


Prepositions Matter in Quantifier Scope Disambiguation


Aleksander Leczkowski

Faculty of Psychology
University of Warsaw

ale.leczkowski@gmail.com 


Justyna Grudzińska

Faculty of Philosophy
University of Warsaw

j.grudzinska@uw.edu.pl 


Manuel Vargas Guzmán

Faculty of Philosophy
University of Warsaw

m.vargas-guzman@uw.edu.pl 


Aleksander Wawer

Institute of Computer Science
Polish Academy of Sciences

axw@ipipan.waw.pl 

Aleksandra Siemieniuk

Faculty of Polish Studies
University of Warsaw

aleksandra.siemieniuk@uw.edu.pl 

Abstract

Although it is widely agreed that world knowledge plays a significant role in quantifier scope disambiguation (QSD), there has been only very limited work on how to integrate this knowledge into a QSD model. This paper contributes to this scarce line of research by incorporating into a machine learning model our knowledge about relations, as conveyed by a manageable closed class of function words: prepositions. For data, we use a scope-disambiguated corpus created by AnderBois, Brasoveanu and Henderson, which is additionally annotated with prepositional senses using Schneider et al's Semantic Network of Adposition and Case Supersenses (SNACS) scheme. By applying Manshadi and Allen's method to the corpus, we were able to inspect the information gain provided by prepositions for the QSD task. Statistical analysis of the performance of the classifiers, trained in scenarios with and without preposition information, supports the claim that prepositional senses have a strong positive impact on the learnability of automatic QSD systems.

1 Introduction

QSD is a problem in natural language processing that arises in connection with sentences that contain multiple quantified NPs:

- (1) Every kid climbed a tree.

Sentence (1) can be understood to mean that every kid climbed a possibly different tree. This is the so-called surface scope reading where the first quantified NP has wider scope than the second,

corresponding to the surface ordering of the two NPs in the sentence: *every kid* > *a tree*. The other, and usually less preferred, reading is the one in which there is a single tree that all the kids climbed. This is the inverse scope reading where the second quantified NP has wider scope than the first, reversing the order of the two NPs in the sentence: *a tree* > *every kid*. Many studies on quantifier scope have dealt with the issue of generating the set of possible scope readings for a sentence like (1), both from a theoretical perspective (May, 1978; Cooper, 1983; May, 1985; Hendriks, 1993; Steedman, 2012; Barker and Shan, 2014) and computationally (Woods, 1987; Hobbs and Shieber, 1987; Bos, 1996; Copestake et al., 2001; Egg et al., 2001; Bos et al., 2004; Koller et al., 2008; Evang and Bos, 2013; Sayeed, 2016). A much smaller number of studies have focused on statistical and automatic QSD and the problem of identifying the set of factors relevant to scope preferences (Higgins and Sadock, 2003; AnderBois et al., 2012; Manshadi and Allen, 2011; Manshadi et al., 2013). These studies have shown, mostly in line with what was proposed in the semantics literature and borne out in psycholinguist work (Ioup, 1975; Micham et al., 1980; Gillen, 1991; Kurtzman and MacDonald, 1993; Tunstall, 1998; Anderson, 2004; Radó and Bott, 2011; Dotlačil and Brasoveanu, 2015; Capelier-Mourguy et al., 2015), that the grammatical role (i.e., subject and object) and lexical realization of a quantifier have an effect on scope-taking; linear precedence in a sentence has an effect as

well.¹

The above factors are certainly not sufficient to predict quantifier scope. It has been repeatedly stressed in previous work that world knowledge also plays a significant part in real world QSD, and any successful model for the QSD task should make use of it (Saba and Corriveau, 2001; Srinivasan and Yates, 2009; Manshadi and Allen, 2011; AnderBois et al., 2012; Tsiolis, 2020). To the best of our knowledge, however, Srinivasan and Yates (2009) have been unique in using a model explicitly geared towards world knowledge (in particular, numerical typicality) in the QSD task. Drawing on Saba and Corriveau (2001), they decided on the preferred scoping by comparing the size of two classes, e.g., *Person* and *City*, standing in a relation such as the *living-in* relation. For example, the surface scope reading is dispreferred in a sentence such as *A person lives in every city* because it would require a person to live in an atypically large number of cities. The present study seeks to contribute to this scarce line of research through incorporating into a QSD model our knowledge about relations, as conveyed by a manageable closed class of function words: prepositions.

