Is Twain's 'Awful German Language' really that Awful? Quantifying the Randomness of Linguistic Gender in 19 Languages

By: Ethan Amato

Boston College Psychology Department

Advisors: Joshua Hartshorne, PhD

A Senior Honors Thesis Submitted in Fulfillment of the Arts and Sciences

Honors Program

May, 2022

Abstract

Noun classes are morphosyntactic structures found in languages worldwide that partition nouns into two or more groups. Over the past two centuries, researchers have argued tirelessly over whether a noun's meaning has any role in which of those groups it's assigned to. Some argue that noun class is decided arbitrarily in relation to semantics while others lean towards an intertwining system where noun class assignment is a result of the morphological structure of a word, how it sounds, in *addition* to its meaning. To address a lack of quantitative evaluations on the role of semantics in Noun Class assignment and a literature-wide bias towards evaluating languages from the Indo-European language family, we used machine learning techniques called density-based clustering analysis and a K-Nearest neighbor search on large datasets of numbers that represent the semantics of words to determine whether proximity in semantic space has any bearing on noun class assignment. When excluding noisy data, the noun class assignments of 19 of 19 clustered languages from 3 different language families strongly outperformed 1,000 permutation tests of random noun class assignment. Afterwards, interactive visualizations were published to https://ardianor-ops.github.io/semanticorewebsite/ for public exploration. The K-nearest neighbor analysis revealed that 17 of 19 languages had at least a moderate (Cramer's V ≥ 0.3) effect size between a noun's noun class and the noun class of its nearest neighbor in semantic embedding space but all languages exhibited some degree of noun class-lexical semantics systematicity. Indirectly, the results of this study reflect an innate predisposition for humans to consider semantics when assigning noun classes worldwide and therefore provides further evidence for the Semantic Core Hypothesis which argues for a cognitive tendency for humans to relate syntax and semantics.

Is Twain’s ‘Awful German Language’ really that Awful? Quantifying the Randomness of Linguistic Gender in 19 Languages

Introduction

Published in an incensed 1880 essay about the struggles of learning German, famous American writer Mark Twain makes the following claim: “Every noun has a gender, and there is no sense or system in the distribution; so the gender of each must be learned separately and by heart. There is no other way. To do this one has to have a memory like a memorandum-book. In German, a young lady has no sex, while a turnip has. Think what overwrought reverence that shows for the turnip, and what callous disrespect for the girl.” (Twain, 1880) Humorous as Twain’s argument may be, he brings up a valid concern for the syntactic structure of not only German, but all languages with grammatical gender systems: If it truly is the case that noun class is assigned irrespective of meaning, how in the world do language speakers all around the world remember the genders associated with tens of thousands of distinct concepts? To Twain’s point, such a way of categorizing ‘young lady’ and ‘turnip’ is certainly unexpected and doesn’t seem like a particularly smart way to organize a language. Therefore, gaining a better understanding of the role of semantics in noun class assignment—not only in German but across many languages—will make such a feat of memory a bit more believable. Additionally, cross-linguistic evidence of systematicity between noun class and semantics holds the potential for elucidating a cognitive predisposition to associate syntax (nominal classifiers in this case) and lexical semantics—referred to in the field as the ‘Semantic Core Hypothesis.’

Qualitative surveys of one or a couple languages make up a large portion of the current literature searching for semantics in noun class assignment (Bergen, 1980; Steinmetz, 2006). This typically consists of choosing a handful of nouns from a single language that belong to the same class and assigning some semantic property that loops them together. This works, but at the expense of using an abstract framework of semantics that doesn't always consider what may be relevant for the true motivation of assignment. Other studies argue for a more active role of phonology and morphology in nouns that may also solve the memory issue involved with noun classes (Dinnes, 1971; Erichsen, 2020).

As these qualitative surveys only deal with small numbers of nouns, quantitative studies that evaluate noun classes serve to increase the sample size of nouns drastically while working within a less abstract framework. From these, evidence for cross-linguistic systematicity between gender assignment and semantics has been found before using machine learning, but the article only provides a correlation coefficient of the association strength rather than contributing the domains that the systematicity occurred in (Williams et al., 2019). Other quantitative papers focus on the overlap of assignment across languages and confirmed the role of historical linguistic contact in noun class assignment by computing phylogenetic trees (McCarthy et al., 2020).

Finally, empirical studies of noun class such as artificial language tasks reveal that people have the cognitive capacity to retain memory of gender for nonsense words when semantics played a role in their assignment and developmental studies show that children will learn 'general' classifiers faster than more specific ones in Japanese (Mirkovic et al., 2011; Yamamoto & Keil, 2000). Both of these hint towards the presence of languages where semantics play a role in noun class assignment, but the question of where this systematicity can be found exactly is not entirely clear. Furthermore, these studies can, realistically, only be conducted in one language (real or fake) so generalizing becomes quite difficult.

Our study builds upon the limitations of all three of these approaches: we use word vectors, numerical representations of semantics, as a quantitative operalization of semantics to address the issues of abstraction in qualitative papers. We then employ machine learning algorithms (density-based clustering and K-Nearest Neighbor) that allow us to not only quantify the association strength between noun class and semantics, but lets us build on the current literature by showing where the connection between noun class and semantics is most prominent. Finally, we build upon the experimental studies (although all three approaches suffer from this to some degree) by evaluating the largest number of nouns in such a study to date, but also by evaluating languages from an unprecedented number of language families. Both of these not only improve our study's capacity for generalization, but the latter addresses a general bias in the literature towards evaluating Indo-European languages (Kidd & Garcia, 2021). This disproportionate attention is particularly prevalent in the study of noun classes, as Indo-European gender systems are very rarely placed in conversation with the noun classes from other language families. By including non-Indo European languages in our study, we grant deserved attention to largely underrepresented languages and contribute to correcting the current imbalance in the noun class literature.

Background

We start this literature review with an overview of numerical classifiers to demonstrate that there exists systems of nominal classification that appear to be more motivated by semantics but that even they have cases where it's not entirely clear what determines assignment. We then move on to explaining grammatical gender and Niger-Congo noun classes and what the literature has to say about the linguistic processes involved with assignment. Finally, we provide an overview of the methodologies typically adhered to in the literature when studying noun classes and what each has discovered before detailing our hypotheses.

Types of Nominal Classification and their Function

Numerical Classifiers

While our study focuses primarily on grammatical gender languages, it is necessary to explain numerical classifiers to provide foundational knowledge surrounding other morphosyntactic systems that are more-obviously motivated by semantics. Furthermore, understanding the domains that numerical classifiers are based on may grant insight into where systematicity in gender systems may occur.

Numerical Classifiers are most common in East Asian languages (but are found in Caucasian, West African, and American Indian ones as well) and generally distinguish a relatively large number of classes relative to other nominal classification systems such as grammatical gender (Croft, 1994). In these languages, numerical classifiers are used when counting objects. Take the following example from Japanese:

a. ‘san-bon pen aru’
three-clf.long-thin-object pen are
‘There are three pens’

Note that alongside the number ‘three,’ there is a classifier that is used for a semantically specific type of noun (there are peculiar exceptions to this rule but they are outside the scope of this project). Given a different head noun, the classifier would change as well:

b. ‘san-mai origami aru’
three-clf.flat-object origami paper are
‘There are three sheets of origami paper’

These numerical classifiers that point towards some specific, innate quality of the head noun are referred to as discrete-noun classifiers (Adams & Conklin, 1973). In contrast, mass-count classifiers do not reveal anything about the quality of the head noun. The distinction should become clear when we consider that one could grammatically make the claim of having a barrel of items that are qualitatively quite different (a barrel of cherries or pencils is just as grammatical as a barrel of rum), but to use the discrete-count classifier with anything that doesn't fit within the ‘long-thin object' category would result in confused looks from local speakers of a numerical-classifier language.

Motivations behind Numerical Classifier assignment. As numerical classifier languages tend to have many more noun classes than other systems of nominal classification, it is a lot easier to pinpoint the relevance of semantics in the decision to include or exclude certain nouns (e.g. it makes sense that pen is included in the 'long-thin-object' noun class because it's a long and thin object). In fact, Adams & Conklin (1973) proposed a 'semantic core' of grouping concepts together by animacy, plants, shapes, and function after surveying numerical classification systems from 37 Asian languages. However, this is not to say that noun class assignment is always intuitive in numerical classifier languages. There are apparent subregularities found across languages that make these systems much less simple. For example, we have already discussed that the Japanese classifier—'hon'—is primarily used for long-thin objects (and this generally holds true), but it can also be used to count connected phone calls, things with clear starts and ends, and even points in sports. With these more abstract cases of classifiers, the argument for a consistent, semantics-driven system becomes much fuzzier. Scholars try to maintain such an ideal system: George Lakoff explains these cases as reflective of innate concepts and views them as a blend of human cognitive propensities for things such as metaphorical chaining. To explain, he would reason that the use of the 'long-thin object' classifier for points in sports as a metaphorical extension of the numerical classification from a baseball bat (a long, thin object) to the points that the bat contributes towards (Lakoff, 1983).

Further arguments against numerical classifiers are that they have some degree of redundancy in their function (i.e. why is it necessary to specify that a long-thin object is a long-thin object when counting it?) while others argue that the system serves to provide more salient descriptions that specify the exact translation of the head noun (Greenberg, 2012; Zhang, 2007).

Numerical classifiers are a good starting point to understanding the relationship between semantics and syntax and grant preliminary insight into what types of concepts are grouped together that we can search for in less-clear morphosyntactic systems like grammatical gender. Furthermore, the fact that these more semantically cohesive groups still have some degree of arbitrariness (although theories that maintain semantic cohesion in these cases exist) highlight the importance of learning whether languages constrained to two or three classes total will lean towards one or the other.

Grammatical Gender

Grammatical Gender is an obligatory system that is implicit in the noun regardless of the function of the sentence. Within these systems, there are typically two or three classes that are named after gender categories (masculine, feminine, or neuter) that are typically represented in the morphology of the word. Take the following two examples from Spanish:

c. *'El perro negro'*
ART.masc dog black.masc-ending
'The Black Dog'

d. *'La perra negra'*
ART.feminine dog.feminine black.fem-ending
'The Black (Female) Dog'

Illustrated here are the morphosyntactic effects of grammatical gender in discourse. 'Perro' and 'Perra' both refer to the concept of 'dog' but encode the gender of that dog in their morphology such that nouns with the suffix '-o' are generally masculine and nouns that end in '-a' are typically feminine. Furthermore, we can see these morphological effects extending to postpositive adjective 'negro' and 'negra' who take on the qualities of their head noun. Thus, we can understand that grammatical gender not only affects the noun, but the way a sentence is syntactically formed can rely on the nominal classification of the head noun.

Motivations behind Grammatical Gender assignment. Factors underlying grammatical gender assignment of a noun is where the role of semantics becomes fuzzy: some scholars argue for a semantic motivation for the gender classification of nouns while others take a more anomalist approach by claiming that the gender of a noun hinges on morphological or phonological characteristics. They argue that even in the case where there appears to be a correlation between semantics and gender assignment, there is great speculation as to whether that specific instance can even be called a 'semantic-rule' in the first place (Enger, 2009). For example, in Spanish there is a consistent assignment (that occurs across several dialects) of nouns referring to trees to the 'masculine' category while their fruits are 'feminine' (Bergen, 1980). Bergen also argues for contrasting grammatical gender as semantically motivated by change in size and indicative of natural vs altered states. This evaluation seems to be compatible with Zubin & K-M (1986) which provides evidence for neuter nouns in German acting as superordinate concepts (e.g. vehicle) that encompass basic and subordinate concepts in the feminine and masculine genders (e.g. car and Volkswagen respectively) which is further supported by prior linguistic work by Eleanor Rosch that argues for a hierarchical distribution of concepts based on imageability (Rosch, 1981).

On the morphological and phonological side, Dinnes (1971) and Erichsen (2020) point towards a number of patterns in nouns that native Spanish speakers can rely on to adequately recall a noun's gender such as the construction '-ción' being a reliable morphological indicator of a feminine noun. In fact, Bergen himself used to be a proponent of a phonology-heavy approach to gender assignment and even criticized others for "classifying nouns to an excessive degree on irrelevant semantic and other non-phonemic bases rather than primarily according to phonemic criteria." (Bergen, 1978)

A final account for motivations behind noun class assignment is efficiency. Dye, Milin, Futrell, & Ramscar (2017) provides an information-theory approach to gender assignment in German nouns and makes a case for opposing gender assignment within semantically-similar nouns as a function not of arbitrary assignment but of efficiency in determining head noun ahead of time. For example, in a conversation about heating kitchen appliances, it would be easier to predict the word ‘la tostadora’ (Spanish feminine noun for toaster) over ‘el microondas’ (Spanish masculine noun for microwave) because the speaker uses the feminine article ‘la’ before actually mentioning the concept of toasters. With this in mind, semantics certainly play a role in noun class assignment but rather provides a backdrop for clear demarcation of nouns into different genders.

Therefore, while it would be ideal for there to be a clear answer to what will linguistically predominate the categorization of a word and for there to be a straightforward, semantically-based distribution of gender assignment, current research suggests that—when considered from a cross-linguistic perspective—semantics, morphology, phonology, and even efficiency all interact to produce the noun class assignments we see today.

