Lexical and Morpho-syntactic Features in Word Embeddings
A Case Study of Nouns in Swedish

Ali Basirat and Marc Tang
Department of Linguistics and Philology, Uppsala University, Box 635, 75 126, Uppsala, Sweden

Keywords: Neural Network, Nominal Classification, Swedish, Word Embedding.

Abstract: We apply real-valued word vectors combined with two different types of classifiers (linear discriminant analysis and feed-forward neural network) to scrutinize whether basic nominal categories can be captured by simple word embedding models. We also provide a linguistic analysis of the errors generated by the classifiers. The targeted language is Swedish, in which we investigate three nominal aspects: uter/neuter, common/proper, and count/mass. They represent respectively grammatical, semantic, and mixed types of nominal classification within languages. Our results show that word embeddings can capture typical grammatical and semantic features such as uter/neuter and common/proper nouns. Nevertheless, the model encounters difficulties to identify classes such as count/mass which not only combine both grammatical and semantic properties, but are also subject to conversion and shift. Hence, we answer the call of the Special Session on Natural Language Processing in Artificial Intelligence by approaching the topic of interfaces between morphology, lexicon, semantics, and syntax via interdisciplinary methods combining machine learning of language and general linguistics.

1 INTRODUCTION

The continuous vector representation of words, known as word embeddings or word vector, has been widely used in different areas of natural language processing. The studies on the distribution of word vectors show that different types of word similarities are captured by word vectors, e.g., the semantically similar words are clustered together. However, it is not still clear what types of linguistic information (e.g., semantic or syntactic word classes) are encoded in these vectors. By way of illustration, the semantic and syntactic information of different types of nominal classes is mirrored through language via grammatical systems of nominal classification. Thus, the investigation of such linguistic structures in word vectors is of high interest.

Nominal classification refers to how a language classify nouns of the lexicon. The most frequent grammaticalized system of nominal classification is grammatical gender (Corbett, 1991; Seifart, 2010; Corbett, 2013), e.g., all nouns in French are affiliated to either masculine or feminine. Nouns may equivalently be divided in categories according to different semantic and/or syntactic criterion, i.e., count/mass and common/proper (Delahunty and Garvey, 2010).

We study different types of nominal classifications in Swedish with regard to the syntactic and semantic information encoded to their word vectors. As a research question, we aim at evaluating the performance of word embeddings on the nominal classification task. The outputs of such investigation are not only expected to serve as a baseline for future research in terms of computational models of nominal classification, but also be compared with the findings of cognitive and linguistic studies with regard to nominal classification (Aikhenvald, 2012; Kemmerer, 2017). Our studies are on the basis of the accuracy of a classifier in predicting different classes of nouns from their word vectors. To this end, we train a classifier with word vectors as input and nominal classes as output.

As a case study, we select three binary nominal features in Swedish: uter/neuter (i.e., grammatical gender), count/mass, and proper/common nouns. These distinctions represent three different types of nominal classification. First, grammatical gender is a typically grammaticalized feature, which is reflected in language via grammatical agreement with other elements of phrase. For instance, in Swedish, the article and the adjective varies in terms of form depending on the grammatical gender (uter or neuter) of the following noun, c.f., *ett stor-t äpple* (a.SG.NEUT big.SG.NEUT apple.SG.NEUT) ‘a big apple’ and *en stor-∅ häst* (a.SG.UTER big.SG.UTER horse.SG.UTER)

Basirat, A. and Tang, M.
Lexical and Morpho-syntactic Features in Word Embeddings - A Case Study of Nouns in Swedish.
DOI: 10.5220/0006729606630674
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 663-674
ISBN: 978-989-758-275-2

'a big horse'. Every noun can only be associated to one grammatical gender, regardless of register and context. Hence, grammatical gender is considered as a static nominal feature in Swedish. [1]

Second, the distinction between common and proper nouns (proper names) is considered as a static semantic feature. Common nouns generally refer to classes of things (e.g., *printer*, *desk*) while proper nouns designate particular individual entities such as *Stockholm*, *Paris*, among others (Delahunty and Garvey, 2010, p.149). To be more precise, proper nouns do not necessarily refer to only one specific individual. By way of illustration, the proper name *Smith* may be given to different people (or even objects) which do not share any property or quality in common. On the other hand, common nouns refers to a set of entities which share specific properties. Thus, common nouns may not be used in the same arbitrary way of proper nouns. As an example, naming a desk as a *fridge* in English would not be semantically valid, since the properties of a desk do not concord with the intrinsic semantic content of the noun *fridge* (e.g., a fridge must have a cold temperature compartment). Following this logic, nouns are either common or proper and do not fluctuate between the two categories. Hence, it is defined as semantically static. Few exceptions of conversion are attested but they are context specific (Gillon, 1999, p.58), thus we do not consider them in the current paper.

Finally, the count/mass distinction is recognized as an intermediary between syntactic and semantic nominal features (Doetjes, 2012), as "the brain differentiates between count and mass nouns not only at the syntactic level but also at the semantic level" (Chiarelli et al., 2011). With regard to semantics, count nouns point to objects which represent a discrete entity and may be counted, e.g., *computer*, *book* in English. On the other hand, mass nouns (also named non-count nouns) generally refer to objects which are interpreted as a uncountable mass and are not specified as how to individuate or divide them, e.g., *sand*, *milk* in English. In terms of morpho-syntax, the categories of count and mass nouns are also differentiated, as mass nouns cannot occur in the plural form even if the language has plural inflection (among other criteria). However, it is considered difficult to define absolute syntactic and semantic parameters to entirely isolate count and mass nouns. As an example, a noun may undergo conversion between count and mass, e.g., *pizza* may be used as a mass noun or or a count noun depending on context, c.f., *did you order pizza?* and *I ordered a pizza* (Gillon, 1999, p.51). Moreover, the presence/absence of grammatical number marking (i.e., plural) on the noun does not completely correlate with the count/mass distinction (Corbett, 2000; Dryer, 2005). For instance, certain mass nouns may have a count interpretation but be morphologically mass nouns as they only occur in the singular form, e.g., luggage (Pelletier and Schubert, 1989).