Relations between objects (but also times, events) are often signaled with prepositions (Pustejovsky, 1991; Srikumar and Roth, 2013; Abzianidze and Bos, 2017; Schneider et al., 2018). Prepositions serve, among other things, to convey place and time (*There is a restaurant at every corner, John taught on each Monday*), to express configurational relationships like possession or part/whole (*someone with every key, a day of every month*), and to indicate semantic roles in predicate–argument structure like agent or instrument (*a study sponsored by a consumer group, a store filled with lots of food*). Recent work argues that certain prepositional senses are special in that they encode dependencies that have an effect on scope-taking (Grudzińska and Zawadowski, 2019, 2020). For example, the preposition *of* expressing ‘part-whole sense’ — as in *a day of every month* — introduces a dependency between each whole (month) and its respective parts (days). By quantifying over this dependency, we obtain the inverse scope read-

ing for the example in question: for every month, there is a different day that belongs to it (*every month > a day*). The surface scope reading (*a day > every month*) is excluded because of what we know about parts and wholes, namely that we can have many parts (days) belonging to the same whole (month), but a single part (day) cannot belong to more than one whole (month). Conversely, the ‘whole-containing-part sense’ of the preposition *of* — as in *a group of four homeowners* — encodes a dependency between a group and its respective members, thus only allowing surface scope (*a group > four homeowners*). Furthermore, universal quantification in locative and temporal prepositional phrases tends to support inverse scope. For example, the locative preposition *at* — as in *a restaurant at every corner* — implies ‘disjointness’ (objects do not occupy more than one place at a time), and hence can be interpreted as a dependency between each corner and the respective restaurants located at that corner. Quantifying over this dependency yields the inverse scope reading: for every corner, there is a different restaurant located at it. The surface scope reading is excluded because one restaurant cannot occupy more than one place (every corner) simultaneously.

Our study takes its theoretical inspiration from the above work and contributes to research on automatic QSD by examining the previously unexplored predictors of quantifier scope: prepositions and their senses. For the experiments undertaken in this study, we use a scope-disambiguated corpus created by AnderBois et al. (2012), additionally annotated with prepositional senses using the Semantic Network of Adposition and Case Super-senses (SNACS) scheme proposed in Schneider et al. (2018). Our results indicate that prepositional senses have a strong role in the QSD task and encourage further research and deeper analysis in this area. The structure of the paper is as follows. Section (2) introduces our corpus and discusses its annotation process. Section (3) explains the methodology of our study. In (4), we introduce our experimental setup and discuss the features we have used in our models. Section (5) presents our results and (6) concludes with a summary and some directions for future work.

2 Corpus

The present study uses a scope-disambiguated corpus which was created for the purposes of the 2012

¹The effect of linear precedence has been debated in previous works, with some authors arguing against it (Ioup, 1975; Micham et al., 1980; Kurtzman and MacDonald, 1993), and it needs to be further explored, especially in free word order languages with case marking (as is the case, e.g., in Sayeed et al. (2019)).

study by AnderBois, Brasoveanu and Henderson (2012). It consists of 680 sentences with multiple quantified NPs from the reasoning section of the Law School Admission Test — the so-called logic puzzles. Logic puzzles provide a good corpus for QSD for they use quantifiers frequently, providing a fair number of sentences containing scopally interacting quantifiers.

Every sentence of AnderBois et al.’s corpus is labeled with the relative scope of the quantified NPs involved. Scope is coded numerically, with **1** corresponding to widest scope and smaller numbers indicating narrower scope; cases with no relative scope like logical equivalence (e.g., two universals or two existentials) are co-tagged with the same number. The scope predictors incorporated into the annotation in the corpus include sentence order (it is not explicitly tagged, since it can be recovered from the linear order of the tags themselves), grammatical function (Subject, Object, Adjunct, etc.) and lexical realization of quantifiers. The beginning of the tag is marked by $\&$ and the end is marked by $\#$:

- (2) Hannah visits at least one $\&3_O_at.least.one\#$ city in each in each $\&2_in_each\#$ of the three of the three $\&1_of_the.three\#$ countries.