Niger-Congo Nominal Classifiers

The nominal classification system used in Niger-Congo languages such as Swahili resembles that of grammatical gender with some key deviations. First, they are characterized by a larger number of genders which are generally not relevant to human concepts of biological sex (which is generally why they are referred to as 'Noun Classes' rather than genders). Like grammatical gender, Niger-Congo noun classes appear to act as a dynamic morphosyntactic system. For example, Swahili noun classes are presented in an almost alliterative fashion (note that *ki* refers to noun class 7):

e. *'ki-kapu ki-kubwa ki-moja ki-languka'*
7-basket 7-large 7-one 7-fell
'One large basket fell.'

However, Swahili Noun class prefixes can indicate plurality as some Noun Classes come with a 'plural equivalent:'

f. *'vi-kapu vi-kubwa vi-tatu vi-languka'*
7-basket 7-large 7-three 7-fell
'Three large baskets fell.'

Examples such as these illustrate that it is appropriate to separate grammatical gender and Niger-Congo Noun Class systems as they differ to such a degree that to treat them as identical phenomena undermines their function in the discourse of their respective languages and furthers the aforementioned issue of succumbing to an Indo-European-centric approach to linguistics.

Motivations behind Noun Class assignment. Similar to grammatical gender, semantics appear to play a role in noun class assignment in Niger-Congo languages or at the very least a subset of them (Denny & Creider, 1976). However, pigeonholing a single noun class to a single abstract idea (human, kinship, animal, etc.) is always met with counterexamples, furthering the need for a more thorough approach of analyzing the degree of semantic relevance in noun class assignment (Hepburn-Gray, 2020).

Summary of Semantic Core Classifier Literature

We will now provide an overview of the types of research on noun class and semantics systematicity along with their respective limitations. There are three prominent ways in the literature that researchers will evaluate noun class: First, there are qualitative studies which involve delving into the lexicon of one or a couple languages and discovering sets of nouns assigned to the same gender that also share some semantic property. Next are quantitative studies that allow for more holistic analyses of language that document trends in nominal classification systems. Finally, there are empirical studies that put noun class systems into more practical contexts such as evaluating cognitive biases associated with them.

Qualitative Studies

Historically, the primary approach to identifying a relation between noun class and semantics has been through qualitative analyses of individual or a few languages. Consequently, there appears to be a tradeoff between the amount of languages discussed in a paper and the depth at which each classifier or grammatical system can be explored.

To be brief, Wang (1994) uses numerical classifiers from numerous Chinese dialects to argue for innate distinctions between animacy, shape, consistency, and size. On the grammatical gender side, there are the aforementioned papers from Bergen (1980) and Zubin & K-M (1986).

The next step in diversifying the languages tested for analysis of classifiers would be to test several languages with little linguistic contact and extrapolate innate structures from the consistencies therein. Denny & Creider (1976) looked for underlying semantic systems in Proto-Bantu by comparing field linguist Malcolm Guthrie's notes on the language's noun classes to Toba, Burmese, and Ojibway and concluded that there was consistency despite a lack of historical contact. Corbett (1991) performed a similar study of semantic-based gender systems by surveying a wide array of language families to identify a tendency to distinguish between male human / other (also encompasses animacy/inanimacy), insects, size, and even edibility. Kramer (2020) looked at Akɔɔse, Amharic, and Sochiapan on top of more familiar Indo-European languages to look at the relative importance of animacy, social gender, and 'humanness' in gender assignment. These broader studies, while better than single-language studies and identifying fascinating hierarchies of innate concepts, still only scratch the surface of the breadth of languages in the world and it is for this reason that quantitative studies are particularly useful.

Limitations. Qualitative studies of noun classes, while providing an important foundation of the modern literature, certainly have issues with validity. Instances of apparent semantic cohesion across nouns are often met with counterexamples and it's understandable that researchers are unable to consider the entire lexicon of a language to evaluate whether their claims are air-tight. Fortunately, quantitative studies provide a logical next step to address some of these concerns.

Quantitative Studies

As mentioned previously, by focusing on solely a few languages in the search for a set of concepts that can be considered innate there comes a risk of missing the forest for the trees. To extensively search through the lexicon of and examine the structure of many languages for evidence of these concepts would be far too time consuming and resource-intensive to do within a feasible time-frame. Fortunately, there have been more macroscopic scale studies of nominal classifiers that address the 'linguistic universality' issues of the aforementioned Qualitative studies.

The aforementioned Adams & Conklin (1973) is the most notable and cited of these studies for its efforts to describe the numerical classifier systems of 37 Asian languages that covered Malayo-Polynesia, Austro-Asiatic, Mon-Khmer, Sino-Tibetan, Altaic, Dravidian, and Indo-Aryan language families. Although the paper did not reveal the numbers behind these generalizations, it paved the way for future quantitative endeavors by showing that fascinating co-occurrence of human categorization was possible across a diverse range of languages. Studies that prioritize the presence of particular morphological/grammatical systems include Corbett (2013) which counts and provides a geographical distribution of languages that have gender or classifier systems or even count the number of genders and system of assignment, but do little to analyze them in their entirety. McCarthy, Williams, Liu, Yarowsky, & Cotterell (2020) provided evidence for a relationship between gender classification and concept across their subset of 20 Indo-European languages. However, they (1) evaluated very few language families and (2) assumed consistency in lexical semantics across languages using Swadesh Lists, cross-linguistic databases of relatively stable concepts, rather than word embeddings (Kaplan, 2017). Williams, Cotterell, Wolf-Sonkin, Blasi, & Wallach (2019) studied 18 primarily Indo-European languages and used word embeddings between gendered and non-gendered languages to determine the correlation between lexical semantics and grammatical gender using canonical correlation analysis—a machine-learning algorithm that finds the correlation between two random multivariate variables. Their study, aptly named "Quantifying the Semantic Core," provides evidence for systematic tendency to assign noun class by lexical semantics which may reflect cognitive predispositions, but the paper only used frequent, inanimate nouns in Indo-European Languages with Grammatical gender and the correlation coefficients they uncovered do little to elucidate domain-specific systematicity. In a critique paper of Boroditsky & Schmidt (2000), Foundalis (2002) provided correlational data pointing against cross-linguistic consistency in gender assignment using 14 Indo-European languages but made the assumptions that concepts evaluated were equivalent cross-linguistically and that ‘masculine’ and ‘feminine’ noun classes could be considered stable cross-linguistically as well. Finally, perhaps the most broad-scale evaluation of numerical and nominal classifier systems was Lobben, Bochynska, Tanggaard, & Laeng (2020) which looked at ~330 languages across 51 language families to see if there was a correlation between the presence of particular semantic categories in nominal classifier systems and the recorded specific semantic deficits of 120 neurologically-impaired individuals. They found that there was a statistically significant overlap between certain intermediate semantic categories (akin to 'basic' concepts like 'fruits and vegetables' in Rosch's Superordinate Theory) in classifiers and impairments, which provides evidence for a hierarchical conceptual system like the ones proposed by Mervis & Rosch (1981) and Adams & Conklin (1973).

Limitations. The most prominent drawbacks of quantitative approaches is that there aren't enough of them. Obtaining data for many languages at once for quantitative analysis, while easier now, was simply not feasible in the late 20th century when this topic was particularly popular. That being said, these studies are very much at the mercy of their data and typically focus on Indo-European languages because they are the ones with the most data available. Whereas native or highly-proficient speakers typically run qualitative studies of noun classes, the temptation of studying over 20 languages simultaneously using computers doesn't necessitate proficiency in all or even most of them. As a consequence, qualitative studies are not as constrained to Indo-European languages as quantitative ones (although they still tend to skew that way regardless). Also, in terms of numerical classifiers and Niger-Congo languages, there is a severe lack of documentation on which classifiers or noun classes can be used for which concepts. This further skews the quantitative literature towards gendered languages and exacerbates the already-present bias towards Indo-European languages. Therefore, while generally providing a wider net to cast in the search for semantics in noun class, quantitative studies sacrifice fine-grained details and are particularly constrained by available data.

Experimental Studies

As explored here, qualitative and (to a lesser extent) quantitative analyses of languages with nominal and numerical classifiers mark two primary methods of evaluating the existence of a semantic core. However, evidence for a semantic core through classifiers and grammatical gender has also been shown in the form of experimental studies. Mirkovic, Forrest, & Gaskell (2011) demonstrated in an artificial language paradigm that participants had the capacity to use phonological and distributional features (similar to what is found in grammatical gender languages) of nonsense words to identify noun class and were also capable of utilizing semantic regularities implicit in the stimuli to productively assign noun classes to novel terms. Lobben & D'Ascenzo (2015) was a reaction time study where 25 Mandarin-speaking individuals demonstrated attention bias towards abstract 'graspable object' classifier 'ba' and provided evidence for semantic content contained in some classifiers as perhaps being more readily processed. Yamamoto & Keil (2000) provided evidence that Japanese children acquire classifiers much faster than previously anticipated and had corroborating evidence for a hierarchy of general to more specific classifier acquisition.

Taken together, these studies demonstrate an ability to handle Noun Class assignment by semantics and even point towards a predisposition towards some over others through the lens of numerical classifiers and grammatical gender.

Limitations. Experimental studies are similar to qualitative ones in that they can only really be constrained to one or two languages (real or fake) at a time and can only deal with a relatively small number of nouns. They also have to deal with language speakers rather than language itself which adds another layer of variability into the process.

Background of Utilized Quantitative Tools

The Distributional Hypothesis

Computational methods for representing abstract fields such as semantics generally are founded in the 'distributional hypothesis.' The underlying idea is that words that appear close in context to one another tend to be closer in meaning than those that don't (Harris,1954). Take a conversation about kitchen appliances for example: in such a situation, terms like 'microwaves,' 'toasters,' or 'sinks,' would likely appear in tandem with temperature words like 'hot,' 'warm,' or 'cold' at a rate much higher than African animals such as 'elephant.' Therefore the distributional hypothesis would posit that kitchen appliances are closer to each other and to temperature words in meaning, and that elephants are further away in meaning from both categories.

One benefit to relying on a systematic approach to semantic quantification is that it avoids 'over-abstracting' a set of concepts to fit them together by noun class. For example, Steinmetz (2006) argued for a systematic assignment of 'neuter' to German nouns that are 'functionally hollow' (e.g. wheels, eggs, ears, and yoke) but such a rule shows its cracks quickly when considering that the words for 'stomach,' 'head,' 'throat,' and 'sack' are all masculine despite appearing to be viable candidates for such a semantic rule (Enger, 2009). Put simply, quantifying a space where one can clearly operationalize semantic domains is certainly useful.

Word Embedding Models

Word embedding models such as Word2Vec take large corpora of text and create a high-dimensional embedding space where a single word (referred to as a 'token') can be represented as a series of numbers referred to as a vector (Bojanowski at al., 2017). There are two types of Word2Vec: skip gram and continuous bag of words (CBOW). The primary difference between the two is the method by which the models predict a given word. While being trained on large bodies of corpora, CBOW models calibrate weights between their underlying neural network by using context words to predict a one-hot vector that pertains to a word closest to said context words in high-dimensional space. Either way, the Word2Vec models use the distributional hypothesis and make it possible to create a Semantic Embedding space where words (represented as vectors) closer together in space are also closer in meaning. To measure distance between 2 high-dimensional vectors, typical metrics include Euclidean distance or cosine similarity, which determines semantic closeness by measuring the angle between them.

Cluster Analysis and HDBSCAN

In machine learning, cluster analysis is the methodology of using some algorithm that takes a set of observations and attempts to group them together in ways that are meaningful. It is a method of unsupervised learning, meaning that the data that it categorizes is unlabeled. This makes it perfect for our purposes since we can feed it our high-dimensional Word Vectors, see what words it clusters together based solely on semantics, and compare to the actual noun class labels as evidence for or against noun class-lexical semantics systematicity.

One particular branch of clustering algorithms is called 'density-based clustering.' As a brief baseline, typical clustering algorithms like K-Means means will randomly assign points to act as centroids, iteratively assign other points to a cluster based on their distance from these centroids, and reassign centroids based on the new clusters until the centroids stop moving over iterations. While useful in some cases, these types of algorithms assume that clusters are spherical, which isn't always the case. Furthermore, the number of clusters must be specified by the user a priori either based on a theoretical number of groupings or through running clusterings with k centroids and analyzing the quality of those clusters.

For our purposes, the amount of possible theories of semantic domains that underlie noun class assignment make it quite difficult to even make a guess. Furthermore, due to the high-dimensional nature of the Word Embeddings, it is certainly unreasonable to simply assume that clusters are spherical in shape.

Fortunately, we can circumvent both of these issues by using density-based clustering: Density-based clusters decide on groupings based on adjacent regions of high point density and low point density. Hierarchical DBSCAN (HDBSCAN) is a powerful density-based clustering algorithm which is indifferent to the shape of clusters, does not require the number of clusters to be specified, and is robust with respect to clusters with different density (Campello et al., 2013). Further, HBDSCAN is very attractive because it has only one hyperparameter 'minimum points' which is the minimal number of points in a cluster. It then optimizes on a 'mutual reachability distance' to apply globally that generally encompasses that number of minimum points to determine what is a high or low density region. It is relatively fast for large data sets, detects outlying cells, and for each cell it reports a probability of assignment to a cluster. A further advantage of this algorithm is that, unlike K-means clustering, it doesn't have to assign every data point into a cluster. This means that if there are no cohesive clusters of lexical semantics in our data, it would simply tell us our dataset is too noisy and not assign any clusters (meaning we would fail to reject the Null Hypothesis).