As a summary, the count/mass distinction is more versatile than the uter/neuter and proper/common nouns classification. Therefore, we estimate that the classifiers should not have difficulties interpreting the uter/neuter and proper/common features of the nouns, since they provide transparent syntactic and semantic clues. As for count/mass, we expect that the classifiers will be able to distinguish between the two classes, but with less accuracy than grammatical gender and proper names.

The structure of this paper is as follow. Section 2 presents a literature review of the three nominal features approached in our research question. Section 3 explains the structure of the selected computational models. Section 4 lists the setting of our experiments. Section 5 details how the classifiers performed with regard to the three nominal features involved in our study. In Section 6, we provide an explanation to the performance of the classifier, along with an error analysis. Finally, we conclude in Section 7.

2 LITERATURE REVIEW

In this section we describe the linguistic definition and examples of grammatical gender, common/proper, and count/mass distinction. Moreover, we provide an overview of the difficulties of classification observed from a linguistic approach.

2.1 Grammatical Gender

Nominal classification, i.e., how languages classify nouns of the lexicon, reflects cognitive and cultural facets of the human mind (Aikhenvald, 2012; Contini-Morava and Kilarski, 2013; Kemmerer, 2017). Grammatical genders are one of the most common systems

[1] Static gender systems are not obligatorily found in every gender language. By way of illustration, Pnar (Austro-Asiatic) applies a grammatical gender system based on the masculine/feminine/neuter distinction. However, a noun in Pnar may have different genders depending on the properties attributed to it. For instance, the noun *deiñ* 'tree' can be marked by both masculine and feminine proclitics. The masculine form *u=deiñ* refers to a standing, upright tree, while *ka=deiñ* refers to a fallen tree (log) or wood. Phenomena of gender variation are likewise attested in languages such as Hindi, Nepali, among others (Nespital, 1990; Hall, 2002; Pokharel, 2010).

of nominal classification (Corbett, 1991). They are widely distributed in Africa, Europe, Australia, Oceania, while their presence is sporadicly attested in the Pacific, Asia and Americas (Aikhenvald, 2000, p.78).

The definition of grammatical gender involves grammatical agreement between a noun and other syntactic units which appear with the noun in a sentence. In other words, all nouns of the lexicon are assigned to certain classes. For instance, a language has two genders if two classes of nouns can be differentiated by their agreement marker (Senft, 2000). As an example in French, *frigo* 'fridge' is masculine and *chaise* 'chair' is feminine. Thus, grammatical gender is reflected via agreement on the articles, adjectives, and verbs, c.f., *un grand frigo* (one.MASC big.MASC fridge.MASC) 'A big fridge' and *une grande chaise* (one.FEM big.FEM chair.FEM) 'A big chair'. Likewise in Swedish, all nouns are affiliated to either uter or neuter gender (Bohnacker, 2004, p.198), which is reflected through grammatical agreement (see examples in Section 1). Thus, the syntactic clues provided by gender agreement are expected to be easily captured by the word embedding model and mirrored in the generated word vectors.

While the main functions of gender are relatively transparent, e.g., to facilitate referent tracking in discourse (Dixon, 1986; Nichols, 1989; Contini-Morava and Kilarski, 2013), gender assignment is considered as much more opaque (Corbett, 1991, p.57). However, contradictory observation occur for Swedish. While grammatical gender is commonly considered as arbitrary (Andersson, 1992; Teleman et al., 1999), several semantic principles are attested, i.e., nouns referring to human and non-human animates tend to be affiliated to the uter gender, while inanimates are more likely to be neuter (Dahl, 2000). Moreover, nouns pointing at concrete or countable entities are generally uter while abstract or mass nouns are favored by the neuter gender (Fraurud, 2000).

2.2 Common/Proper Nouns

Nevertheless, not all classification of nouns in the lexicon are interpreted via agreement marking. As an example, the distinction between common and proper nouns is based on semantics. Common nouns refer to a general things which share certain properties, while proper nouns name specific individual things (Delahunty and Garvey, 2010, p.149). By way of illustration for common nouns, one of the inherent properties of a *printer* is 'a device which can print materials'. Proper nouns, on the other hand, do not carry such intrinsic properties. For instance, any person, animal, and object may be named after *Simon*. The differentiation between common and proper nouns is realized via capitalization in languages such as English. However, such marking is only represented in the writing system, which is an additional invention of language. Moreover, the rules of capitalization of proper nouns are not consistent in a cross-language manner. English capitalizes names of the months (e.g., *December*) while Swedish does not. Furthermore, languages do not apply the same parameters, e.g., in German, all nouns are capitalized. Finally, the rule of capitalizing sentence initial words in certain languages may also interfere with the recognition of proper nouns solely based on capitalization.

Similar observations are made in terms of syntax, as no absolute criterion is valid to differentiate between common and proper nouns. For instance, in English, proper nouns generally appear as bare forms, e.g., *John*, *Apple*. However, the definite article is preferred for certain types of proper nouns, e.g., *the United States*. Moreover, the use of article with proper noun is not universal cross-linguistically. By way of illustration, country names are used with the definite article in French, e.g., *la France* (the.FEM France), while in English it is not required (e.g., *France*). Therefore, the classification of common and proper nouns is a relevant issue for machine translation (Lopez, 2008), sentiment analysis (Pang and Lee, 2008), topic tracking (Petrovic et al., 2010), web data search (Baeza-Yates and Ribeiro-Neto, 2011), case restoration (Baldwin et al., 2009), among others topics (Preiss and Stevenson, 2013).

2.3 Count/Mass Distinction

The third main subdivision of nouns is the contrast between count and mass. Count nouns are commonly perceived as entities which can be individuated and counted, while mass nouns incarnate things as a mass whose parts are not considered as discrete units (Delahunty and Garvey, 2010, p.156). As an example, a piece of a cake is still cake, however, a piece of a desk is not a desk. The reflection of the cognitive principle of individuation have been studied with regard to its connection to human cognition and how such cognitive concept is mirrored through language (Quine, 1960; Chierchia, 1998; Chierchia, 2010; Doetjes, 2012). For instance, morphological marking of grammatical number has been attested as one of the main linguistic realization of count/mass marking (Gillon, 1999; Borer, 2005). Hence, we expect that the word embedding model is capable of retrieving the syntactic information and encode it into the vectors.