Since the sentences in the corpus were chosen for quantified NPs, they would be expected to provide no bias with respect to prepositions. The most common prepositions in English identified by Litkowski and Hargraves (2007) do indeed overlap in eight cases with those in the corpus, although in some cases the frequency distribution is different: *of*, *in*, *on*, *at*, *to*, *for*, *with*, and *from*. While individual prepositions in prepositional phrases are tagged separately in the corpus (as illustrated by example (2)), the prepositional senses are not. It is, however, prepositional senses (rather than prepositions) that induce or block inverse scope. For example, as discussed above, the ‘part-whole sense’ of the preposition *of* induces inverse scope, while its inverse ‘whole-containing-part sense’ blocks it. For our study, we additionally annotated the corpus with prepositional senses, using the Semantic Network of Adposition and Case Supersenses (SNACS) scheme proposed in Schneider et al. (2018; 2020).

2.1 Proposition-sense annotation

The SNACS scheme provides a hierarchy of 50 supersenses, divided into three main subhierarchies that loosely correspond to adverbial adjuncts, event arguments, and adnominal complements:

- CIRCUMSTANCE: TIME, LOCUS, MEANS, MANNER, PATH, ...
- PARTICIPANT: AGENT, THEME, RECIPIENT, BENEFICIARY, INSTRUMENT ...
- CONFIGURATION: WHOLE, ORG, QUANTITYITEM, POSSESSION, STUFF ...

Furthermore, the scheme deploys the construal analysis proposed in Hwang et al. (2017), i.e., it introduces a distinction between a SCENE ROLE (marked by SS), which expresses the preposition’s meaning in context, and a FUNCTION (marked by SS2), which denotes the preposition’s lexical meaning. Both SCENE ROLE and FUNCTION are drawn from the supersense hierarchy and are often identical. The SNACS scheme was applied to prepositions in the STREUSLE corpus, a collection of online consumer reviews taken from the English Web Treebank (Bies et al., 2012). Each preposition token in the STREUSLE corpus is annotated with SS and SS2 (SS \rightsquigarrow SS2):

- (3) Dan arrived at 10 am. TIME \rightsquigarrow TIME
 (4) The team at Max’s is great. ORG \rightsquigarrow LOCUS

In example (3), the preposition *at* is unambiguously temporal — SS and SS2 are congruent. In example (4), there is an overlap between organizational belonging meaning (marked by ORG) and locational meaning (marked by LOCUS) — SS and SS2 differ. The construal analysis helps with cases where multiple supersenses seem to fit and contributes to reducing disagreement among annotators, who are not forced to pick a single label in cases of meaning overlap.

Our scope-disambiguated corpus has been annotated with prepositional senses by three annotators, all non-native speakers with linguistic training. The annotators were familiar with the annotation manual — guidelines for English including description of the 50 supersenses, with examples and criteria for borderline cases (Schneider et al., 2020). Across all targets, there was good agreement on SS between the three annotators, $k = .68, p = .000$,

and there was very good agreement on ss2 between the three annotators, $k = .79, p = .000$. Agreement was higher on the function slot than on the scene role slot. This is expected considering the fact that the function of a preposition reflects its prototypical and more stable meaning, whereas the scene role depends on context and can vary more from person to person. Our results are only slightly lower than the SNACS IAA numbers found in Schneider et al. (2018) ($k = .73$ and $k = .80$). Our agreement is so strong most likely due to the simple literal language of the logic puzzles. The remaining differences were adjudicated in meetings involving all of the three annotators.