Hypotheses

Hypothesis 1

As we have discussed, the literature appears entirely uncertain about the role of semantics in nominal classification. Even in the case of numerical classifiers, which are supposed to have 'semantic cores,' there are subregularities that call into question the processes involved with noun class assignment. In terms of grammatical gender and Niger-Congo noun classes, the smaller number of available classes further calls the role of semantics into question, especially when assignment can be explained by morphological, phonological, and even functional factors. Despite these counterarguments, linguists continue to identify semantically-motivated trends in noun class assignment which suggests that semantic proximity between two nouns will be a reliable predictor of noun class. With this in mind, we provide the null hypothesis that the effect size between the closeness of two nouns in semantic space and noun class assignment is no different than 0 and the alternative hypothesis that the effect size between semantic proximity and noun class assignment is different than 0.

Hypothesis 2

Given the overview of possible ways that noun class could be distributed in terms of semantics (by generality, animacy, imageability, all of the above etc.), it is particularly necessary to understand what such theories of noun class-semantic systematicity might look like in the quantitative context of a Word Embedding Space.

Researchers have debated for some time whether noun class systems are truly as arbitrary with regards to meaning as Twain claims. One extreme is that these systems are truly arbitrary (Fig. 1, "Completely Random Distribution”). For the Indo-European languages, this is may not the case because, if nothing else, there really is a correlation between noun class and biological sex although it may turn out to be very weak. They are called ‘gender’ systems for this exact reason. However, this may not turn out to be the case for Swahili / Niger-Congo languages where there is no such clear distinction between biological sexes by noun class. On the other side of the spectrum, there lies the possibility of nouns being completely consistent by semantics bar perhaps a couple exceptional cases like where noun class is driven by cross-linguistic borrowing (Barken, 1980).

If neither of these turn out to be the case, there must be some possible explanation for a middle ground. As mentioned previously, sets of seemingly semantically coherent nouns found in the qualitative literature are usually met with counter examples, but is it truly correct to throw out an entire theory due to a few potential outliers? With this in mind, we must account for the chance of there being clusters of a single gender that are *generally* bound to a single semantic constraint, but also include nouns that contradict that systematicity (e.g. Twain’s example of ‘young lady’ being neuter in German). We must fall back on the two extremes in this case: total randomness or total systematicity.

The former would involve nouns of the same noun class being clustered around a central concept but a select few nouns within that cluster are randomly assigned to another class (Fig 1. “The Random Backfill Hypothesis”). While the thought of there being random-assignment within a systematic tendency of assignment is perplexing, Dye et al. (2017)’s functional account may allow for a compelling argument towards the possible use of such a counterintuitive system because, in theory, whichever noun in a semantically-coherent cluster that differs from the norm need not be chosen purposefully to achieve the benefits in efficiency.

The final possibility would be that, like "The Random Backfill Hypothesis," nouns are centered around a central concept but instead of being randomly assigned into other noun classes, they are systematically assigned into semantically-coherent groups (Fig 1. "Systematic Backfill Hypothesis"). We have already run into this in Spanish with the example of trees being male while their fruits are female.

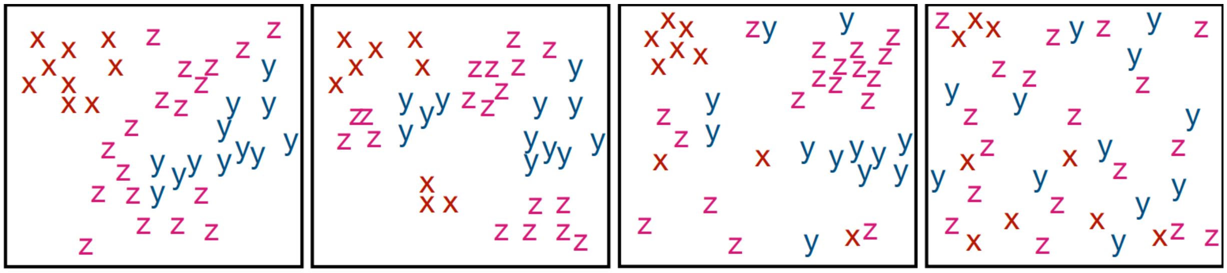


Figure 1: From the left: Semantic Coherence Hypothesis, Systematic Backfill Hypothesis, Random Backfill Hypothesis, and Completely Random Distribution (H0)

How does this project improve on previous research?

Above, we outlined the limitations of the literature that prevent it from clearly distinguishing between the hypotheses. We make use of several recent developments that allow us to go beyond the work to date. First, using density-based clustering allows us to identify whether or not languages can be separated into a set of clusters in the first place purely based on semantics. If they can, we can visualize which areas in semantic embedding space are more prone to systematization which has never been done before.

As density-based clustering involves throwing away a lot of our data, we implement a K-Nearest Neighbor search that is guaranteed to use all of the data to quantify the relationship between noun class assignment and semantic proximity but won't show groupings of semantics like clustering. Therefore, our two analyses account for each others' weaknesses while highlighting their strengths.

We further improve upon the literature by using a wide breadth of languages and language families to adequately determine innate conceptual representations from a token-level. Whereas previous research has generally focused on Indo-European languages, our project aims to broaden the scope of what can be evaluated simultaneously by placing grammatical gender languages and broader nominal classification methods such as Swahili's 7-class system and by including as many non-Indo European languages as there is sufficient public data for. It addresses the issue of 'overabstracting' concepts of the same noun class by using an intuitive and intra-linguistically consistent measure such as the distributional hypothesis to. It serves to build upon Williams, Cotterell, Wolf-Sonkin, Blasi, & Wallach (2019) by replicating their results with a larger, more diverse dataset and identifying which exact tokens were responsible for the relationship between Noun Class and Lexical Systematicity that they discovered. Finally, further exploration of such a tool can be used to support one of the aforementioned theories of innate conceptual structure, cognitive distinctions between concepts (such as animacy, shape, or function), or at the very least systematically and identify a large array of clusters of semantically related terms that share noun classes which will be of use to the linguistic community going forward.

Methodology

Data and Models

The data was retrieved from Wiktionary, an open-source linguistic resource where users can contribute information regarding a language of interest. This resource is admittedly not perfect: while there is a list of criteria that govern a term's inclusion in the site, the sheer magnitude and breadth of the database make this incredibly difficult to enforce. Quantitative analyses by Spitzer & Wolfer (2016) point out concerns of a significant portion of Wiktionary is edited by a relatively few number of contributors and that there is a disproportionate amount of attention (operationalized by number of 'edits' made to a given page) paid to a subset of concepts. However, despite these concerns, Wiktionary remains an essential resource that has contributed greatly to many recent peer-reviewed articles in quantitative linguistics largely due to its open nature and extensive library.

We retrieved token and gender information by using a web scraper that accessed the respective pages for gendered nouns in our 19 included languages. Words that clearly did not follow the orthography of the language being scraped were removed (i.e. words with Chinese characters in the Spanish nouns) and a language was excluded from the study if the sum of the number of nouns across all classes was less than 1,000.

For our Word2Vec models, we used the models produced by Grave et al. (2018) which used large sources of multilingual corpora (Wikipedia and the common crawl project) to produce high-quality 300-dimension word embedding representations for 157 languages. This project made use of the FastText models for Arabic, Bulgarian, Dutch, French, German, Hebrew, Hindi, Icelandic, Italian, Maltese, Polish, Portuguese, Russian, Sanskrit, Serbo-Croatian, Slovak, Spanish, Swahili, and Urdu. Overall this project covered 3 language families: 15 Indo-European, 3 Afroasiatic, and 1 Niger-Congo. It should be briefly noted that when training these models, the more corpora / textual information you can provide the better and therefore less-spoken languages may suffer by virtue of their respective Wikipedia having less data overall.

For each language, we import the nouns into a dataframe with another column that keeps track of their noun class. We then use the corresponding FastText model to retrieve the associated 300-dimension vector for each noun.

To prepare the data, we first removed duplicate terms to prevent the possibility of inflating a single epicene noun/concept's representation in semantic space (e.g.we remove the possibility of a noun's 'next door neighbor' being itself). Next, we scaled the data as is customary for preparing for dimensionality reduction and applied PCA to the data in each language such that the remaining number of dimensions maintained 95% of the original variance. These few steps serve to minimize the influence of outlier dimensions, improve computation time, and to reduce computation complexity all while minimizing information lost. A summary table containing basic information regarding our data and the corpora used to train our models can be found in Table 1 and a map that displays the geographic distribution of where the analyzed languages are officially spoken can be found in Figure 2.

Table 1.
Sample Size of Nouns by Language and Family

Language	Language Family	# of Wikipedia Articles	N Tokens	N Dimensions
Arabic	Afro-Asiatic	1,162,234	11,538	250
Bulgarian	Indo-European	280,446	4,749	238
Dutch	Indo-European	208,602	28,862	251
French	Indo-European	2,410,118	43,686	247
German	Indo-European	2,676,801	42,829	245

Table 1.
Sample Size of Nouns by Language and Family

Language	Language Family	# of Wikipedia Articles	N Tokens	N Dimensions
Hebrew	Afro-Asiatic	313,333	5,914	246
Hindi	Indo-European	153,538	9,238	253
Icelandic	Indo-European	54,121	11,019	252
Italian	Indo-European	1,747,435	66,038	251
Maltese	Afro-Asiatic	4,529	3,803	197
Polish	Indo-European	1,516,037	25,898	245
Portuguese	Indo-European	1,088,035	5,318	232
Russian	Indo-European	180,6335	26,488	239
Sanskrit	Indo-European	11,644	3,490	187
Serbo-Croatian	Indo-European	456,423	30,218	255
Slovak	Indo-European	239,950	3,823	239
Spanish	Indo-European	1,763,372	53,784	244
Swahili	Niger-Congo	70,752	3,854	248
Urdu	Indo-European	169,085	2,472	239

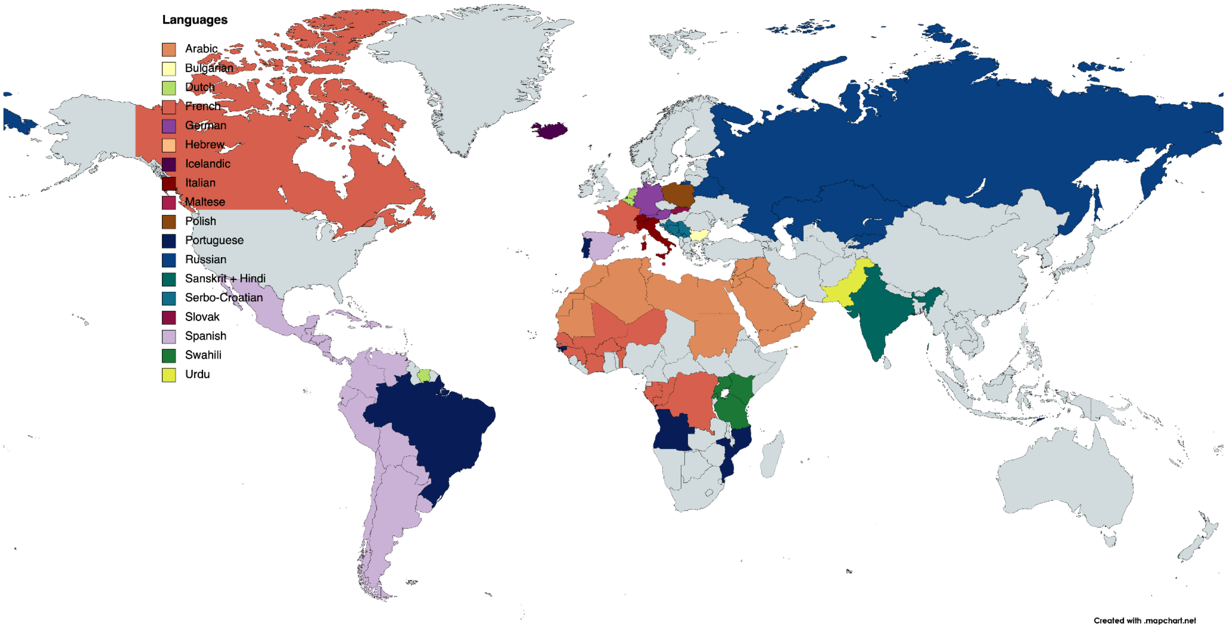


Figure 2: A Map Containing Geographic Distribution of all Languages studied. Inclusion of a country was based on that country labeling a language as an 'Official Language'

HDBSCAN and Cluster Purity Comparison

For each language, we ran HDBSCAN 3 times with minimum cluster sizes of 5, 7, and 9 tokens to prevent the possibility of the algorithm optimizing on a mutual reachability distance too small to fairly capture as many clusters as possible or minimizing all tokens as noise (i.e. we approximate the clusterings the algorithm *usually* gets). Larger minimum cluster sizes should count to decrease cluster purity as a trade off of capturing what the algorithm usually lands on for an optimal mutual reachability distance, so as a sanity check a correlation analysis between cluster size and purity was performed. We used cosine similarity as a similarity metric for this procedure.

