overlap assessment of ease of use construct

In contrast to the above example, which does not raise any major questions about the instrument, Table 2 presents an assessment of the perceived behavioral control construct employed by Venkatesh et al. (2003). The pattern of scores in Table 2 reveals a distinct point of contrast. Specifically, the definition-item SOP scores are lower, ranging from 0.29 to 0.50, as compared to a range of 0.56 to 0.68 for the ease of use instrument above. This suggests a need to review the construct validity of the instrument, i.e. whether the items measure the construct they purport to measure. A perusal of the text of the construct definition statement (Definition: "Perceptions of internal and external constraints on behavior and encompasses self-efficacy resource facilitating conditions and technology facilitating conditions") against the text of Item 4 (I4: "Given the resources, opportunities and knowledge it takes to use the system, it would be easy for me to use the system") does suggest that the I4 item statement shares little overlap with the construct definition statement and illustrates the value of the analysis conducted using NLP techniques. A perusal of the definition statement and the item statement also suggests that they may not meet the criteria for good definition and item statements identified by MacKenzie et al. (2011).

Construct: Perceived behavioral control (Venkatesh et al., 2003) **Definition:** Perceptions of internal and external constraints on behavior and encompasses self-efficacy resource facilitating conditions and technology facilitating conditions.	**Definition semantic overlap**	**Item semantic overlap**			
		I1	I2	I3	I4
I1: I have control over using the system.	0.34				
I2: I have the resources necessary to use the system.	0.37	0.81			
I3: I have the knowledge necessary to use the system.	0.38	0.83	0.95		

I4: Given the resources, opportunities and knowledge it takes to use the system, it would be easy for me to use the system.	0.29	0.57	0.55	0.54	
I5: The system is not compatible with other systems I use.	0.50	0.68	0.72	0.72	0.66

Table 2. Semantic overlap assessment of the perceived behavioral control construct

Another example that illustrates the value of the NLP-based algorithm we have developed comes from an assessment of the instrument for the website navigability construct employed by Pavlou and Fygenson (2006). The assessment results presented in Table 3 flag two concerns. One, the definition-item SOP scores for the four items are 0.38, 0.39, 0.42 and 0.51. This may signal a low degree of content validity for some of the items. Two, the inter-item SOP scores, 0.08, 0.86, 0.07, 0.39, 0.16, and 0.38, are all 'low' except for one score of 0.86. This suggests a low degree of content validity for the instrument. This is further corroborated through a careful read of Pavlou and Fygenson (2006). While they list the four items shown in Table 3 as part of the construct, Pavlou and Fygenson (2006) only retained two of the items as indicated in their factor analysis results (p. 142).

We would like to clarify here that we have employed the above examples only to illustrate the use of NLP-based techniques and their ability to flag issues with constructs and items, rather than to critique any specific study. We might also note here that our initial assessments of a large number of constructs and instruments based only on the definition-item and inter-item SOP scores corroborates the suspicion voiced by MacKenzie et al. (2011) that the quality of our constructs and instruments can be improved by adding greater rigor to the process. The analysis presented here serves to illustrate the contribution that an NLP-based tool to supplement human judgments can make to the construct and instrument development process.

Name: Website navigability (Pavlou and Fygenson, 2006) **Definition:** The natural sequencing of web pages, well-organized layout, and consistency of navigation protocols.	**Definition semantic overlap**	**Item semantic overlap**		
		I1	I2	I3
I1: I expected the sequencing of hyperlinks in this website to be clear.	0.38			
I2: Having a clear sequence of hyperlinks would make it: much more difficult-easier for me to get information about this product.	0.51	0.08		
I3: I expect the layout of this website to be intuitive.	0.42	0.86	0.07	
I4: A website with an intuitive layout would make it: much more difficult-easier for me to get information about this product.	0.39	0.39	0.16	0.38

Table 3. Semantic overlap assessment of the website navigability construct

6 Discussion and Conclusion

The validation of constructs and measurement instruments is still heavily reliant on modelling respondent data (Borsboom, 2008). The prevalent use of Likert-type scales in measurement practice has been included in the standard methodology of construct validation along with all other types of measurement (Bagozzi, 2011). The Likert-type scales now belong to a group of methods where the predominant source of information on their measurement properties is respondent data. Extant literature on construct and instrument development contains recommendations for how to develop such measurement instruments (MacKenzie et al., 2011). However, there is still an obvious gap in sophistication between the techniques applied to the respondent data (e.g., factor analysis) compared to the relative manual and crude work suggested for developing the measurement items themselves. By today's standards, the field does not know any statistical techniques or other inter-subjective approaches to determine the properties of items, scales and multi-scale properties in absence of respondent data. In other words, there is a clear discrepancy between the sophisticated computational techniques available *after* the collection of data and the crude, manual work performed *prior* to administering the measurement scales to subjects. NLP now allows computational procedures and precision to be applied on both sides of this time scale divide. This is where we believe that our NLP-based semantic overlap approach fills a need and a gap.