2.2 Scope annotation

The number of tagged quantified NPs in a sentence ranges from two to eight in AnderBois et al.’s corpus. Manshadi and Allen (2011, 2013) developed a method that can deal with an arbitrary number of quantifiers per sentence in the QSD task. They define the task as learning to build a partial order over the set of quantifiers in the sentence. In adapting the scope coding in AnderBois et al.’s corpus to Manshadi and Allen’s method, we thus consider three relations between each pair of quantifiers $\{q_1, q_2\}$, with q_1 occurring before q_2 in a given sentence: *wide scope* ($q_1 > q_2$), *narrow scope* ($q_2 > q_1$) and *incomparability* (q_1, q_2). The three relations are used in order to determine the quantifier scoping of each sentence from the corpus, based on the relative scopings provided. Determining the scopes of the tagged quantifiers in example (2) is straightforward. The third quantifier outscopes the second ($q_3 > q_2$); the second one outscopes the first one ($q_2 > q_1$). Moreover, since outscoping is a transitive relation, the third one also outscopes the first one ($q_3 > q_1$). Hence, the formula describing the sentence’s scoping looks as follows: $q_3 > q_2 > q_1$. Each of the 680 sentences in the corpus is annotated following that method. According to the formula $\sum_i n_i * (n_i - 1) / 2$, where n_i denotes the number of quantified NPs in a sentence, there are 1451 relations between quantifiers in the corpus.

3 Method

As mentioned above, Manshadi and Allen build their method on the fact that quantifier scopings (QS) form partial orders. Hence, they define QSD as a task of creating partial orders and show that

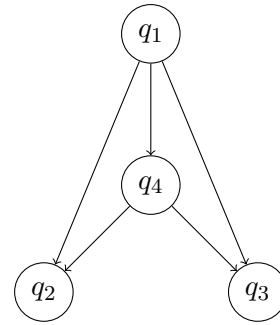


Figure 1: TDAG representing quantifier scopings in example (5): $q_1 > q_4 > q_2, q_3$.

it is equivalent to a pairwise comparison problem (see Manshadi (2014) for definitions and proofs).

3.1 Manshadi and Allen’s approach

Partial orders can be represented as Directed Acyclic Graphs (DAGs). In fact, since outscoping is a transitive relation, Transitive Directed Acyclic Graphs (TDAGs) have a one-to-one correspondence with quantifier scopings — each has exactly one TDAG representing it. Hence, every sentence’s QS is analysed in its transitive closure form and TDAGs are used for visualisation purposes.

Figure 1 depicts a TDAG which is a representation of a typical, for the examples in the corpus, quantifier scoping: $q_1 > q_4 > q_2, q_3$.² A sentence from the corpus which is defined by this order is provided in example (5).

- (5) Each&1_S_each# member of the Kim family sits in a&3_in_a_Locus_Locus# seat adjacent to, and in the same row&3_in_the.same_Locus_Locus# as, at least one other&2_as_at.least.one.other_ComparisonRef_ComparisonRef# member of the family.

Since the QSD task is reduced to a problem of pairwise comparisons, a sentence containing n quantifiers results in $n * (n - 1) / 2$ samples. There are four quantified NPs in example (5) which results in six different observations for the classifier. For each pair of quantifiers (an observation), a classifier has to predict one of three relations: *wide scope*, *narrow scope* or *incomparability*. From the perspective of a TDAG (G), those relations are defined as follows for every pair $\{q_1, q_2\}$, where q_1 precedes q_2 in a given sentence:

²This notation is equivalent to $q_1 > q_4 > q_2$ and $q_1 > q_4 > q_3$.

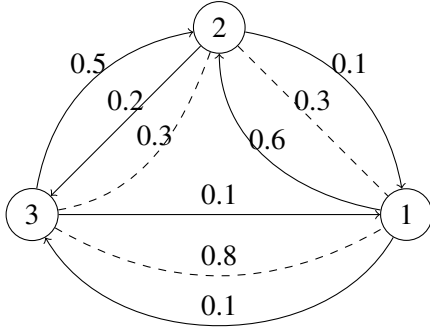


Figure 2: Example of a preference graph (Manshadi, 2014, p. 136).

1. *wide scope* (+) if $(q_1, q_2) \in G$
2. *narrow scope* (−) if $(q_2, q_1) \in G$
3. *incomparability* (ϵ) otherwise

That is, if two quantifiers q_1 and q_2 are characterised by *wide scope*, there is a directed edge from q_1 to q_2 . In the case of *narrow scope*, a directed edge goes from q_2 to q_1 . *Incomparability* is represented by the lack of an edge between two nodes.