Formally, count nouns may bear singular and grammatical plural marking (c.f., *the book is here* and *the books are here* in English), while mass nouns may solely occur in singular (c.f., the furniture is new, *the furnitures are new). Moreover, only count nouns can apply indefinite article (c.f., *a table* and **a luggage*), among other syntactic criteria (Chierchia, 1998). Semantically, count and mass nouns may be distinguished on the basis of cumulativity (Quine, 1960), divisibility (Cheng, 1973), and specificity (Gillon, 1999, p.51-53). Mass nouns are unspecified as how to be cumulated and divided, while count nouns are specified for how to be cumulated or divided. By way of illustration, *coffee* may be counted in terms of cup, glass, barrel, brand, among other measure terms (Kilarski, 2014, p.9), with no intrinsic specification. On the opposite, count nouns such as *book* can inherently be counted by the mean of cardinal numbers and cannot be divided.

Nevertheless, even though both syntactic and semantic criteria are available to distinguish count and mass nouns, the fact that nouns may undergo conversion (shift) and migrate to the other category represents a challenge for classification (Gillon, 1999). For instance, *coffee* in English may be used as a mass noun when referring to a type of beverage, e.g., *the coffee is good*. Nonetheless, the same word form can equivalently be employed as a count noun when referring to ‘a cup of coffee’, e.g., *I would like to have two coffees*. Such conversion is extremely productive and common within languages of the world (Doetjes, 2012, p.14) and represents one of the major difficulty of identifying count and mass nouns in terms of computational linguistics (Katz and Zamparelli, 2012).

3 METHODOLOGY

We selected Swedish as language of analysis due to the contradictory observation attested on grammatical gender (see Section 2.1) and its availability of first hand data. For each nominal classification task, we train a classifier on a set of word vectors which are labeled with their corresponding nominal classes. The process is divided into three steps: vector generation (word embedding), data labeling, and classification.

During the first stage, word embedding, a corpus of raw sentences with word segmentation is fed to a word embedding model. The word embedding model assigns a vector to each word in the corpus. This vectors assignment is done in such a way that semantic similarities of words is modeled by the correlation between their corresponding vectors, i.e., the semantically similar words are assigned to the similar vectors.

Then, in the data labeling step, a dictionary is used to associated a subset of word vectors with their corresponding nominal classes. These are the word vectors associated with some nouns in the dictionary whose nominal classes could be extracted from the dictionary. Later on, in Section 4, we elaborate how the nominal classes are extracted from the dictionary. This gives us a list of word vectors labeled with nominal classes. This list is partitioned into train, and test sets, to be used in the classification step.

Finally, in the classification, we train a classifier on the pairs of word vectors and nominal classes. The classifier takes word vectors as input and predict the nominal classes in its output. The train data is used to train the classifier, and the test set is used to evaluate the classification model. Since, this research aims to study the information provided by word vectors for different nominal classification tasks, we use simple classification methods without performing complicated tuning step. This is why we don’t keep a part of data as development set to tune the classifier.

The evaluation is on the basis of the performance of word vectors on each of the nominal classification tasks, e.g., the higher the accuracy of the classification is, the more the information about the nominal classes is encoded into the word vectors. In the next section, we elaborate the tools, models, and data used in each of these steps, word embedding, data labeling, and classification.

4 EXPERIMENTAL SETTINGS

Our word vectors are generated via RSV (Real-valued Syntactic Word Vectors) model for word embedding (Basirat and Nivre, 2017). RSV extracts a set of word vectors from an unlabeled data in three major steps: First, it builds a co-occurrence matrix whose elements are the frequency of seeing words together. The elements of this matrix are the frequency of seeing words in the domain of different context words. The columns are associated with words and the rows are associated with contexts. Each column forms a high dimensional word vector that describes the word with respect to its occurrence frequency in different contexts. In the second step, the elements of the high dimensional column vectors are transformed in such a way that the distribution of the vectors is more close to the Gaussian distribution with zero mean. Finally, in the third step, it forms the low dimensional data from the top K right singular vectors of the transformed co-occurrence matrix. Within this process, the RSV model has the following parameters:

- Context type: the context of a word may refer to the preceding words (symmetric), following words or include both directions (asymmetric).

- Context size: how many words does the system count in the context. As an example, the most popular setting is one preceding word.
- Dimensionality: the quantity of dimensions the model may use to represent the word vectors. The amount of dimensions is generally positively correlated to the accuracy, but negatively correlated with the processing time and memory.

Within our experiments we set context type and context size as the immediate preceding word, as proposed by (Basirat and Nivre, 2017). The number of dimensions is set to 50.

We include two types of classifiers in our experiments: linear discriminant analysis, and neural network. The purpose of this selection is to verify the complexity of nominal properties and to see if the nominal classes are linearly separable. Linear discriminant analysis (LDA) is a linear generative classifier that fits a Gaussian distributions on the training data and predict the test data classes through the likelihood ratio of the data given the distributions. Neural network (NN) is a non-linear discriminative classifier that make no assumption on the data distribution. It searches for a boundary between the data points with regard to their classes in such a way that the classification accuracy is minimized. LDA is more simple in terms of structure and processing time but may be less accurate depending on the complexity of the task. While Neural Network is the more elaborate type of classifier but also costly in terms of processing time.

Our computational model is fed with two types of data, which both originate from the Swedish Language Bank (Språkbanken) at the University of Gothenburg. [2] The word vectors are generated from a corpus of Swedish raw sentences. This corpus is compiled by Språkbanken and involves data from Swedish Wikipedia (available at Wikipedia Monolingual Corpora, Swedish web news corpora (2001-2013) and Swedish Wikipedia corpus). The OpenNLP sentence splitter and tokenizer are used for normalizing the raw corpus. [3] We replace all numbers with a special token NUMBER and convert uppercase letters to lowercase forms. Due to the high ratio of compound nouns in Swedish (Carter et al., 1996; Ostling and Wirn, 2013; Ullman and Nivre, 2014), we only include nouns which have more than 100 occurrences in the corpus.