The field of psychometrics has advanced with new technologies, be that in statistics, computation hardware, computation software or data harvesting tools. The NLP-based techniques we refer to in this study allow assessments of multiple quantitative features of items, measurement scales and construct definition statements prior to or independently of respondent data analysis. We suggest that this comes equal to introducing a sort of operationalism on the side of construct definition and scale exploration and scale construction. This would imply an important step forwards in methodology from the present reliance on previous traditions.

Moreover, our need for adopting machine learning in the construct and instrument development process stems from the inherent frailty in the human semantic validation process. Somewhat paradoxically, our hesitance and resistance in using machines to that purpose may stem from the same problem, i.e., human cognitive constraints in determining the meanings of words and sentences. Despite decades of research in cognitive psychology, humans remain "competent without comprehension" in linguistic skills: We know how to speak but cannot really explain how we do it (Dennett, 2018). This leaves us with three methodological challenges:

1. Human error: Using human judges in determining semantic overlap, similarities, and discriminant boundaries, we can only rely on approximate assessments of rater agreement – itself a type of psychometric approach. We have no independent way of assessing the accuracy and biases of the judges and the processes they use.
2. Machine error: When digital tools are put to the same use, we have no previously established criteria for assessing the technology's performance. Issues of trust and agreement in the resulting statistics may arise.
3. Philosophical error: As long as we do not develop standards for points 1 and 2 above, we will struggle to determine if the construct we are looking for has an existence independently of its semantic construction. This problem is discussed below.

The meaning of words and utterances is not fixed. Languages develop continuously, allowing the meanings of words to change between groups of speakers across space and time (Gumperz, 1996). Studies have found that linguistic habits are sufficiently different between groups of people based on their professions to skew semantic similarity scores on work-related measurement scales (Arnulf et al., 2020).

Our lack of meta-linguistic skills has been a challenge to the use of NLP since its inception. Should machines try to assimilate the way we think humans use language, or should NLP-based techniques shortcut human syntax and dictionaries by allowing computers to do what they are good at doing? The first approach is possible by using lexical databases such as WordNet (Poli et al., 2010) and syntactic

rules (Mohler and Mihalcea, 2009). Examples of the second approach is to use a "bag-of-word-approach" such as Latent Semantic Analysis (LSA) (Landauer et al., 1998). Landauer and Dumais (1997) argue that LSA (a key NLP-based technique) actually comes very close to simulating the way humans learn the meaning of words through exposure to vast amounts of language. Previous studies show that such techniques can predict the correlations between items in survey responses fairly well, with LSA allowing predictions in excess of 80% of explained variance (Arnulf et al., 2014; Arnulf and Larsen, 2020). This research stream demonstrates that NLP can predict human semantic processes involved in human survey responses.

Moreover, humans seem to have a limited memory and attention span that may be the cause why humans find the Hinkin & Tracy procedure cumbersome and difficult (Hoehle and Venkatesh, 2015). The number of items that humans can possibly compare is usually restricted to three or four measurement scales at a time. To the NLP approaches, in contrast, the number of scales is in itself no practical limitation as all items will be compared against all other items with the same systematic rigor. One practical application of this possibility can be found in the so-called "Semantic scale network" which is publically available and published as a psychometric study (Rosenbusch et al., 2020). This approach allows a direct comparison of the item overlaps between all scales that are uploaded into the network. The benefit of this possibility is to compare scale properties across domains of which the researchers are not even aware. This may help alleviate the construct identity fallacy—also known as the jingle/jangle problem (Larsen and Bong, 2016).

Through NLP approaches, the following types of computations are available for statistical modelling prior to collecting and analysing respondent data:

- *Semantic overlap and redundancy:* Quantitative assessments of likely redundancies in measurement items within scales and overlaps with other scales.
- *Coherence:* Quantitative estimates of the semantic relationships among all items within a scale, allowing an assessment of the likely coherence of the items within the scale.
- *Construct adherence:* Quantitative assessment of the strength of semantic relationship between items, scales and their purported construct definitions. It is possible to assess each item for its representativity of the construct definition, and a group of items for their saturation of the item definition domain.
- *Research field positioning:* It is possible to estimate the a priori relationships of a scale to all other known scales (e.g., when a new measurement scale is proposed for an existing construct), or a construct definition to all other construct definitions with a literature corpus (e.g., when a new construct definition is developed). This is a type of assessment that is rarely considered by researchers, who are usually content to consider relationships among the scales included in the study in question. Such assessments could minimize the proliferation of constructs that overlap and are potentially redundant within a research stream.