A ternary soft classifier³ predicts probabilities for each observation, for each of the three possible classes. Once those probabilities are predicted, a preference graph for each sentence’s scoping can be built, as in Figure 2.

The goal at this point is to find a subgraph of that preference graph which satisfies the following condition: that it maximizes the sum of weights with the constraint that the resulting subgraph is a TDAG. The algorithm that performs this task is presented in the following section, and it is based on finding an approximately optimal ordering (Cohen et al. 1999).

3.2 Approximation algorithm

Let $(u, v)^+$, $(u, v)^-$ and $(u, v)^\epsilon$ be the probabilities that the nodes u and v are in a wide scope, narrow scope, and incomparability relation, respectively. The algorithm takes a preference graph Γ_p and stores its vertices in a set V . The difference between outgoing and incoming edges for each vertice is computed (lines 3 and 4) and the highest value is selected (i.e., the node with the widest scope) to store it as t (line 5). From lines 6 to 8, a

³Note that a hard classifier cannot be used here as then there is no guarantee that the resulting predicted graph will be either acyclic or transitive.

Algorithm 1 Creates a TDAG

Input: a preference graph Γ_p
Output: a transitive directed acyclic graph G

```

1:  $V \leftarrow \text{get\_vertices}(\Gamma_p), r \leftarrow 0, G \leftarrow \emptyset$ 
2: while  $V$  is non-empty do
3:   for each  $u \in V$  do
4:      $\pi(u) \leftarrow \sum_{v \in V} (u, v)^+ - \sum_{v \in V} (u, v)^-$ 
5:    $t \leftarrow \text{argmax}_{u \in V} \pi(u)$ 
6:   if  $\exists v \in G : \rho(v) = r$  and  $(v, t)^+ > (v, t)^\epsilon$  then
7:      $r \leftarrow r + 1$ 
8:    $\rho(t) \leftarrow r$ 
9:   for each  $v \in G$  do
10:    if  $\rho(v) < r$  and  $(v, t)^+ > (v, t)^\epsilon$  then
11:       $G \leftarrow G \cup \{(v, t)\}$ 
12:    $V \leftarrow V - \{t\}$ 
13:    $G \leftarrow G \cup \{t\}$ 
14: end while

```

rank (starting from 0) is assigned to t , the algorithm checks before if there is a node with the current rank that outscopes t , in which case the rank is incremented by one. From lines 9 to 11, edges (v, t) are added to the final graph G by checking all v nodes from previous ranks that have a wide scope relation with t . Finally, t is removed from V and added to G . The process repeats until V is empty.

4 Experimental setup

A Support Vector Machines (SVM) classifier, Python’s scikit-learn implementation (Pedregosa et al., 2011), was fitted to the data ($n = 1451$) in order to predict probabilities of three different relations between each pair of quantifiers: *wide scope*, *narrow scope* or *incomparability*. Once the probabilities were predicted by the classifier, in order to restore a full sentence’s quantifier scoping, a predicted TDAG was built.⁴

4.1 Features

A small set of features was selected for the purpose of the experiment: only those that were manually annotated in the corpus or could be computed in a simple manner. Listed below are the extracted features, each with a brief explanation:

- *Quantifier lexicalization* — quantifier lexicalizations are combined into groups in order to limit the dimensionality of this feature. For instance, all bare numerals are grouped together, all exactly-modified numerals (e.g., *exactly one*) are combined together, superlative

⁴Therefore, the breakdown of data was made at the sentence level and not at the observation level. Otherwise, this could result in observations from the same sentence being placed in both the training and test set, which would not allow restoring sentence’s quantifier scoping.

and comparative modified numerals (e.g., *at least/most three* and *more/less than three*) are assigned to one group, and so on.