Second, the information on grammatical gender, common/proper, and count/mass distinction is extracted from the SALDO (Swedish Associative Thesaurus version 2) dictionary. The categorization of SALDO is rather "generous" and lists diverse properties of nouns (Borin et al., 2008, p.27). By way of illustration, nouns which were assigned two genders according to speaker variation are affiliated to neither *uter* nor *neuter* but rather to a third type *vacklande*. Moreover, some nouns are annotated as *blank* if their gender was difficult to interpret. The creation of these categories was driven by specific pragmatic and semantic classification purposes. In our study, we only incorporate the relevant categories. For instance, with regard to grammatical gender, we solely cover nouns annotated as *uter* and *neuter* since only these two classes fulfill the conditions of grammatical gender. Furthermore, the frequency and quantity of peripheral nouns such as *vacklande* and *blank* is not significant compared to the whole dataset (1%). Hence, we leave these patterns of variation for further studies to explore. We also filter out those nouns that are not in word vectors' vocabulary set. The result of our filtering is shown in Table 1.

Table 1: Nominal features in Swedish based on SALDO.

Category	**Quantity**	**Example**
uter	13540	*bok* 'book'
neuter	5518	*hus* 'house'
count	16181	*kontor* 'office'
mass	3085	*bagage* 'luggage'
comon	18549	*bord* 'table'
proper	3142	*Alyssa*

The distinction of grammatical gender and common/proper nouns is rather straightforward, as the information is annotated in SALDO. However, the count/mass distinction is not a category transparently specified in the dictionary. Thus, we follow the formal syntactic definition and consider that only count nouns may have plural inflection (Chierchia, 1998; Chierchia, 2010).

5 RESULTS

The results of our experiments are evaluated according to the *Rand index* (Rand, 1971) (accuracy) and *F-score* (Ting, 2010). To obtain the general accuracy of the model on the entire dataset, the Rand index is calculated by dividing the total number of correctly retrieved tokens by the total number of retrieved tokens. In the remaining parts of this paper, Rand index and accuracy are used interchangeably. As as side note, we do not compare directly our results with other studies as we are not aware of methodologies and approaches being applied in a way similar to ours.

Due to the lack of balance between the investi-

[2] spraakbanken.gu.se/eng/resources/corpus

[3] opennlp.apache.org

gated classes (e.g., 71% uter words vs 29% neuter words), we use majority label prediction, called *Zero rule*, as our baseline. In this case, our accuracy baseline for each of the nominal classifications is equal to relative size of the larger class, i.e., 71.0% for grammatical gender prediction, 84.0% for count/mass prediction, and 85.5% for common/proper noun prediction. Moreover, we expect to obtain adequate measures not only for the overall accuracy of the classifier, but also for the detailed performance on every single class. For instance, in the classification task of uter/neuter, did one of the two classes represent more difficulties for the classifier. Hence, we generate from the classifier's output the two values of *precision* and *recall*. Precision evaluates how many tokens are correct among all the output of the classifier, i.e., precision is equal to the total number of correctly retrieved tokens divided by the total number of retrieved tokens. Recall quantifies how many tokens are correctly retrieved among all the expected correct output, i.e., recall is obtained by dividing the total number of correctly retrieved tokens with the total number of correct tokens in the dataset. The two measures evaluate different facets of the output, thus they are merged into the F-score, which is equal to the harmonic mean of the precision and recall, $2\frac{\text{recall}\times\text{precision}}{\text{recall}+\text{precision}}$.

Furthermore, we also provide two figures for every class of nouns we targeted. First, we display how the noun classes are distributed in the distributional semantic space formed by the word vectors. Second, we show the histogram of the entropy of the neural network's output for each class of nouns. The entropy measures the uncertainty involved in the neural network's output to identify the noun classes. High values of the entropy can be interpreted as more uncertainty in the classifier's outputs, which itself show the weakness of the information provided by the input word vectors with regard to the nominal classes. The skewness of the histogram toward right or left shows the certainty of the classifier for a particular nominal class. Once the histogram is skewed toward right, the classifier is uncertain about its outputs. However, the left skewness means higher certainty in the output.

5.1 Grammatical Gender

The overall accuracy and processing time of the classifiers is shown in Table 2. As expected, neural network reaches higher accuracy (Rand index), as it was able to identify correctly the grammatical gender of 93.6% of the nouns in the test set. Such performance is conjointly higher than our baseline accuracy, which is 71.0%. LDA could reach relatively high accuracy as neural network but in significantly shorter time.

Table 2: The performance of LDA and neural network (NN) on the grammatical gender prediction.

	LDA	NN
accuracy	92.7%	93.6%
uter recall	94.8%	96.3%
uter precision	94.8%	95.0%
uter f-score	94.8%	95.6%
neuter recall	87.9%	87.4%
neuter precision	87.9%	90.1%
neuter f-score	87.9%	88.7%

This shows that the word vectors could encode the information about the grammatical genders of the nouns and they are almost linearly separable with respect to the grammatical genders of the nouns. Such statement is supported by the semantic space of neural network in Figure 1. The uter nouns (black) and the neuter nouns (red) are forming two clusters which only overlap at a small area. As expected, this intermediary zone is precisely where most of the errors generated by the neural network (green and blue) are located. Moreover, the classifiers had more difficulties identifying neuter nouns. For instance, neural network could interpret uter nouns with higher f-score (95.6%) compared to neuter nouns (88.7%).

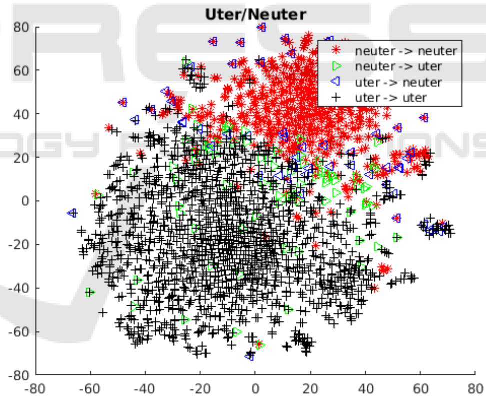

Figure 1: tSNE representation of the word vectors with regard to their grammatical genders associated predicted by the neural network.