Hence, we are proposing that NLP techniques could be used in future research for assessing the *semantic network* of measurement scales and construct definitions, against which the *nomological network* of construct validation can be subsequently matched empirically. NLP assessments of semantic structures can take many forms and types, and the quantitative estimates will vary. However, common to all NLP procedures is that the input values and algorithms are transparent and will yield the same outputs to all participants. This allows an intersubjectively valid methodology for exploring the textual side of measurement: construct definitions, measurement scale operationalizations, and item relationships.

While we do not yet possess standardized metrics to this use, we believe that adoption of NLP techniques will help move this field forward. As we briefly mentioned in our methods section, there is already a variety of machine learning approaches to choose from (Arnulf et al., 2021). We could also conceivably develop more sophisticated statistical approaches to text modeling once this field generates more traction. For this reason, we think that the findings we present in the present study are but a promising invitation for the field to join in refining our understanding and toolkits.

In conclusion, this study proposes a semantic-based approach to construct and instrument development process, as well as to the semantic overlap between construct definitions and measurement items. Importantly, we show that the semantic-based view developed here can be operationalized using recent advances in NLP techniques. This is an important issue because standards in terms of construct validation procedures and expectations for the application of those standards have increased over time, thereby putting an increasingly greater burden on researchers tasked with the development of constructs and instruments. Hence, the use of supplementary techniques based on NLP could help alleviate this burden. NLP techniques can supplement human judgments required in various stages of the construct development process and reduce the need for human judgement in the process. Just as developments in statistics and psychometrics underpinned significant advances in research and theory building and testing in the social sciences, we see developments in NLP as supplementing the current repertoire of psychometric techniques and contributing to further developments in research and theorizing.

References

Anderson, J. C. and D. W. Gerbing (1991). "Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities." *Journal of Applied Psychology* 76(5), 732-740.

Arnulf, J. K. and K. R. Larsen (2020). "Culture blind leadership research: How semantically determined survey data may fail to detect cultural differences." *Frontiers in Psychology* 11(176).

Arnulf, J. K., K. R. Larsen and Ø. L. Martinsen (2018a). "Respondent robotics: Simulating responses to likert-scale survey items." *Sage Open* 8(1), 1-18.

Arnulf, J. K., K. R. Larsen, Ø. L. Martinsen and C. H. Bong (2014). "Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour." *PloS One* 9(9), e106361.

Arnulf, J. K., K. R. Larsen, Ø. L. Martinsen and T. Egeland (2018b). "The failing measurement of attitudes: How semantic determinants of individual survey responses come to replace measures of attitude strength." *Behavior Research Methods*, 1-21.

Arnulf, J. K., K. R. Larsen, Ø. L. Martinsen and K. F. Nimon (2021). "Editorial: Semantic algorithms in the assessment of attitudes and personality." *Frontiers in Psychology* 12(3046).

Arnulf, J. K., K. Nimon, K. R. Larsen, C. V. Hovland and M. Arnesen (2020). "The priest, the sex worker, and the CEO: Measuring motivation by job type." *Frontiers in Psychology* 11, 1321.

Bagozzi, R. P. (2011). "Measurement and meaning in information systems and organizational research: Methodological and philosophical foundations." *MIS Quarterly* 35(2), 261-292.

Borsboom, D. (2008). "Latent variable theory." *Measurement* 6(1), 25-53.

Burton-Jones, A. and A. S. Lee (2017). "Thinking about measures and measurement in positivist research: A proposal for refocusing on fundamentals." *Information Systems Research* 28(3), 451-467.

Churchill Jr, G. A. (1979). "A paradigm for developing better measures of marketing constructs." *Journal of Marketing Research*, 64-73.

Colquitt, J. A., T. B. Sabey, J. B. Rodell and E. T. Hill (2019). "Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness." *Journal of Applied Psychology* 104(10), 1243.

Cronbach, L. J. and P. E. Meehl (1955). "Construct validity in psychological tests." *Psychological Bulletin* 52(4), 281-302.

Davis, F. D. (1989). "Perceived usefulness, perceived ease of use, and user acceptance of information technology." *MIS Quarterly* 13(3), 319-340.

Dennett, D. (2018). *From bacteria to bach and back: The evolution of minds*. London: Penguin Books.