- *Complex* — a binary feature that denotes whether a quantifier lexicalization consists of one token (e.g., *one*) or more than one token (e.g., *more than one*).
 - *Grammatical function* — whether a tagged NP plays the role of, for instance, a *subject* or an *object*.
 - *Appositive* — a binary feature which denotes whether a tagged NP is followed by an appositive (e.g., *four people - Grace, Heather, Josh, and Maria*).
 - *Prepositions* — preposition lexicalizations or preposition supersenses depending on the system (see Section 4.4 for an explanation). We focus on SUPER SS2 and SUPER SS \rightsquigarrow SS2 combinations only, i.e., we drop the less stable and more idiomatic SUPER SS.
- *Distance* — a gap between a pair of quantifiers in a given observation. For instance, if a sentence has three tagged quantificational expressions q_1 , q_2 and q_3 , occurring in that order in the sentence, then the distance between q_1 and q_2 equals 1 and the distance between q_1 and q_3 equals 2.

Since an observation is a pair of quantifiers, each feature⁵ was defined twice for a given observation.

Linear precedence, a much-discussed predictor of quantifier scope, is not provided here as a separate feature. It is inherently encoded due to the manner in which the task is formulated, as each observation is a pair of quantifiers $\{q_1, q_2\}$, where q_1 occurs before q_2 in a given sentence.

Feature selection was performed using the Mutual Information (MI) measure. First, all of the features with MI equal to zero were deleted. In fact, this led only to the removal of features that were duplicated as a result of defining an observation as a pair. For instance, supersense (SS2) ENDTIME occurs only once in the data and only as a property of the first quantifier in an observation — denoted as ENDTIME_1; hence, feature ENDTIME_2 was deleted. Second, features occurring

⁵Except for the feature *distance* which is a property of the relation between quantifiers, not a property of a quantifier itself.

only once in the corpus were deleted as well; as a result, ENDTIME_1 was also removed.

There are 27 different prepositions, 26 different SS2 supersenses and 67 different SS \rightsquigarrow SS2 combinations in the corpus. Since each observation is a combination of two quantifiers, these numbers correspond to 54, 52 and 134 different columns in the feature vector. After the feature selection, we get 39, 41 and 82, respectively.

4.2 Training and optimization

Training and optimisation were performed using nested cross-validation. Hyperparameter selection was executed in the inner loop using the 5-fold technique. Kernel, among other SVM’s hyperparameters, was considered in the optimization process and selected from *linear*, *polynomial*, *rbf* and *sigmoid*. The outer loop was repeated 30 times with different random data splits — Monte Carlo cross-validation. This way a standard 20 percent of the data was used in both inner and outer loops for the purpose of the validation of the models and the final results are an average of 30 independent runs.

4.3 Evaluation

Three different evaluation metrics, adapted to the QSD task by Manshadi and Allen (2011), were used in order to assess the performance of the models and all three of them, similarity, precision and recall, are based on the notion of the similarity of two graphs which represent gold $G_g = (V, E_g)$ and predicted $G_p = (V, E_p)$ sentence’s quantifier scopings. Let $G^+ = (V, E^+)$ be the transitive closure of the graph and $\bar{G} = (V, \bar{E})$ be the complement of the undirected version of G , where V denotes the set of nodes and E corresponds to the set of edges. The most general one of the three, the similarity metric (Equation 1), was used for hyperparameter selection during the optimization process as well as for the purpose of statistical testing.

$$\sigma^+ = \frac{|E_p^+ \cap E_g^+| \cup |\bar{E}_p^+ \cap \bar{E}_g^+|}{|V|(|V| - 1)/2} \quad (1)$$

The similarity measure treats outscoping and incomparability relations equally. In practice, it is the outscoping relation that should be most important in classification. That is because if the outscoping relation is mislabeled, it leads to a different interpretation of the sentence. Hence, Manshadi and Allen also adapt to the task a form of precision (Equation 2) and recall (Equation 3) which

are based on the number of outscoping relations identified correctly.

$$P^+ = \frac{|E_p^+ \cap E_g^+|}{|E_p^+|} \quad (2)$$

$$R^+ = \frac{|E_p^+ \cap E_g^+|}{|E_g^+|} \quad (3)$$

One might point out that precision or recall should be the metric selected to assess the model’s performance during optimisation and to report final results. Note, however, that there are a number of sentences in our corpus where the incomparability relation is the dominant or only relation present. Sentences defined by QS that only consist of incomparability relations are not considered in the computation of precision and recall. Hence, even though informative, those metrics do not result in a comprehensive evaluation.