We use an intuitive and commonly-used metric for purity taken from Manning et al. (2008): For each cluster defined by HDBSCAN (that isn't labeled as noise), we take the modal noun class and note the amount of nouns that correctly match the mode. Finally, we summate the number of matching noun classes across all clusters and divide by the total number of non-noise data. This can be mathematically formalized as:

$$purity(\Omega, \mathbb{C}) = \frac{1}{N}\sum_{k} max_{j}|\omega_{k} \cap c_{j}|$$

Where $\Omega = \{\omega_1, \omega_2 \ldots \omega_k\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2 ... c_J\}$ is the set of possible Noun Classes, ω_k serves as the set of concepts in cluster k and c_j referring to the noun classes of those concepts.

To understand whether each language's purity score is different from chance, we ran 1,000 permutation tests of randomly-assigned noun classes to the same clusters and performed a Z-test to understand where the real purity score stands in the distribution of purity scores of randomly-assigned languages.

UMAP, Visualization, and Translation

To check for obvious clustering of tokens by semantics, UMAP dimensionality reduction was performed on all data down to 2 dimensions, translated using Google Cloud Translate, and plotted using the interactive Python package Bokeh for public, post-hoc, qualitative analysis.

Nearest Neighbor Search

Since HDBSCAN will recognize any words that have an insufficient number of points within a given mutual reachability distance as noise, a follow up test to evaluate all nouns without a necessity for a minimum distance was performed. For each noun, the concept that functioned as its k=1 nearest neighbor was calculated using cosine similarity and its noun class was recorded. For each language, a confusion matrix, classification model metrics, and a

Chi-Square test with Cramer's V effect size were performed on the results. Cramer's V (or Cramer's Phi) is analogous to Pearson's r in that it measures the effect size between two variables but can measure the association strength of nominal variables in complex contingency tables (e.g. actual noun class vs nearest neighbor noun class). It ranges from 0 to 1 where 0 means no effect and 1 is a perfect correlation. The size of an effect found is also analogous to Pearson's r where .1-.3 is a weak effect, .3 to .5 is a moderate effect, and >.5 is a large effect .

It should be noted that the p value is misleading and that, as sample size increases the probability of a given outcome becomes arbitrarily small. Focusing on the effect size is the best way to interpret our results.

Results

HDBSCAN and Cluster Purity Comparison

First of all, our sanity check of Cluster Quality measured by minimum cluster sizes versus Purity was significant in 14 out of 19 languages. As this was counterintuitive, further analyses were performed by correlating the number of clusters found by HDBSCAN in each iteration and cluster purity which was significant in 19 out of 19 languages (see Table 2). The first set of results indicate that cluster purity increases as cluster size decreases for most languages (which is intuitive). The latter results reveal that in the languages where the first check didn't pan out, the number of clusters HDBSCAN decided on for each iteration of minPts were significantly correlated with purity. Therefore, the initial results can be explained by HDBSCAN occasionally optimizing on a mutual reachability distance that led to a very small number of clusters in a select few languages which led to a corresponding decrease in purity, but there was indeed a connection between cluster quality and purity in all languages.

This confirms that our algorithm follows the common tradeoff of cluster quality and cluster purity and, more importantly, reveals that as regions of semantic space are widened or tightened, there is a corresponding shift in the relative density of that region's modal noun class. Therefore, this assures some degree of internal validity and can be taken as preliminary evidence towards a relation between semantic proximity and noun class assignment.

Table 2.
Cluster information from HDBSCAN Analysis

	Correlation Between Minimum Points and Purity			Correlation Between Number of Clusters and Purity		
Language	Pearson's r	r^2	p	Pearson's r	r^2	p
Arabic	-0.925	0.857	*$p < 0.001$	0.718	0.516	*$p < 0.001$
Bulgarian	-0.927	0.858	*$p < 0.001$	0.634	0.402	*0.003
Dutch	-0.224	0.05	0.343	0.677	0.458	*0.001
French	0.13	0.017	0.585	0.685	0.469	*0.001
German	0.054	0.003	0.82	0.568	0.322	*0.009
Hebrew	-0.85	0.722	*$p < 0.001$	0.889	0.79	*$p < 0.001$
Hindi	-0.883	0.779	*$p < 0.001$	0.746	0.557	*$p < 0.001$
Icelandic	-0.938	0.88	*$p < 0.001$	0.855	0.73	*$p < 0.001$
Italian	-0.609	0.371	*0.004	0.851	0.724	$p < 0.001$
Maltese	-0.854	0.729	*$p < 0.001$	0.573	0.328	*0.008
Polish	-0.119	0.014	0.618	0.472	0.223	*0.036
Portuguese	-0.717	0.514	*$p < 0.001$	0.773	0.597	*$p < 0.001$
Russian	-0.4	0.16	0.081	0.664	0.442	*0.001
Sanskrit	-0.813	0.66	*$p < 0.001$	0.966	0.933	*$p < 0.001$
Serbo-Croatian	-0.958	0.917	*$p < 0.001$	0.786	0.617	*$p < 0.001$
Slovak	-0.465	0.216	*0.039	0.787	0.62	*$p < 0.001$
Spanish	-0.701	0.491	*0.001	0.836	0.699	*$p < 0.001$
Swahili	-0.804	0.646	*0.003	0.949	0.901	*$p < 0.001$
Urdu	-0.565	0.319	*0.009	0.925	0.855	*$p < 0.001$
*: Significant at $\alpha = 0.05$						

As for the clustering themselves, a large degree of the tokens were deemed as noise and a particularly large number of clusters were identified for each language (see Table 3). This can be taken one of two ways: the first explanation is that as you increase the dimensions of a dataset, the volume of the space that dataset depicts becomes exponentially sparse (referred to as the ‘curse of dimensionality’). Therefore HDBSCAN could have simply found it difficult to place some words in clusters because they were too far away from their neighbor and also inflated the number of clusters by being particularly attentive to small regions of semantic space accentuated by their distance from each other.

The second theoretical explanation is in line with Kramer (2020), where they argue for both semantic rules that govern noun class in addition to simple, arbitrary assignment. Under the assumption that there is some relation between noun class and lexical semantics and that these noise points are representative of arbitrary assignment, this appears to be possible evidence for the Random Backfill Hypothesis.

When compared to 1,000 permutation tests of randomly-assigned noun classes using Z-Tests, the purity of the HDBSCAN-identified clusters were statistically significant in 19 out of 19 languages (see Table 3). While it is important to keep in mind that the data for some languages was quite noisy, the statistical significance of the cluster purities allows us to reject the null hypothesis of there being a complete lack of noun class-lexical systematicity and the lack of perfect purities allows us to reject the Semantic Coherence Hypothesis. This is therefore evidence for some type of middle ground such as the Systematic Backfill Hypothesis and the Random Backfill Hypothesis. Unfortunately, given the unexpected amount of clusters identified, it is unfeasible to parse out which of the two actually reflects reality, but the sheer amount of perceived noise that the algorithm excluded (and therefore didn't assign into new, cohesive clusters) appears to weigh more in favor of Random Backfill.

Table 3.
Z Tests between 1,000 Randomly-Assigned Permutation Tests against Real Language Purity

				1,000 Permutation Tests			
Language	# Of Clusters	% of Tokens Deemed as Noise	Real Purity	M Purity	SD Purity	z	p
Arabic	270	79.7%	0.819	0.6652	0.005	30.47	***p <.00001
Bulgarian	150	71.57%	0.7244	0.5612	0.0075	21.75	***p <.00001
Dutch	1018	67.35%	0.7029	0.4921	0.003	70.93	***p <.00001
French	1361	71.06%	0.8208	0.6369	0.0026	71.77	***p <.00001
German	1419	69.84%	0.7672	0.4959	0.0026	104	***p <.00001
Hebrew	151	78.09%	0.8079	0.6417	0.008	20.69	***p <.00001
Hindi	270	76.07%	0.7879	0.6938	0.0046	20.29	***p <.00001
Icelandic	403	66.93%	0.7827	0.5209	0.0046	56.95	***p <.00001
Italian	2120	71.44%	0.8499	0.6332	0.0021	104.79	***p <.00001
Maltese	131	58.11%	0.806	0.6124	0.0075	25.93	***p <.00001
Polish	828	68.92%	0.8455	0.5399	0.0031	99.38	***p <.00001
Portuguese	155	65.49%	0.885	0.6434	0.0056	42.88	***p <.00001
Russian	487	72.61%	0.8994	0.554	0.0026	131.63	***p <.00001
Sanskrit	99	59.43%	0.6186	0.5356	0.0055	15.12	***p <.00001
Serbo-Croatian	748	74.01%	0.8654	0.5878	0.003	92.28	***p <.00001
Slovak	129	68.58%	0.7927	0.5564	0.008	29.41	***p <.00001
Spanish	1853	67.88%	0.806	0.6324	0.0022	79.89	***p <.00001
Swahili	124	74.55%	0.5403	0.4454	0.0066	14.46	***p <.00001
Urdu	47	83.5%	0.7451	0.6912	0.0104	5.18	***p <.00001

***: p<.001
Note: Histograms displaying the distribution of simulation purities relative to real language purity can be found in Appendix A

UMAP, Visualization, and Translation

After reducing the PCA vectors to 2 dimensions using UMAP and translating all tokens to English in 18 of the 19 languages (due to Google Cloud not supporting Sanskrit), visualizations were made that represented the global structure of our word embedding space. As hinted at by the HDBSCAN results, there are no clear clusters as one would expect based on the literature (e.g. Rosch's Superordinate, animacy, humanness, etc). Rather, it is generally a large blob of words without cohesive separation except for a few cases (which tend to be composed of human names rather than theory-driven semantic fields). All visualizations with translations were made to be interactive in that they allow users to select a specific cluster to view, hover over tokens in semantic space to see the concept it represents and its translation, and view the noun class of those concepts. All visualizations are hosted online for public exploration at https://ardianor-ops.github.io/semanticorewebsite/.

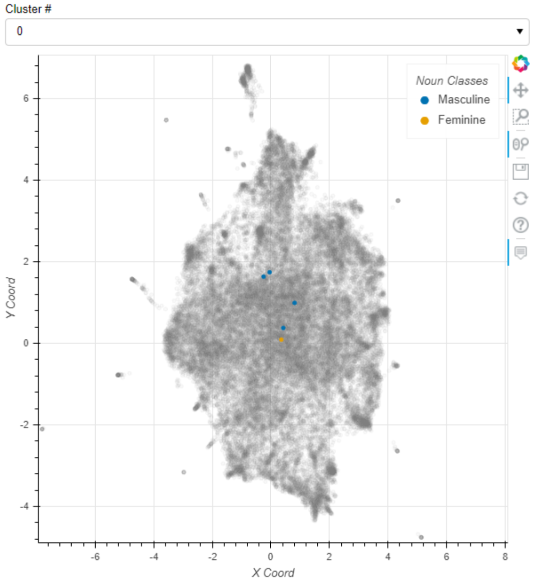


Figure 3: The interactive UMAP embedding for French hosted on https://ardianor-ops.github.io/semanticorewebsite/

Nearest Neighbor Search

As one of the drawbacks of using density-based clustering was the consequence of throwing away a large chunk of the data, the Nearest Neighbor Search not only acts as a separate measure of evaluating noun class-lexical semantics systematicity, but also makes use of the entire dataset because it looks at the closest neighbor of all nouns regardless of its distance away in semantic space. As expected of datasets of this size, all Chi-square tests reached significance, but more importantly all languages had at least a small effect size and 17 of the 19 had moderate to high effect sizes for Cramer's V (see Table 4a).

Table 4a.
Chi Square Tests of Real Noun Class Labels against Nearest Neighbor Noun Class Labels

Language	Pearson-Chi Square	DF	p	Cramer's V
Arabic	2637.7524	1	p <.00001	**0.4781
Bulgarian	2969.6406	4	p <.00001	***0.5592
Dutch	16546.7063	9	p <.00001	**0.4272
French	16342.4231	1	p <.00001	***0.6116
German	34977.2351	4	p <.00001	***0.639
Hebrew	1721.7596	1	p <.00001	***0.5396
Hindi	1226.7637	1	p <.00001	**0.3644
Icelandic	7434.7975	4	p <.00001	***0.5808
Italian	30119.2561	1	p <.00001	***0.6753
Maltese	666.5157	1	p <.00001	**0.4186
Polish	33507.3498	4	p <.00001	***0.8043
Portuguese	2842.0299	1	p <.00001	***0.731
Russian	36414.2639	4	p <.00001	***0.8431
Sanskrit	302.3628	4	p <.00001	*0.2081
Serbo-Croatian	17942.7852	4	p <.00001	***0.5449
Slovak	3454.0785	4	p <.00001	***0.6721
Spanish	17326.5204	1	p <.00001	***0.5676

Table 4a.
Chi Square Tests of Real Noun Class Labels against Nearest Neighbor Noun Class Labels

Language	Pearson-Chi Square	DF	p	Cramer's V
Swahili	4532.9992	36	p <.00001	**0.4428
Urdu	96.4994	1	p <.00001	*0.1976
*: Low Effect Size (0.1 ≥ V > 0.3) **: Moderate Effect Size (0.3 ≥ V > 0.5) ***: Large Effect Size (V ≥ 0.5)				

The classification model metrics revealed that using the noun class of a noun to approximate the noun class of its closest neighbor performs better than just always guessing the majority noun class in 17 of 19 languages, although the differences between some were very slight (See Table 4b, 4c, and 4d). Regardless, we are able to reject the null hypothesis effect size between the closeness of two nouns in semantic space and noun class assignment is no different than 0 and adopt the alternative hypothesis that the effect size between the two is different from 0.