DeVellis, R. F. (2016). *Scale development: Theory and applications*. 4th edition. Los Angeles, CA: SAGE publications.

Ferrucci, D., A. Levas, S. Bagchi, D. Gondek and E. T. Mueller (2013). "Watson: Beyond jeopardy!" *Artificial Intelligence* 199, 93-105.

Gefen, D. and K. R. Larsen (2017). "Controlling for lexical closeness in survey research: A demonstration on the technology acceptance model " *Journal of the Association for Information Systems* 18(10), 727-757.
Gerbing, D. W. and J. C. Anderson (1988). "An updated paradigm for scale development incorporating unidimensionality and its assessment." *Journal of Marketing Research*, 186-192.
Goldberg, Y. (2017). "Neural network methods for natural language processing." *Synthesis Lectures on Human Language Technologies* 10(1), 1-309.
Gregor, S. (2006). "The nature of theory in information systems." *MIS Quarterly* 30(3), 611-642.
Gumperz, J. (1996). The linguistic and cultural relativity of inference. In: J. Gumperz and S. C. Levinson (Eds.), *Rethinking Linguistic Relativity*, p. 374-406. Cambridge, UK.: Cambridge University Press.
Hannah, M., T. Geetha and S. Mukherjee (2011). "Automatic extractive text summarization based on fuzzy logic: A sentence oriented approach." *Swarm, Evolutionary, and Memetic Computing*, 530-538.
Hevner, A., S. March, J. Park and S. Ram (2004). "Design science in information systems research." *MIS Quarterly* 28(1), 75-105.
Hinkin, T. R. and J. B. Tracey (1999). "An analysis of variance approach to content validation." *Organizational Research Methods* 2(2), 175-186.
Hochreiter, S. and J. Schmidhuber (1997). "Long short-term memory." *Neural Computation* 9(8), 1735-1780.
Hoehle, H. and V. Venkatesh (2015). "Mobile application usability: Conceptualization and instrument development." *MIS Quarterly* 39(2), 435-472.
Kahane, H. (1982). *Logic and philosophy: A modern introduction*. 4th. Belmont, CA: Wadsworth, Inc.
Kakkonen, T., N. Myller, E. Sutinen and J. Timonen (2008). "Comparison of dimension reduction methods for automated essay grading." *Educational Technology & Society* 11(3), 275-288.
Katz, D. M., M. J. Bommarito, S. Gao and P. Arredondo (2023). "Gpt-4 passes the bar exam." *Available at SSRN: https://ssrn.com/abstract=4389233*.
Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran and R. Zens (2007). "Moses: Open source toolkit for statistical machine translation." In: *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics. 177-180.
Kontostathis, A. and W. M. Pottenger (2006). "A framework for understanding latent semantic indexing (lsi) performance." *Information Processing & Management* 42(1), 56-73.
Lake, B. M., T. D. Ullman, J. B. Tenenbaum and S. J. Gershman (2017). "Building machines that learn and think like people." *Behavioral and Brain Sciences* 40, 1-72.
Landauer, T. K. (2007). LSA as a theory of meaning. In: (Eds.), *Handbook of Latent Semantic Analysis*, p. 3-34. NJ: Larence Erlbaum Associates
Landauer, T. K. and S. T. Dumais (1997). "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological Review* 104(2), 211-240.
Landauer, T. K., P. W. Foltz and D. Laham (1998). "An introduction to latent semantic analysis." *Discourse Processes* 25(2-3), 259-284.
Larsen, K. R. and C. H. Bong (2016). "A tool for addressing construct identity in literature reviews and meta-analyses." *MIS Quarterly* 40(3), 529-551.
Larsen, K. R., Z. A. Voronovich, P. F. Cook and L. W. Pedro (2013). "Addicted to constructs: Science in reverse?" *Addiction* 108(9), 1532-1533.
Lewis, B. R., G. F. Templeton and T. A. Byrd (2005). "A methodology for construct development in MIS research." *European Journal of Information Systems* 14(4), 388-400.
Likert, R. (1932). "A technique for the measurement of attitudes." *Archives of Psychology* 140, 3-55.
MacKenzie, S. B. (2003). "The dangers of poor construct conceptualization." *Journal of the Academy of Marketing Science* 31(3), 323-326.
MacKenzie, S. B., P. M. Podsakoff and N. P. Podsakoff (2011). "Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques." *MIS Quarterly* 35(2), 293-334.

Maul, A. (2017). "Rethinking traditional methods of survey validation." *Measurement: Interdisciplinary Research and Perspectives* 15(2), 51-69.