Figure 2 shows the histogram of the entropy of the neural network's outputs. The top left histogram, the neuter nouns are classified as neuter, and the bottom right histogram, the uter nouns are classified as uter, are skewed toward left. This left skewness in the histograms show that the classifier predicts the correct grammatical genders with high confidence. This confirm that the information about the grammatical genders of the nouns are captured by the word vectors. We also see that the histogram of the entropy of the neural network's outputs for the erroneous items, i.e., the top right and the bottom left graphs, are skewed toward right. This displays the uncertainty involved in the neural network's outputs and indicates the lack of information in the erroneous word vectors. Thus, the analysis of the output's entropy demonstrate that with regard to grammatical gender, the neural network was interpreting the grammatical gender of nouns with high accuracy, with exception to some outliers for which the entropy was unusually high. Further explanation is provided in the following Section.

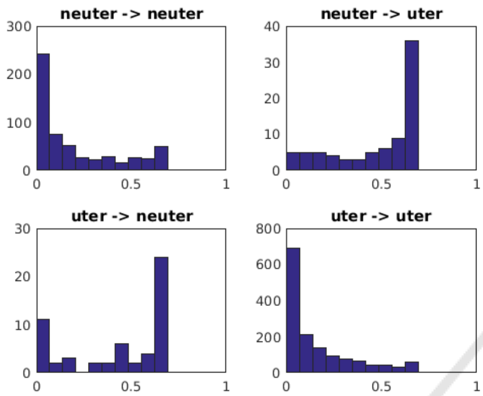

Figure 2: The histogram of the entropy of the neural network's outputs with regard to grammatical genders.

5.2 Common/Proper Nouns

Table 3) summarizes the classification results obtained from both LDA and the neural network to the distinction between common and proper nouns. We see that both classifiers outperform our baseline accuracy (85.5%).

Table 3: The performance of LDA and neural network (NN) on the common noun versus proper noun distinction.

	LDA	NN
accuracy	93.4%	95.2%
common noun recall	96.8%	98.9%
common noun precision	95.5%	93.9%
common noun f-score	96.1%	96.3%
proper noun recall	73.4%	62.4%
proper noun precision	79.9%	91.0%
proper noun f-score	76.5%	74.3%

In terms of accuracy, the neural network is more accurate than LDA. However, most of this achievement is due to the unbalance data distribution which biases the neural network toward the bigger population. In term of f-scores, we see that LDA is more accurate than the neural network in identifying the proper nouns and it is as accurate as the neural network in identifying the common nouns.

It is also worth noting that LDA is much more efficient than neural network when the processing time is of importance. The high accuracy obtained from LDA show that the word vectors should be linearly separable with regard to this nominal classification. We support this by the visualisation of the word vectors with regard to their proper noun versus common noun categories (see Figure 3).

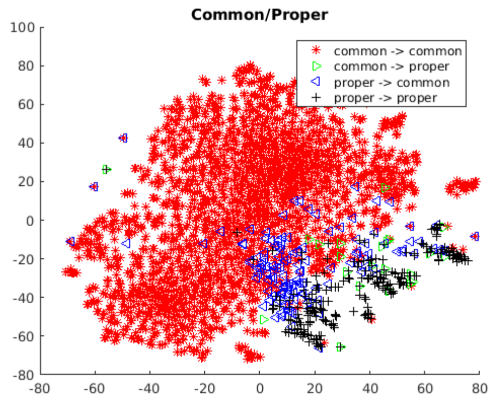

Figure 3: tSNE representation of the word vectors associated with the proper noun and common noun categories by the neural network.

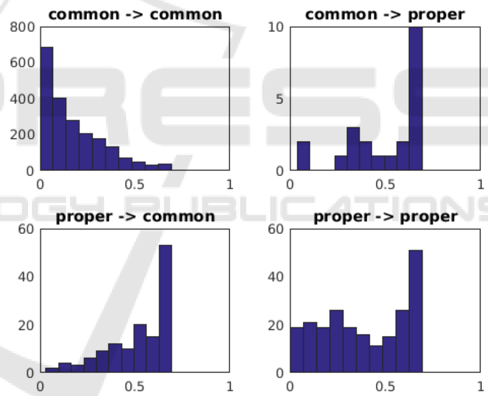

Figure 4: The histogram of the entropy of the neural network's outputs with regard to the proper noun versus common noun classification.

As shown, a certain part of proper nouns are merged into the common nouns, thus, hard to identify. In this case, the recall of the proper nouns is significantly improved if we use quadratic discriminant analysis (QDA) instead of the linear discriminant analysis. However, this improvement is together with a dramatic decrease on the recall of the common nouns. In our experiments with QDA, we obtained the f-score of 93.3% on the common nouns and the f-score of 71.4% on the proper nouns which are smaller than LDA and the neural network.

As an additional evidence, the histogram of the entropy of the neural network's outputs in Figure 4 shows less confidence in the outputs of the neural network when it faces with proper nouns in its inputs. In comparison with Figure 2, we see higher amount of uncertainty in the classifier's outputs which confirms the lack of information in the word vectors with regard to the common/proper noun classes.

5.3 Count/Mass Distinction

As for the classes of count and mass nouns, the classifiers did not perform as well as the two other tasks. Table 4 summarizes the results obtained from the count/mass distinction. As shown, the accuracy obtained from both classifiers are below our baseline, 85.5%. As we mentioned before, our baseline uses the Zero rule for classification which simply predicts the majority class, i.e., count noun in this case. So, it will be completely unable to predict the minority class of mass nouns.

Table 4: The performance of LDA and neural network (NN) on the count noun versus mass noun distinction.