Table 4b.
Classification Model Metrics for Real Label-Nearest Neighbor Comparison

			Noun Class Metrics											
			Masculine				Feminine				Neuter			
Language	Overall Model Accuracy	Majority Baseline Accuracy	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
**Arabic	76.43%	65%	76.43%	0.81	0.83	0.82	76.43%7	0.67	0.65	0.66				
**Bulgarian	73.11%	45.63%	78.08%	0.76	0.76	0.76	77.95%	0.73	0.73	0.73	90.19%	0.63	0.62	0.63
**French	80.88%	55.98%	80.88%	0.83	0.83	0.83	80.88%	0.79	0.,78	0.78				
**German	75.85%	51.31%	82.65%	0.72	0.74	0.73	82.88%	0.77	0.76	0.77	86.18%	0.78	0.77	0.78
**Hebrew	78.14%	61.06%	78.14%	0.82	0.82	0.82	78.14%	0.72	0.72	0.72				
**Hindi	72.03%	67.43%	72.03%	0.79	0.79	0.79	72.03%	0.57	0.57	0.57				
**Icelandic	72.47%	39.01%	83.28%	0.76	0.77	0.76	79.15%	0.74	0.73	0.73	72.47%	0.66	0.66	0.66
**Italian	83.85%	53.43%	83.85%	0.85	0.85	0.85	83.85%	0.85	0.82	0.83				
**Maltese	71.58%	57.03%	71.58%	0.75	0.76	0.75	71.58%	0.67	0.66	0.67				
**Polish	87.25%	47.44%	88.77%	0.87	0.89	0.88	90.20%	0.87	0.86	0.86	95.54%	0.88	0.84	0.86
**Portuguese	87.04%	59.36%	87.04%	0.84	0.84	0.84	87.04%	0.89	0.89	0.89				
**Russian	90.47%	49.60%	91.90%	0.91	0.92	0.92	92.99%	0.9	0.9	0.9	96.06%	0.88	0.85	0.87
Sanskrit	51.58%	54.67%	60.32%	0.65	0.61	0.63	72.69%	0.4	0.43	0.41	70.14%	0.36	0.38	0.37
**Serbo-Croatian	73.68%	49.27%	75.84%	0.75	0.77	0.76	77.64%	0.75	0.73	0.74	93.89%	0.58	0.6	0.59
**Slovak	81.06%	47.38%	84.04%	0.82	0.84	0.83	85.38%	0.83	0.83	0.83	92.70%	0.72	0.66	0.69
**Spanish	78.62%	54.90%	78.62%	0.8	0.81	0.81	78.62%	0.77	0.75	0.76				
Urdu	64.40%	66.63%	64.40%	0.73	0.74	0.73	64.40%	0.47	0.46	0.46				

Note:This is not to argue that noun classes are equivalent cross-linguistically, rather it is an intuitive way to organize this table. Please interpret Noun Class metrics individually.
**Language outperformed baseline model of only predicting majority noun class label

Table 4c.
Classification Model Metrics for Real Label-Nearest Neighbor Comparison for Dutch

			Noun Class Metrics															
			Masculine				Feminine				Neuter				Common			
Language	Overall Model Accuracy	Majority Baseline Accuracy	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
**Dutch	67.47%	35.72%	76.17%	0.66	0.68	0.67	76.34%	0.65	0.65	0.65	83.60%	0.73	0.71	0.72	98.83%	0.1	0.12	0.11

**Language outperformed baseline model of only predicting majority noun class label

Table 4d.
Classification Model Metrics for Real Label-Nearest Neighbor Comparison for Swahili

			Noun Class Metrics															
			Ki-vi				Ma				M-Mi				M-Wa			
Language	Overall Model Accuracy	Majority Baseline Accuracy	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
**Swahili	43.31%	41.41%	89.85%	0.41	0.46	0.43	86.79%	0.24	0.24	0.24	84.38%	0.36	0.33	0.35	76%	0.31	0.32	0.32
			N				Pa				U							
			Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score				
			67.90%	0.61	0.61	0.61	99.97%	1	0.8	0.89	81.71%	0.2	0.19	0.2				

**Language outperformed baseline model of only predicting majority noun class label

Discussion & Conclusion

We set out to resolve a gap in the linguistics and cognitive psychology literature by quantifying noun-class assignment and lexical semantics systematicity across several language families by using density-based clustering and a K-nearest neighbor search. In line with previous qualitative research on individual and subsets of languages that argue for a role of semantics in noun class assignment, we were able to not only identify and quantify this systematicity in 19 languages from 3 separate language families, but we were also able to produce a useful tool for further study in the form of hosting our clusterings online. The HDBSCAN analysis found that the cluster purities of real languages strongly outperform 1,000 simulated languages with random noun-class assignments but do not have any easily discernible global structure. We were able to rule out hypotheses regarding complete systematicity and complete randomness and conclude that reality falls somewhere in between although the nature of which must be left to future study. The K-Nearest Neighbor analysis reinforced these findings and indicated that using the noun class of the noun closest in meaning to another noun as an approximation of that noun's assignment is fairly reliable across most languages studied.

Limitations

One of the major drawbacks of our approach was the source of the data. While the literature appears to indicate that Wiktionary is a fair albeit imperfect approximation of cross-linguistic lexicography, its community-driven approach unfortunately leads to unsystematic compiling of concepts and, therefore, imbalanced datasets for our purposes. Furthermore, there is no a priori way to scrub through the concepts evaluated to partition them into meaningful semantic groups (e.g. animacy, kinship terms, etc.). Finally, likely as a result of decreased attention to non-Indo European languages, there was only one language (Swahili) from the Niger-Congo family with sufficient wiktionary data and FastText model that we could evaluate.

Another consequence of this recurring theme of ‘language obscurity’ was that the quality of our models likely suffered due to a lack of quality training corpora. less articles means less exposure and optimization for different semantic domains.

Finally, while we *were* able to quantify the role of semantics in noun class assignment, we lacked sufficient time to quantify it relative to other relevant features mentioned in the literature such as morphology and phonology.

Areas of Future Study

To address the data drawbacks, we are currently in the process of creating a series of lists to distribute to linguists that specialize in languages with relevant nominal classification systems. By doing so, we will have a more reliable source of data and will gain access to a broader range of languages. With this data, we intend to publicize this data in the form of a Noun Global Atlas (NouGAt for short) which will help support further research in the fields of typological, phonological, and historical linguistics.

Understanding the role of phonology may be accomplished by applying an n-gram approach similar to Pimentel et al. (2019) which identified cross-linguistic tendencies for nouns with certain phonological features to have similar meanings.

Further steps for the information found here are semantically pure clusters that are consistent cross-linguistically. This should be possible given studies such as Conneau et al. (2017) which allow for the rotating of high dimensional word embedding spaces in such a way that the same concept in two or more separate languages would occupy the same location in semantic space. By analyzing cross-linguistic consistency in noun class assignment we would take one step further from the large-scale tendency to classify nouns based on their meaning discovered here and understand whether there are exact concepts/semantic domains that humans tend to group together.

Implications

The geographic diversity and strength of our findings suggest a cognitive tendency to associate noun class with the meaning of a noun. The semantic domains in which this phenomenon occurs, however, while significantly purer than a random distribution of noun class assignments, are also not entirely pure either. If we are to understand that semantic proximity to other nouns is related to the assignment to a certain noun class, it is not entirely inappropriate to conclude that those not close enough to anything are assigned by factors outside of semantics (although this could only be definitively argued for when the dataset comprises the entire lexicon). With this in mind and assuming that we have captured a fair enough scope of the lexicon, the sheer amount of noise in our clusterings appear to be in support of there being random assignment of nouns surrounding core semantic domains of systematic noun class assignment, but due to the large number of clusters it is difficult to parse out whether there is further systematicity in these deviations as well and must be left to future study. Regardless, Twain's claim of there being "no sense or system in the distribution" of grammatical gender does not appear to be supported by our findings and, rather, this tendency is the case not only for German, but for languages from all over the world.

Acknowledgements

I would like to extend a huge thank you to Yujing Huang, Joshua Hartshorne, Wesley Ricketts, Seun Eisape, Aidas Aglinskey, Damian Blasi, and Katie McCauliffe for their endless support and assistance with this project.

References

Adams, K. L., & Conklin, N. F. (1973). Toward a theory of natural classification. *Paper from the Ninth Regional Meeting of the Chicago Linguistic Society,* , 1-10.

Barkin, F. (1980). THE ROLE OF LOANWORD ASSIMILATION IN GENDER ASSIGNMENT. *Bilingual Review / La Revista Bilingüe*, *7*(2),105–112. http://www.jstor.org/stable/25743887

Bergen, J. J. (1980). The semantics of gender contrasts in spanish. *Hispania, 63*(1), 48-57. doi:10.2307/340811

Bergen, J. J. (1978). A Simplified Approach for Teaching the Gender of Spanish Nouns. *Hispania*, *61*(4), 865–876. https://doi.org/10.2307/340934

Bojanowski, Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. Στο J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Επιμ.), *Advances in Knowledge Discovery and Data Mining* (σσ. 160–172). Berlin, Heidelberg: Springer Berlin Heidelberg.

Corbett, G. G. (1991). *Gender*. Cambridge [England] ; New York: Cambridge University Press.

Corbett, G. G. (2013). Number of genders. In M. S. Dryer, & M. Haspelmath (Eds.), *The world atlas of language structures online* (). Leipzig: Max Planck Institute for Evolutionary Anthropology.

Croft, W. (1994). Semantic universals in classifier systems. *Null, 45*(2), 145-171.

doi:10.1080/00437956.1994.11435922

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word Translation Without Parallel Data. doi:10.48550/ARXIV.1710.04087

Denny, J. P., & Creider, C. A. (1976). The semantics of noun classes in proto-bantu. *Studies in African Linguistics, 7*(1), 1.

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). *A functional theory of gender paradigms* doi:10.1163/9789004342934_011

Enger, H. (2009). The role of core and non-core semantic rules in gender assignment. *Lingua, 119*(9), 1281-1299. doi:10.1016/j.lingua.2009.02.004

Erichsen, G. (2020). "Is that noun masculine or feminine? Retrieved from thoughtco.com/noun-masculine-or-feminine-spanish-3079270.

Foundalis, H. E. (2002). Evolution of gender in indo-european languages. *Proceedings of the Annual Meeting of the Cognitive Science Society, 24*

Grave, Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages.

Greenberg, J. H. (2012). *Numeral classifiers and substantival number: Problems in the genesis of a linguistic type* doi:10.1163/9789004242050

Harris, Z. S. (1954). Distributional Structure. WORD, 10(2–3), 146–162. doi:10.1080/00437956.1954.11659520

Hepburn-Gray, R. (2020). *Niger-Congo Noun Classes: Reconstruction, Historical Implications, and Morphosyntactic Theory* (State University of New York at Buffalo, Ann Arbor; J. Good, Επιμ.). Retrieved from https://www.proquest.com/dissertations-theses/niger-congo-noun-classes-reconstruction/docview/2384853607/se-2?accountid=9673

Kaplan, J. (2017). From lexicostatistics to lexomics: Basic vocabulary and the study of language prehistory. *Osiris (Bruges), 32*(1), 202-223. doi:10.1086/694093

Kidd, E., & Garcia, R. (2021, September 6). How diverse is child language acquisition?. https://doi.org/10.31234/osf.io/jpeyq

Kramer. (2020). Grammatical Gender: A Close Look at Gender Assignment Across Languages. Annual Review of Linguistics, 6(1), 45–66. https://doi.org/10.1146/annurev-linguistics-011718-012450

Lakoff, G. (1983). Classifiers as a reflection of mind: The experiential, imaginative, and ecological aspects.

Lobben, M., Bochynska, A., Tanggaard, S., & Laeng, B. (2020). Classifiers in non-european languages and semantic impairments in western neurological patients have a common cognitive structure. *Lingua, 245*, 102929. doi:https://doi.org/10.1016/j.lingua.2020.102929

Lobben, M., & D'Ascenzo, S. (2015). Grounding grammatical categories: Attention bias in hand space influences grammatical congruency judgment of chinese nominal classifiers. *Frontiers in Psychology, 6*, 1299. doi:10.3389/fpsyg.2015.01299

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511809071

McCarthy, A. D., Williams, A., Liu, S., Yarowsky, D., & Cotterell, R. (2020). Measuring the similarity of grammatical gender systems by comparing partitions. *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp.