McGregor, A., R. Sharma, C. Magee, P. Caputi and D. Iverson (2017). "Explaining variations in the findings of presenteeism research: A meta-analytic investigation into the moderating effects of construct operationalizations and chronic health. *Journal of Occupational Health Psychology* 23(4), 584.

Mohler, M. and R. Mihalcea (2009). "Text-to-text semantic similarity for automatic short answer grading." In: *12th Conference European Chapter of the Association for Computational Linguistics (EACL 2009)*. Association for Computational Linguistics. Athens, Greece. 567-575.

Nimon, K., B. Shuck and D. Zigarmi (2016). "Construct overlap between employee engagement and job satisfaction: A function of semantic equivalence?" *Journal of Happiness Studies* 17(3), 1149-1171.

Nori, H., N. King, S. M. McKinney, D. Carignan and E. Horvitz (2023). "Capabilities of gpt-4 on medical challenge problems." *arXiv preprint arXiv:2303.13375*.

Nunnally, J. C. and I. H. Bernstein (1994). *Psychometric Theory*. Third Edition. New York: McGraw-Hill, Inc.

Pang, B. and L. Lee (2008). "Opinion mining and sentiment analysis." *Foundations and Trends in Information Retrieval* 2(1–2), 1-135.

Pavlou, P. A. and M. Fygenson (2006). "Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior." *MIS Quarterly*, 115-143.

Podsakoff, P. M., S. B. MacKenzie, J. Lee and N. P. Podsakoff (2003). "Common method biases in behavioral research: A critical review of the literature and recommended remedies." *Journal of Applied Psychology* 88(5), 879-903.

Podsakoff, P. M., S. B. MacKenzie and N. P. Podsakoff (2016). "Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences." *Organizational Research Methods* 19(2), 159-203.

Poli, R., M. Healy and A. Kameas (2010). Wordnet. In: C. Fellbaum (Eds.), *Theory and applications of ontology: Computer applications*, p. 231-243. New York: Springer.

Rai, A. and X. Tang (2010). "Leveraging IT capabilities and competitive process capabilities for the management of interorganizational relationship portfolios." *Information Systems Research* 21(3), 516-542.

Rivard, S. (2014). "Editor's comments: The ions of theory construction." *MIS quarterly* 38(2), iii-xiii.

Rosenbusch, H., F. Wanders and I. L. Pit (2020). "The semantic scale network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies." *Psychological Methods* 25, 380–392.

Sajja, P. S. and R. Akerkar (2012). "Mining sentiment using conversation ontology." *Advancing Information Management through Semantic Web Concepts and Ontologies*, 302.

Sharma, R., M. Safadi, M. Andrews, P. O. Ogunbona and J. Crawford (2014). "Estimating the magnitude of method bias on account of text similarity using a natural language processing-based technique." In: *International Conference on Information Systems*. Association for Information Systems. Auckland, N.Z.

Suddaby, R. (2010). "Editor's comments: Construct clarity in theories of management and organization." *Academy of Management Review* 35(3), 346-357.

Sumpter, D. M., D. Greenberg and S. Kim (2019). "The dark side of construct convergence: Navigating consensus, evolution, and practical relevance in theory building." *Academy of Management Perspectives* ((In Press)).

Thorndike, E. (1904). *An introduction to the theory of mental and social measurements*: Teachers college, Columbia university.

Venkatesh, V., M. G. Morris, G. B. Davis and F. D. Davis (2003). "User acceptance of information technology: Toward a unified view." *MIS Quarterly* 27(3), 425-478.

Wacker, J. G. (2004). "A theory of formal conceptual definitions: Developing theory-building measurement instruments." *Journal of Operations Management* 22(6), 629-650.

Weber, R. (2012). "Evaluating and developing theories in the information systems discipline." *Journal of the Association for Information Systems* 13(1), 1-30.

Weber, R. (2021). "Constructs and indicators: An ontological analysis." *MIS Quarterly* 45(4), 1644-1678.

Whetten, D. A. (1989). "What constitutes a theoretical contribution?" *Academy of Management Review* 14(4), 490-495.

Winnie, J. A. (1967). "The implicit definition of theoretical terms." *The British Journal for the Philosophy of Science* 18(3), 223-229.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao and K. Macherey (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144*.

Zhang, M., W. Che, G. D. Zhou, A. Aw, C. L. Tan, T. Liu and S. Li (2008). "Semantic role labeling using a grammar-driven convolution tree kernel." *Audio, Speech, and Language Processing, IEEE Transactions on* 16(7), 1315-1329.