	LDA	NN
accuracy	74.8%	82.8%
count noun recall	74.3%	99.0%
count noun precision	93.7%	83.2%
count noun f-score	82.9%	90.4%
mass noun recall	76.9%	7.2%
mass noun precision	39.1%	61.4%
mass noun f-score	51.8%	12.7%

The baseline's recall for the mass nouns, however, is 0.0% since all the mass nouns are classified as count noun. This is similar to the results obtained from the neural network, resulting in the high precision 99.0% on the count nouns but very small precision on the mass nouns. This basically shows that the neural network's prediction on the mass nouns is almost always wrong, i.e., mass nouns are always classified as count noun by the neural network. However we see that the LDA's recall on the mass nouns (76.9%) is significantly higher than the baseline's recall (0.0%) and the neural network's recall (7.1%). This is because of the generative nature of LDA that always give a chance to all classes, regardless of their size, to appear in the prediction task. However, this might not always work well specially when the data is not properly distributed with regard to the classes. In this case, the classifier will result in a small value of precision, as we see for LDA's precision on the mass nouns. In general, the weak performance of the classifiers on the count/mass classification task shows that the word vectors have almost no information about this feature of the nouns. This is because of the migration of words between the two classes (see Section 2.3) which allows a noun to be count or mass depending on its syntactic environment. This dynamic behavior of nouns results in a big overlap between the two classes of word vectors associated with count and mass nouns and makes the word vectors ineffective for this task. The overlap between the two classes of word vectors is seen in Figure 5. We see that mass nouns and the count nouns appear everywhere in the semantic space.

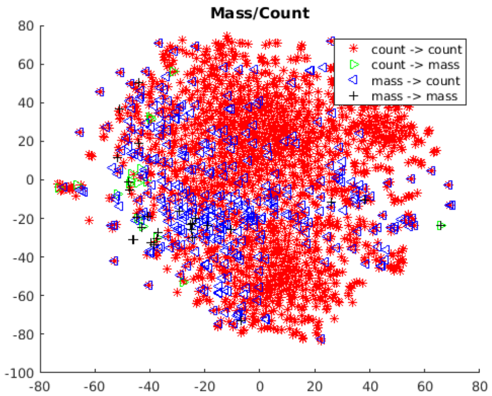

Figure 5: tSNE representation of the word vectors associated with the mass nouns and count nouns by the neural network.

Such observation is further supported by the histogram of the entropy of the neural network's outputs in Figure 6. First of all, with regard to the tasks of uter/neuter and common/proper classes, the neural network is highly certain about the correctly classified tokens (see Figure 2 and Figure 4).

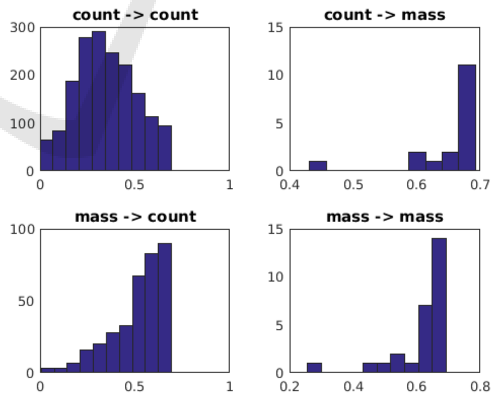

Figure 6: The histogram of the entropy of the neural network's outputs with regard to the proper noun versus common noun classification.

However, the neural network does not perform as well on count/mass categorization. We see that the classifier is very uncertain about the identification of count versus mass nouns. This shows that the classifier could not find any relevant information in the distribution of word vectors and its decisions rely highly on chance. This uncertainty is seen is the four scenarios shown in Figure 6. We observe that the amount of uncertainty involved in the classifier's decisions on the mass nouns that are correctly classifies as mass noun is higher than the count nouns that are correctly classified as count noun. This because the likelihood of migration (shift) from mass category to count category is higher than the migration from the count category to mass category (Gillon, 1999, p.57-58).

6 DISCUSSION

Our research question was to evaluate the performance of word embedding on the nominal classification tasks, uter/neuter, proper/common, count/mass noun distinctions. Through our preliminary analysis, we deduce that word vectors are satisfactory for uter/neuter and proper/common noun distinctions, but they show weak performance on the count/mass nominal classification. Nonetheless, the classifiers still encountered difficulties when categorizing the three nominal classes in our experiment. Hence, we provide a preliminary error analysis of the erroneous output of the classifiers. Due to ranking of performance and space limitation, we only scrutinize the wrongdoings of the neural network. As a preparatory analysis, we do not provide a quantitative error analysis, we rather describe qualitatively the main types of errors attested. Further experiments and statistic measurements are planned in future studies to investigate the recurrent frequency of different types of errors generated by the classifiers.

Our analysis shows that most errors are due to polysemy, i.e., the coexistence of several meanings for a unique word form. For instance, one frequently observed type of error is due to participles which can be used both as adjectives and nouns. By way of illustration with the task of uter/neuter distinction, the word *sovande* 'sleeping' may refer to the action of sleeping as a neuter noun in (1).

(1) *det går att bryta cirkeln*
it.NEUT go.PRS to break.INF circle.DEF.UTER
av dåligt sovande
of bad.NEUT sleeping
'It is possible to break the circle of bad sleeping.'

However, it can also be interpreted as an adjective in (2). Such situation results in incoherent context information when building the word vectors. Thus, the classifier equivalently encounters difficulties to identify the correct class of the noun.

(2) *brasilien kan inte fortsätta att vara*
Brazil can not continue.INF to be.INF
en sovande jätte
one.UTER sleeping giant
'Brazil can not continue to be a sleeping giant.'

(2) and (1) further demonstrates the source of confusion for the word vectors via the different articles leading the target word. In (1), neuter agreement is reflected on the adjective 'bad'. However, in (2), *sovande* is preceded by the uter indefinite article *en* which actually refers to the next uter noun *jätte* 'giant'. Such situation are expected to occur frequently as the majority of nouns in Swedish are affiliated to the uter gender. Since the parameters of word embeddings were set to include the immediate preceding word of the target, polysemous neuter nouns are more likely to be erroneously identified as uter nouns. This is exactly what we observed in Section 5.

Furthermore, polysemy may also occur with regard to proper and common nouns. As an example in (3), the word *springer* may refer to the name of a person, e.g., *Axel Springer*. As a reminder, capitalization effect was removed from our corpora. Thus, the word embeddings could not simply retrieve the information relevant to common and proper nouns by identifying the starting letter of the encountered nouns.