5664-5675). Online: Association for Computational Linguistics.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*(1), 89-115. doi:10.1146/annurev.ps.32.020181.000513

Mirkovic, J., Forrest, S., & Gaskell, G. (2011). Semantic regularities in grammatical categories: Learning grammatical gender in an artificial language.

Muller-Spitzer, & Wolfer, S. (2016). How many people constitute a crowd and what do they do? quantitative analyses of revisions in the English and German wiktionary editions. Lexikos, 26(1), 347–371. https://doi.org/10.5788/26-1-1346

Pimentel, T., McCarthy, A. D., Blasi, D. E., Roark, B., & Cotterell, R. (2019). Meaning to form: Measuring systematicity as information.

Steinmetz, D. (2006). Gender shifts in germanic and slavic: Semantic motivation for neuter? *Lingua, 116*(9), 1418-1440. doi:https://doi.org/10.1016/j.lingua.2004.06.014

Twain, Brown, Walter F. , illustrator, Kelsey, Charles C., former owner, & Kelsey, Jean A., donor. (1880). *A tramp abroad*: 601-620. American Pub. Co. ; Chatto & Windus.

Wang, W. (1994). Chinese classifier systems and human categorization. *In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change,* , 479-494.

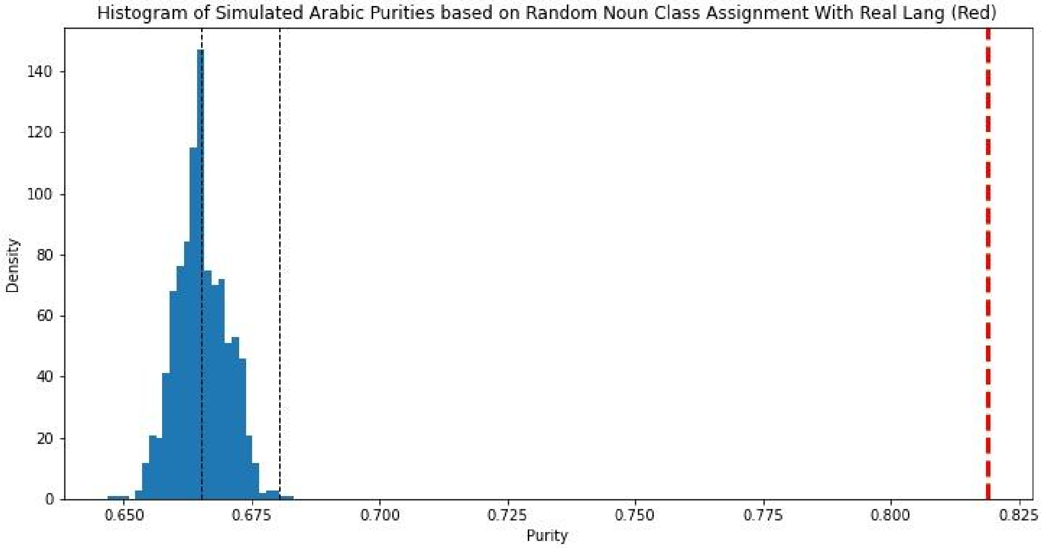
Williams, A., Cotterell, R., Wolf-Sonkin, L., Blasi, D., & Wallach, H. (2019). Quantifying the semantic core of gender systems.

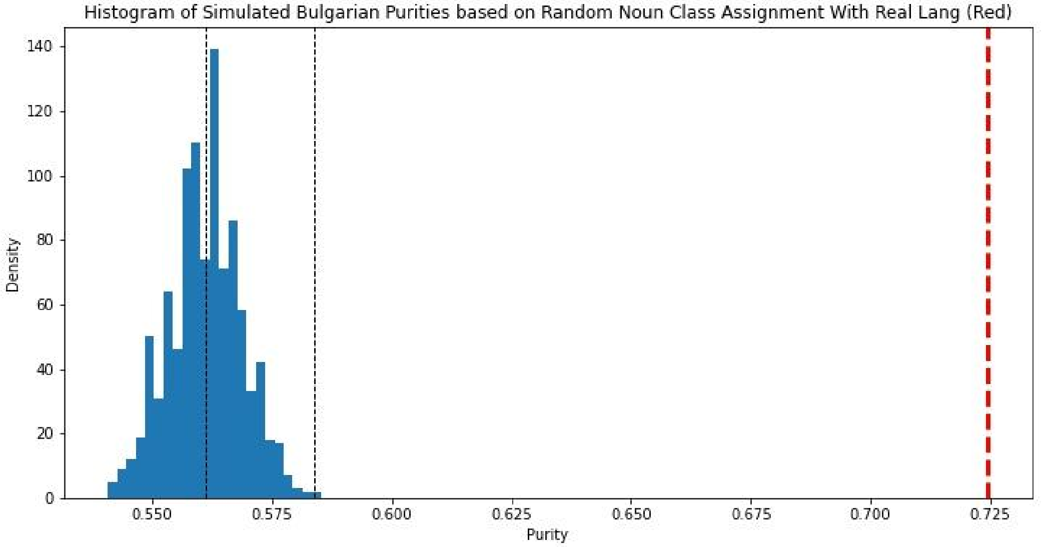
Yamamoto, K., & Keil, F. (2000). The acquisition of japanese numeral classifiers: Linkage between grammatical forms and conceptual categories. *Journal of East Asian Linguistics, 9*(4), 379-409.

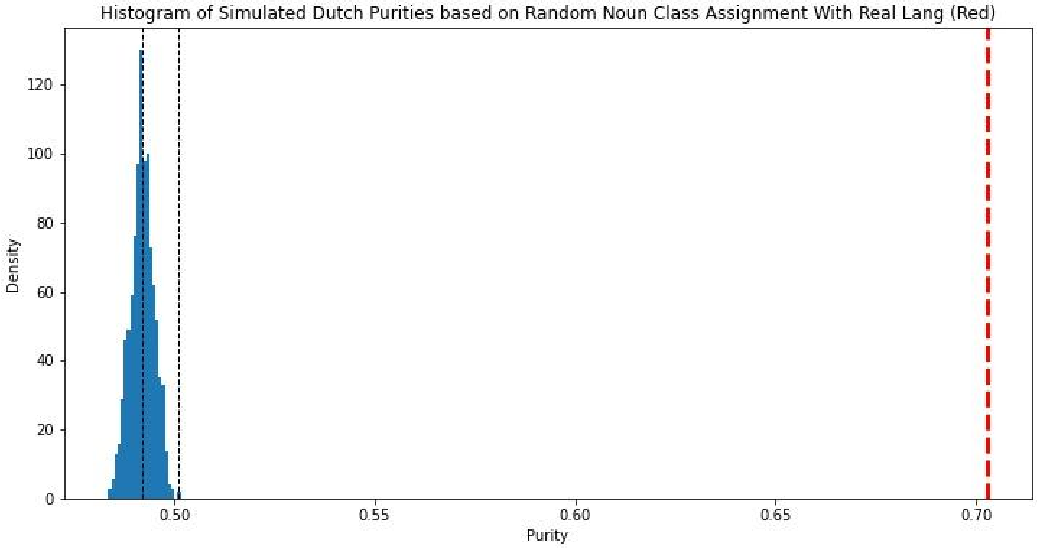
Zhang, H. (2007). Numeral classifiers in mandarin chinese. *Journal of East Asian Linguistics, 16*(1), 43-59. doi:10.1007/s10831-006-9006-9

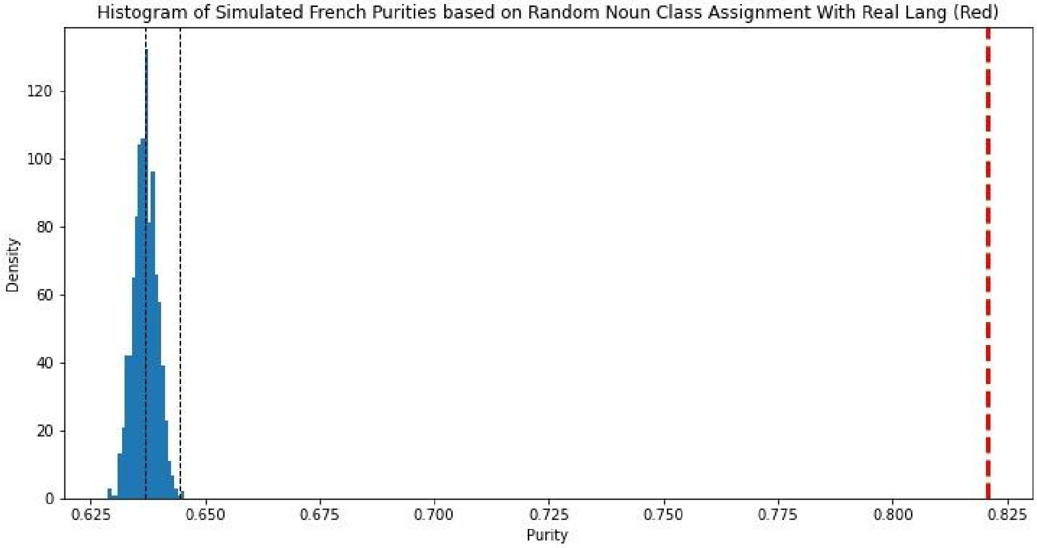
Zubin, D., & K-M, K. (1986). Gender and folk taxonomy: The indexical relation between grammatical and lexical categorization. In C. G. Craig (Ed.), *Noun classification and categorization* (pp. 139-180). Philadelphia: John Benjamins Publishing Company.

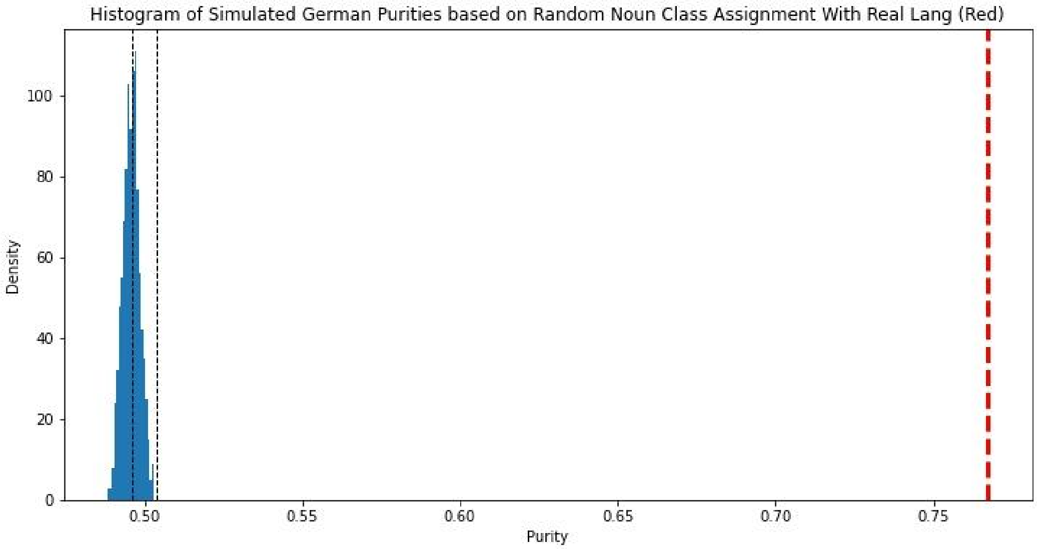
Appendix A

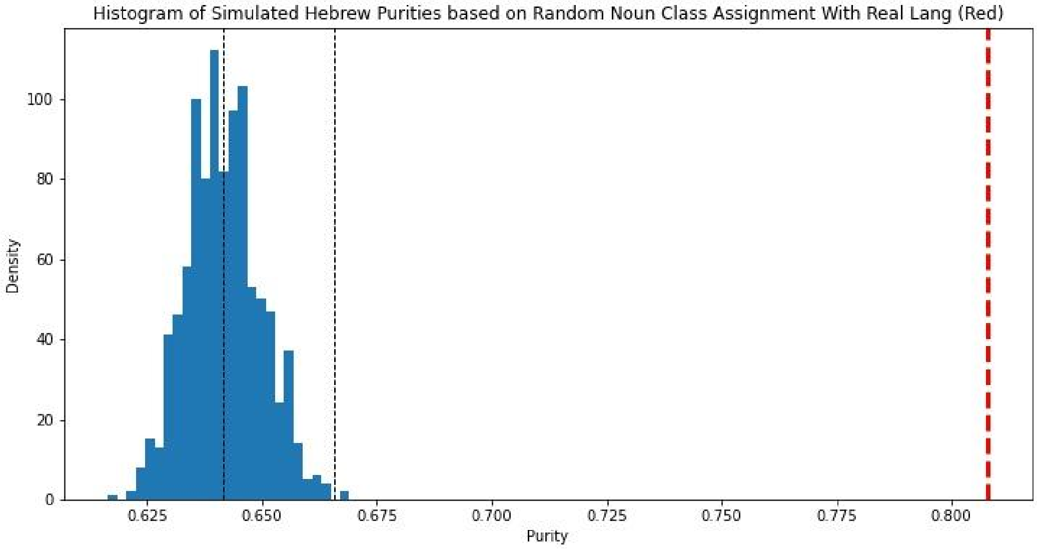

Histogram of Simulated Arabic Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
140
0.650
0.675
0.700
0.725
0.750
0.775
0.800
0.825
Purity

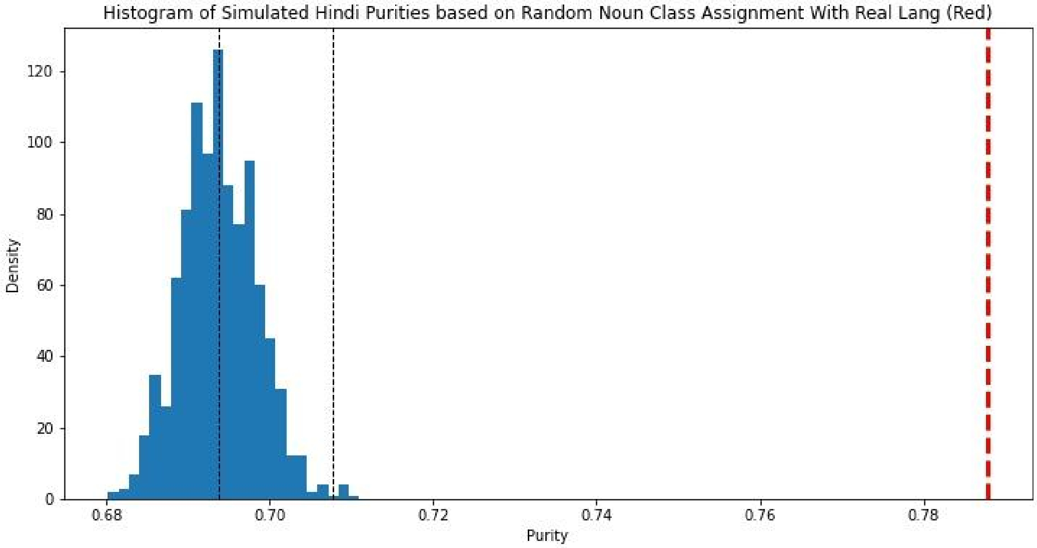

Histogram of Simulated Bulgarian Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
140
0.550
0.575
0.600
0.625
0.650
0.675
0.700
0.725
Purity

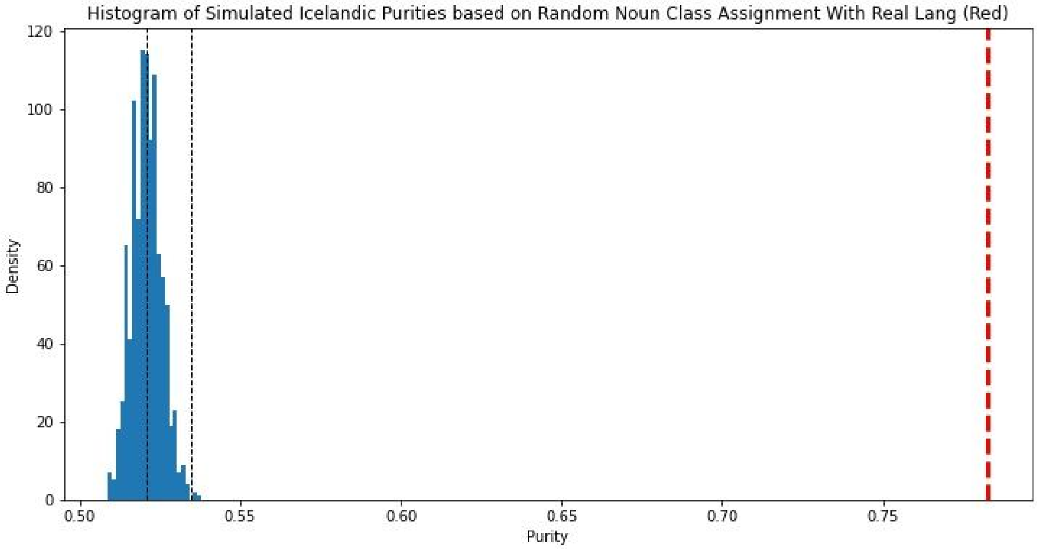

Histogram of Simulated Dutch Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
0.50
0.55
0.60
0.65
0.70
Purity

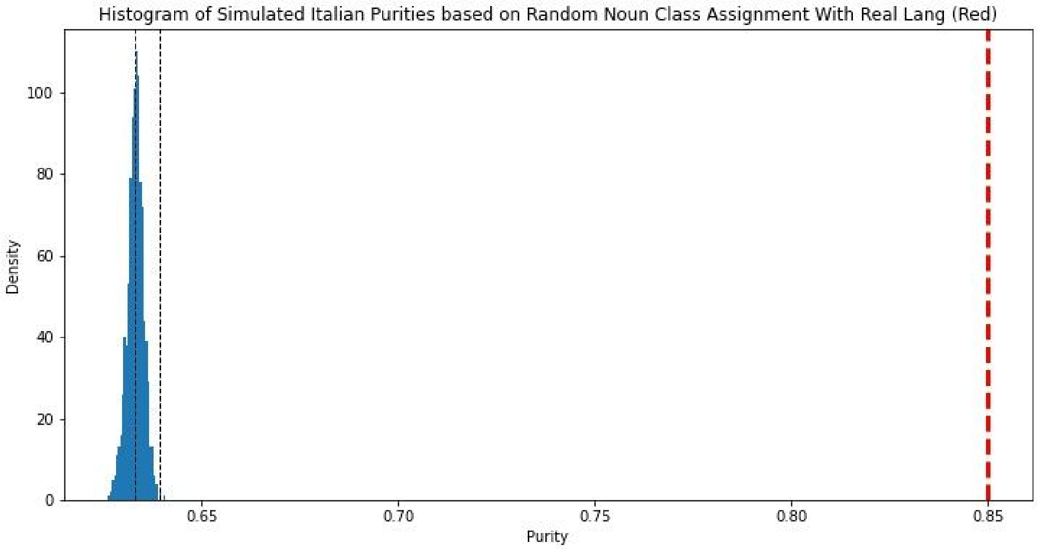

Histogram of Simulated French Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
0.625
0.650
0.675
0.700
0.725
0.750
0.775
0.800
0.825
Purity

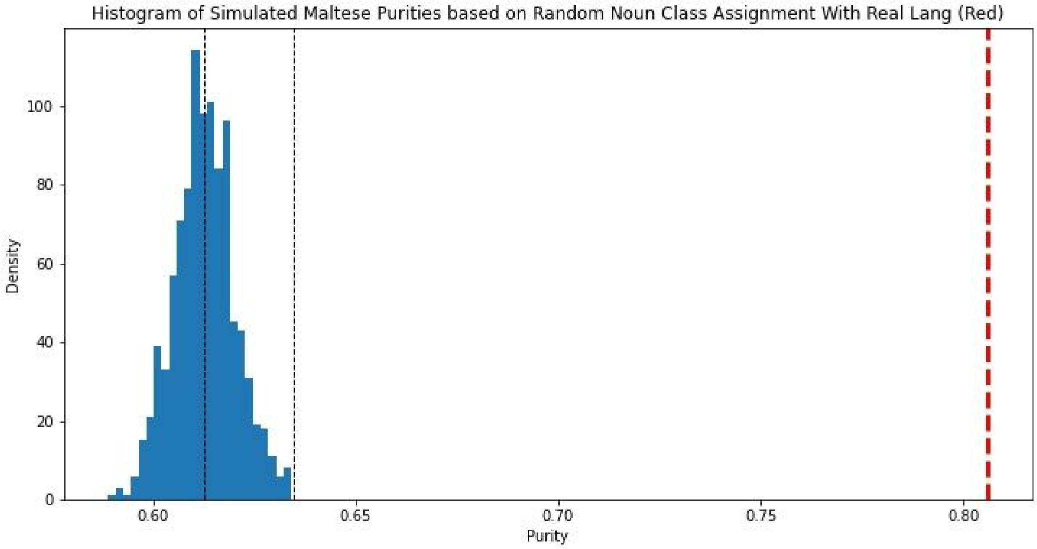

Histogram of Simulated German Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
0.50
0.55
0.60
0.65
0.70
0.75
Purity

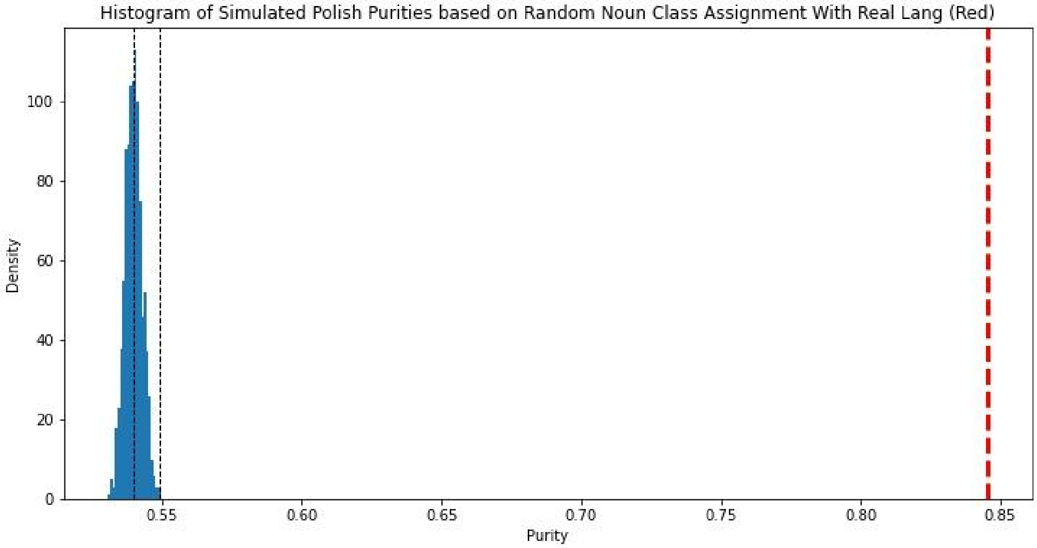

Histogram of Simulated Hebrew Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
0.625
0.650
0.675
0.700
0.725
0.750
0.775
0.800
Purity

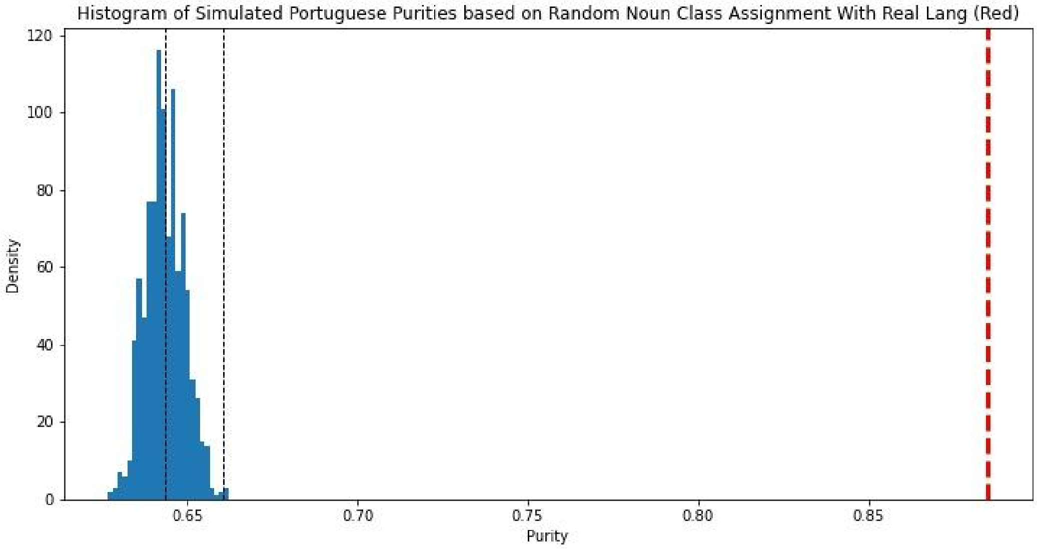

Histogram of Simulated Hindi Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
0.68
0.70
0.72
0.74
0.76
0.78
Purity

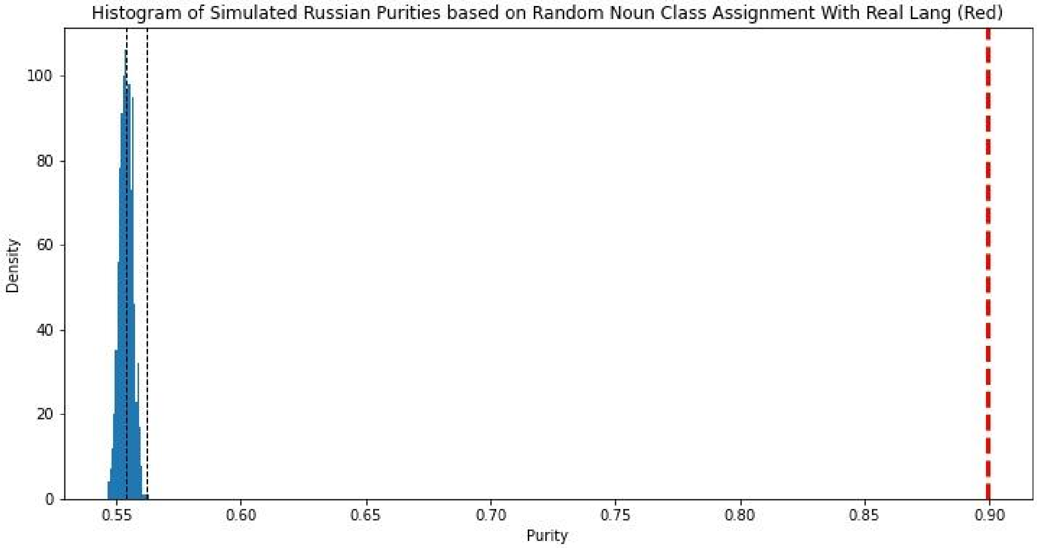

Histogram of Simulated Icelandic Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
0.50
0.55
0.60
0.65
0.70
0.75
Purity