(3) *axel springer är känd som ägare*
Axel Springer be.PRS known as owner
till bild som är tysklands och
to Bild which be.PRS Germany.POSS and
europas största tidning.
Europe.POSS biggest magazine
'Axel Springer is known as the owner of Bild, which is Germany's and Europe's largest magazine.'

Nonetheless, the word form *springer* can also carry the meaning of a verb, i.e., 'run', e.g. in (4). Once more, the verb interpretation of *springer* is more frequent than the reference proper noun. Thus, the word vectors are disrupted when considering the syntactic and semantic context of *springer* as a noun, since it mistakenly involves the syntactic and semantic contexts of *springer* as a verb.

(4) *det är andra året*
it.NEUT be.PRS second year.DEF.NEUTER
som jag springer här
which I run.PRS here
'That's the second year I'm running here.'

Moreover, (3) is in fact a double example of polysemy. As displayed in (5), the word *bild* is usually a noun, i.e., ‘picture’. Yet, *bild* is employed as a proper noun in (3) when referring to the name of a magazine. Hence, such examples demonstrate the ease of category shift for the same word form.

(5)	*en*	*gammal*	*bild*	*av*	*barn*
	one.UTER	old.UTER	picture	of	children

‘An old picture of children.’

Finally, polysemy likewise occurred in terms of count/mass classification. As shown in (6), *tequila* can refer to tequila as a category of liquid, thus a mass. In such situation, the noun is expected to occur without indefinite articles or plural formation.

(6)	*samtliga*	*kan*	*göras*	*med*	*vodka*	*eller*	*tequila*
	all	can	done	with	vodka	or	tequila

‘All can be done with vodka or tequila.’

However, as explained in Section 2.3, mass nouns regularly undergo conversion to count nouns. As displayed in (7), the noun *tequila* is considered as a count noun since it does not refer to the liquid tequila, but rather to a shot of tequila, which can be counted. Hence, *tequila* as a count noun can be preceded by the indefinite article *en*.

(7)	*han*	*får*	*börja*	*med*	*att*	*berätta*	*hur*
	he	get	begin.INF	with	to	tell.INF	how

mycket	*han*	*druckit.*	*ett*	*glas*
much	he	drink.PERF	one.NEUT	glass

vin	*och*	*en*	*tequila.*
wine	and	one.UTER	tequila

‘He gets to start with telling how much he drank. A glass of wine and a tequila.’

Conversion is extremely productive between the count/mass classes of nouns (Gillon, 1999; Doetjes, 2012). Hence, it creates additional difficulties to the word embedding models. Therefore, we may deduce that static syntactic and semantic nominal features such as grammatical gender (uter/neuter) and the common/proper nouns differentiation are more generally interpretable by word embedding. However, cases of polysemy represent a challenge to word embedding, which may require additional tuning to reach high precision, c.f. the accuracy of the count/mass distinction was lower since the count/mass category is more versatile. Such results are in accordance with findings from linguistic and psycholinguistic studies. The brain differentiates count and mass nouns in terms of syntax and semantics. However, “one-to-one mappings between mass-count syntax and semantics is not supported by empirical findings” (Barner and Snedeker, 2005; Chiarelli et al., 2011).

7 CONCLUSIONS

Through the application of word embedding (Basirat and Nivre, 2017) with various classifiers such as linear discriminant analysis (LDA) and feed-forward neural network (NN), we are able to demonstrate that some types of nominal features can be captured by word embedding models. We show that both grammatical and semantic properties of nouns may be identified correctly through word embedding. However, we equivalently point out the importance of polysemy with regard to classification task. By way of illustration, the count/mass categorization represented difficulties for the classifiers not only due to its complex grammatical and semantic parameters, but also for its high occurrences of conversion (shift) across nouns. Such issues do incarnate complication for word embeddings. Thus, the application of word embeddings is considered as adequate for static nominal features such as grammatical genders or semantic features, but less appropriate for fluctuating properties such as count/mass. The high accuracy of LDA on the nominal classification task shows that the word vectors are almost linearly separable with regard to the nominal classes.

The limitations of our study include a lack of diversity in terms of data. For instance, we only included one language (Swedish) in our analysis. It would be necessary to enlarge the sampling and run the experiment on a phylogenetically weighted group of languages. Furthermore, we solely selected one model of word embeddings, i.e., Real-valued Syntactic word Vectors. More existing architectures such as GloVe and word2vec should be involved in our experimentation to further testify the pros and cons of word embeddings. Finally, we identified the issue of polysemy with regard to classification tasks via linear word embedding models. Thus, the next step would be to undergo similar functions with different types of vector generator, e.g., dependency parsing, and compare their respective performance.

ACKNOWLEDGEMENTS

Our work on this paper was fully collaborative; the order of the authors names is alphabetical and does not reflect any asymmetry in contribution. We are grateful for the fruitful discussion with the audience of the Special Session on Natural Language Processing in Artificial Intelligence at the 10th International Conference on Agents and Artificial Intelligence in Funchal, Madeira. We would like to express our gratitude to our colleagues Linnea Öberg, Karin Koltay, Rima Haddad, and Josefin Lindgren for their comments and support. We also appreciate the constructive comments from the anonymous referees. We are fully responsible for any remaining errors.

REFERENCES

Aikhenvald, A. Y. (2000). *Classifiers: A Typology of Noun Categorization Devices*. Oxford University Press, Oxford.

Aikhenvald, A. Y. (2012). Round women and long men: Shape, size, and the meanings of gender in New Guinea and beyond. *Anthropological Linguistics*, 54(1):33–86.

Andersson, A.-B. (1992). *Second language learners' acquisition of grammatical gender in Swedish*. PhD dissertation, University of Gothenburg, Gothenburg.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern information retrieval: The concepts and technology behind search*. Addison Wesley Longman Limited, Essex.

Baldwin, T., Paul, M., and Joseph, A. (2009). Restoring punctuation and casing in English text. *Proceedings of the 22nd Australian Joint Conference on Artificial Intelligence (AI09)*, pages 547–556.

Barner, D. and Snedeker, J. (2005). Quantity judgments and individuation: Evidence that mass nouns count. *Cognition*, 97:41–66.