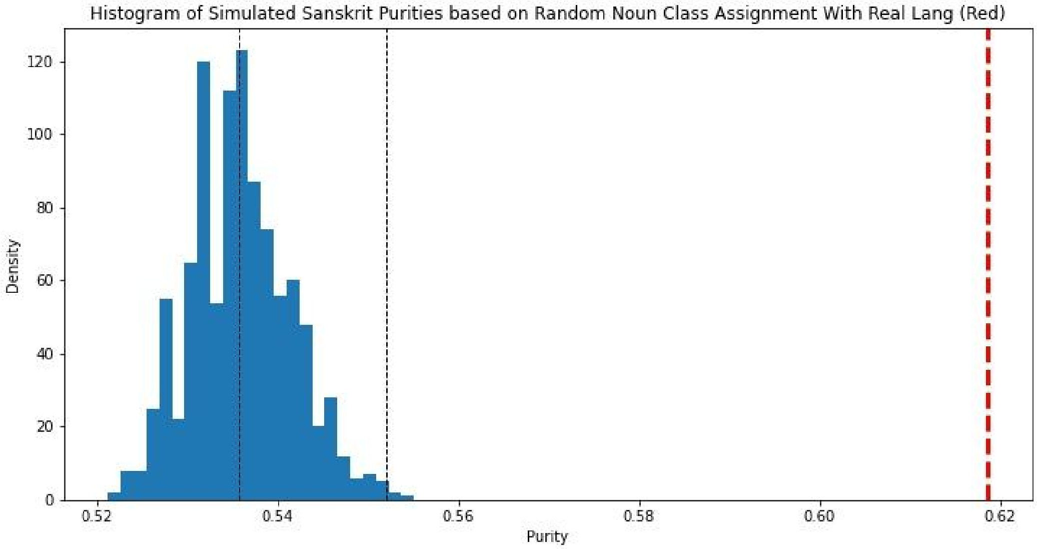

Histogram of Simulated Italian Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
0.65
0.70
0.75
0.80
0.85
Purity

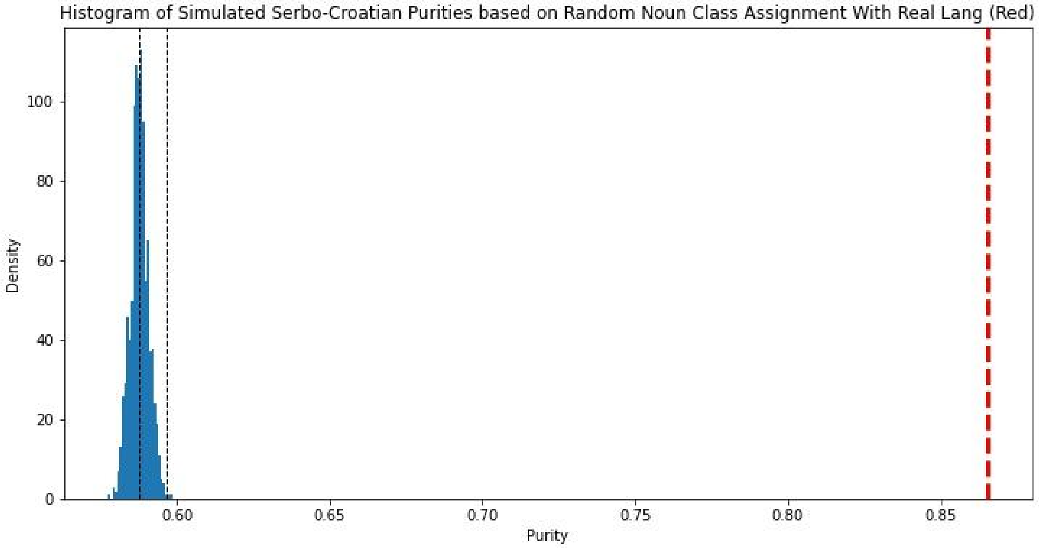

Histogram of Simulated Maltese Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
0.60
0.65
0.70
0.75
0.80
Purity

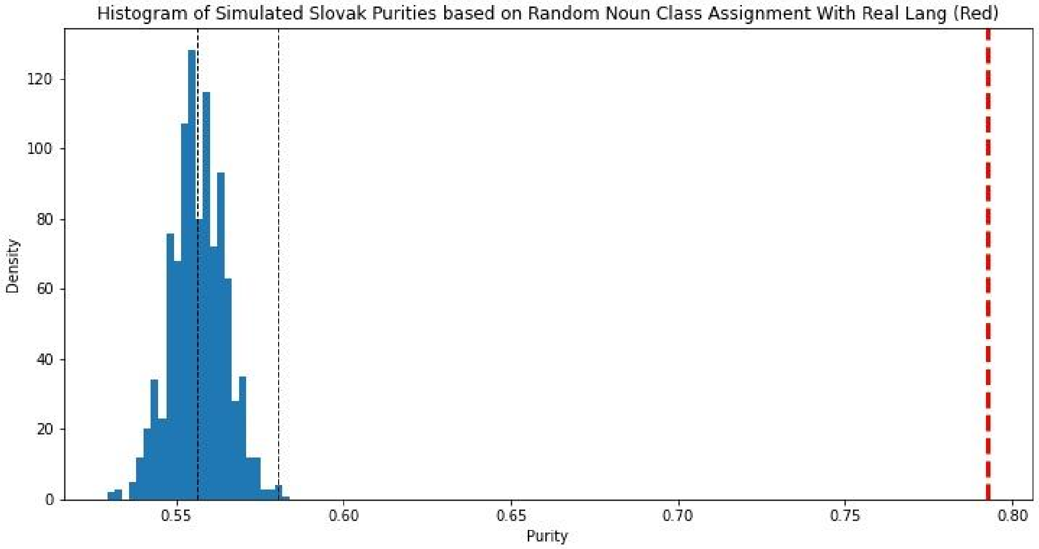

Histogram of Simulated Polish Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
0.55
0.60
0.65
0.70
0.75
0.80
0.85
Purity

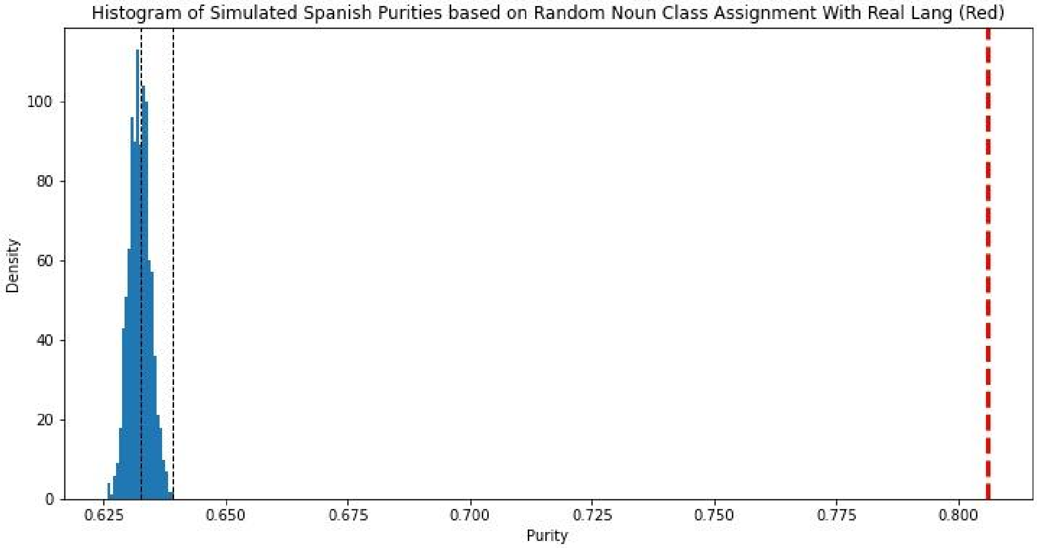

Histogram of Simulated Portuguese Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
0.65
0.70
0.75
0.80
0.85
Purity

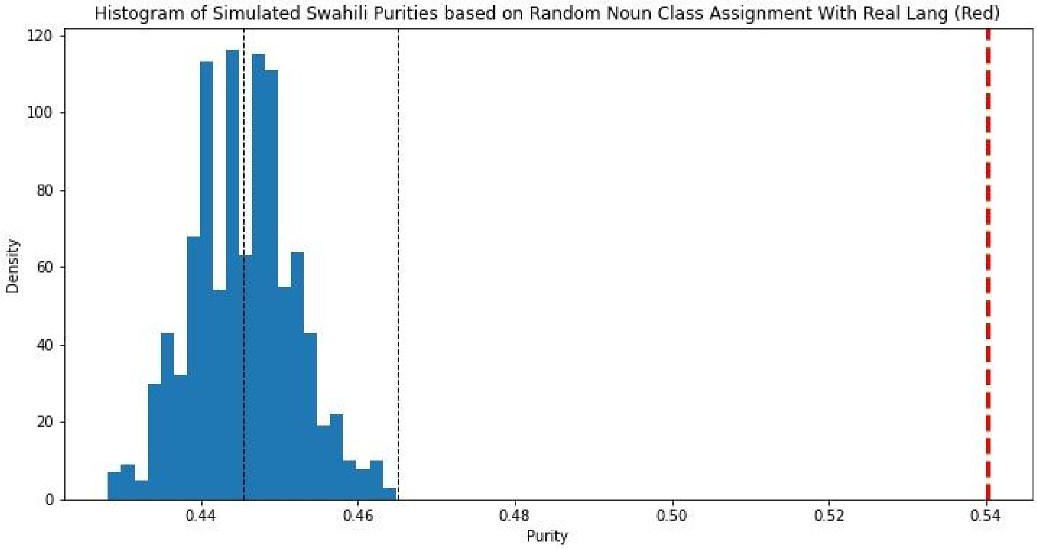

Histogram of Simulated Russian Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Purity

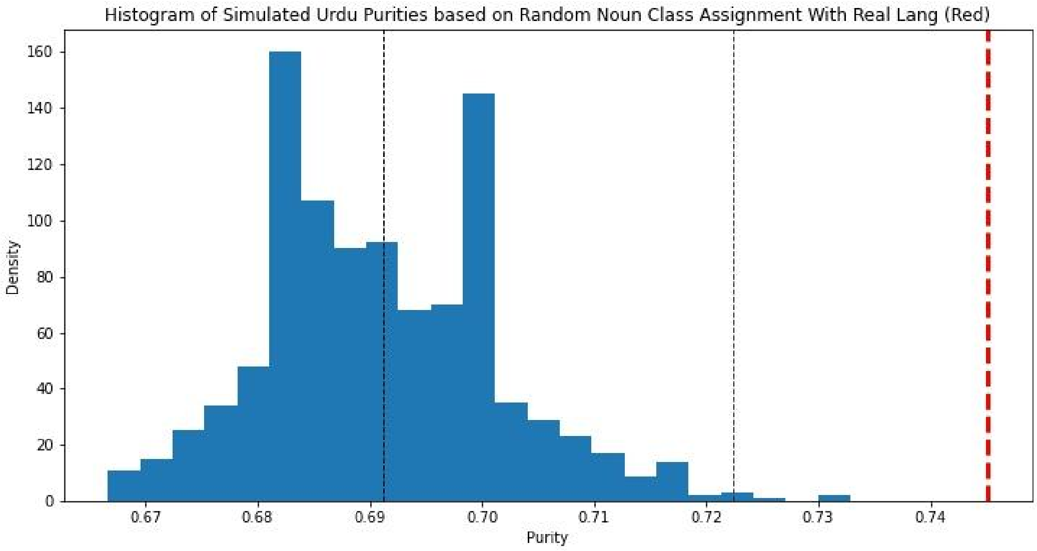

Histogram of Simulated Sanskrit Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
0.52
0.54
0.56
0.58
0.60
0.62
Purity


Histogram of Simulated Serbo-Croatian Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
0.60
0.65
0.70
0.75
0.80
0.85
Purity


Histogram of Simulated Slovak Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
0.55
0.60
0.65
0.70
0.75
0.80
Purity


Histogram of Simulated Spanish Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
0.625
0.650
0.675
0.700
0.725
0.750
0.775
0.800
Purity


Histogram of Simulated Swahili Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
0.44
0.46
0.48
0.50
0.52
0.54
Purity


Histogram of Simulated Urdu Purities based on Random Noun Class Assignment With Real Lang (Red)
Density
0
20
40
60
80
100
120
140
160
0.67
0.68
0.69
0.70
0.71
0.72
0.73
0.74
Purity