Basirat, A. and Nivre, J. (2017). Real-valued Syntactic Word Vectors (RSV) for Greedy Neural Dependency Parsing. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa*, pages 21–28, Gothenburg. Linkping University Electronic Press.

Bohnacker, U. (2004). Nominal phrases. In Josefsson, G., Platzack, C., and Hkansson, G., editors, *The acquisition of Swedish grammar*, pages 195–260. John Benjamins, Amsterdam.

Borer, H. (2005). *Structuring Sense, part I*. Oxford University Press, Oxford.

Borin, L., Forsberg, M., and Lnngren, L. (2008). The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology. *Studia Linguistica Upsaliensia*, pages 21–32.

Carter, D., Kaja, J., Neumeyer, L., Rayner, M., Weng, F., and Wirn, M. (1996). Handling compound nouns in a Swedish speech-understanding system. *Proceedings of the Fourth International Conference on Spoken Language*, 1:26–29.

Cheng, C.-Y. (1973). Response to Moravcsik. In Hintikka, J., Moravczik, J., and Suppes, P., editors, *Approaches to natural language*, pages 286–288. D.Reidel, Dordrecht.

Chiarelli, V., El Yagoubi, R., Mondini, S., Bisiacchi, P., and Semenza, C. (2011). The Syntactic and Semantic Processing of Mass and Count Nouns: An ERP Study. *PLoS ONE*, 6(10):e25885.

Chierchia, G. (1998). Plurality of mass nouns and the notion of semantic parameter. In Rothstein, S., editor, *Events and grammar*, pages 53–104. Dordrecht, kluwer edition.

Chierchia, G. (2010). Mass nouns, vagueness and semantic variation. *Synthese*, 174(1):99–149.

Contini-Morava, E. and Kilarski, M. (2013). Functions of nominal classification. *Language Sciences*, 40:263–299.

Corbett, G. G. (1991). *Gender*. Cambridge University Press, Cambridge.

Corbett, G. G. (2000). *Number*. Cambridge University Press, Cambridge.

Corbett, G. G. (2013). Number of Genders. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dahl, O. (2000). Elementary gender distinctions. In Unterbeck, B. and Rissanen, M., editors, *Gender in grammar and cognition*, pages 577–593. Mouton de Gruyter, Berlin.

Delahunty, G. P. and Garvey, J. J. (2010). *The English Language: From Sound to Sense*. West Lafayette, parlor press edition.

Dixon, R. M. W. (1986). Noun class and noun classification. In Craig, C., editor, *Noun classes and categorization*, pages 105–112. John Benjamins, Amsterdam.

Doetjes, J. (2012). Count/mass distinctions across languages. In Maienborn, C., Heusinger, K. v., and Portner, P., editors, *Semantics: an international handbook of natural language meaning, part III*, pages 2559–2580. Mouton de Gruyter, Berlin.

Dryer, M. S. (2005). Coding of nominal plurality. In Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors, *The word atlas of language structures*, pages 138–141. Oxford University Press, Oxford.

Fraurud, K. (2000). Proper names and gender in Swedish. In Unterbeck, B., Rissanen, M., Nevalainen, T., and Saari, M., editors, *Gender in Grammar and Cognition*, pages 167–220. Mouton de Gruyter, Berlin.

Gillon, B. S. (1999). The lexical semantics of English count and mass nouns. In Viegas, E., editor, *Breadth and depth of semantic lexicons*, pages 19–37. Springer, Dordrecht.

Hall, K. (2002). Unnaturalgender in Hindi. In *Gender across Languages: The Linguistic Representation of Women and Men*, pages 133–162. John Benjamins, Amsterdam.

Katz, G. and Zamparelli, R. (2012). Quantifying count/mass elasticity. *Proceedings of the 29th West Coast Conference on Formal Linguistics*, pages 371–379.

Kemmerer, D. (2017). Categories of object concepts across languages and brains: the relevance of nominal classification systems to cognitive neuroscience. *Language, Cognition and Neuroscience*, 32(4):401–424.

Kilarski, M. (2014). The Place of Classifiers in the History of Linguistics. *Historiographia Linguistica*, 41(1):33–79.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49.

Nespital, H. (1990). On the relation of Hindi to its regional dialects. In Offredi, M., editor, *Language versus dialect: Linguistic and literary essays on Hindi, Tamil and Sarnami*, pages 3–23. Manohar Publications, New Delhi.

Nichols, J. (1989). The origin of nominal classification. *Proceedings of the fifteenth annual meeting of the Berkeley linguistics society*, pages 409–420.

Ostling, R. and Wirn, M. (2013). Compounding in a Swedish blog corpus. *Acta Universitatis Stockholmiensis*, pages 45–63.

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

Pelletier, F. J. and Schubert, L. K. (1989). Mass expressions. In Gabbay, D. and Guenther, F., editors, *Handbook of philosophical logic. Volume IV: Topics in the philosophy of language*, pages 327–408. Reidel, Dordrecht.

Petrovic , S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189.

Pokharel, M. P. (2010). Noun class agreement in Nepali. *Kobe papers in linguistics*, 7:40–59.

Preiss, J. and Stevenson, M. (2013). Distinguishing common and proper nouns. *Second joint conference on lexical and computational semantics: Proceedings of the main conference and the shared task: Semantic textual similarity*, 1:80–84.

Quine, W. v. O. (1960). *Word and object*. MIT Press, Cambridge.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850.

Seifart, F. (2010). Nominal classification. *Language and Linguistics Compass*, 4(8):719–736.

Senft, G. (2000). *Systems of nominal classification*. Cambridge University Press, Cambridge.

Teleman, U., Hellberg, S., and Andersson, E. (1999). *Svenska Akademiens grammatik. Vol. 2: Ord. [The Swedish Academy Grammar, Part 2: Words]*. Norstedts, Stockholm.

Ting, K. M. (2010). Precision and Recall. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning*, pages 781–781. Springer US, Boston, MA. DOI: 10.1007/978-0-387-30164-8_652.

Ullman, E. and Nivre, J. (2014). Paraphrasing Swedish Compound Nouns in Machine Translation. In *MWE@ EACL*, pages 99–103.