4.4 Models

Four different training scenarios are conducted in order to assess the impact of prepositions and preposition supersenses on the learnability of the QSD system:

- **BASELINE** — models trained using all of the features defined, except information about prepositions.
- **PREP** — models trained using all of the features defined, including preposition lexicalizations, but not preposition supersenses.
- **SUPER SS2** — models trained using all of the features defined, including preposition supersenses (SS2 only), but except preposition lexicalizations.
- **SUPER SS \rightsquigarrow SS2** — models trained using all of the features defined, including preposition supersenses (SS \rightsquigarrow SS2 combinations), but except preposition lexicalizations.

The performance of these four systems allows us, first, to study the effect of preposition information on the ability of a system to learn a QSD task and, second, to assess whether this impact is better captured when provided with the SNACS supersense hierarchy.

An additional baseline pseudo-model — **WIDE** — is presented as a reference. It always predicts the most frequent label in the training set.⁶

⁶In total, out of 1451 observations, 307 represent *narrow scope*, 828 *wide scope* and 316 *incomparability*.

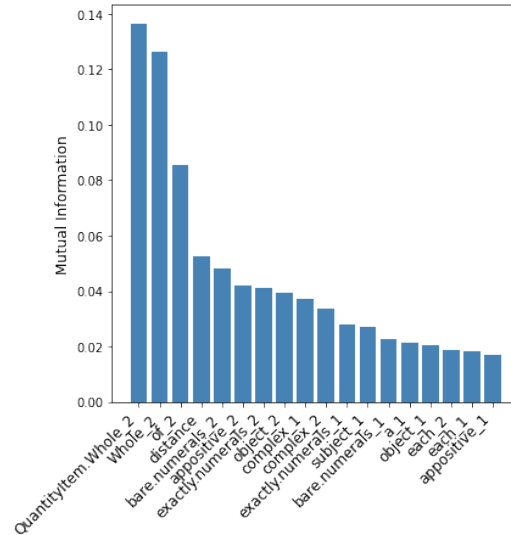


Figure 3: Most significant features according to the Mutual Information analysis.

5 Results⁷

5.1 Feature Importance

Figure 3 presents the 18 most informative features. Both the grammatical roles of *subject* and *object* and certain lexical realizations of quantifiers (*each*, *a*, bare numerals and exactly-modified numerals) rank high in the results, in line with previous findings. One other feature related to lexical realization, *complex*, also ranks high. The feature *appositive* signals the referential function of the NP to which it is related. Its high ranking is in line with the well-known fact that referentially used NPs tend to take the widest scope possible. Notably, the preposition *of* (when present in the second tagged NP in a given observation) ranks third. As expected, the ‘part-whole sense’ of the preposition *of*, marked by **WHOLE** and the corresponding role-function combination **QUANTITYITEM \rightsquigarrow WHOLE**, appears to be even more informative, ranking the highest.

Based on previous findings, one might expect the features of *subject* and *object* to rank higher in the analysis. However, previous studies (e.g., [AnderBois et al. \(2012\)](#)) that reported grammatical function to have a strong impact on scope-taking only focused on sentences with two quantified NPs and did not consider the incomparability relation between quantifiers. Thus, experimental setups previously employed might have inflated the role

⁷The information and code necessary to replicate the results reported in this section are available at the GitHub link: https://github.com/ALeczkowski/prep_matter_in_qsd

of *subject* and *object* in the QSD task.

5.2 Experiments

The results of the experiments are presented in Table 1. The first observation is that standard deviation is substantial in the case of each model and each metric. This is a result of the diverse data set (e.g., sentences with 2 vs. 8 quantified NPs) and shows the importance of reporting final results as a mean over a significant number of runs (in this case, 30). Second, the WIDE pseudo-model that classifies each observation as *wide scope*⁸ achieves significant results, e.g., the similarity measure equals 64.82, which is a result of the domination of that relation in the data. Third, in the case of each model, precision and recall are higher than the similarity metric. This contradicts expectations and is also different than in Manshadi and Allen’s experiments, where the opposite is the case. It appears that models perform classification better with respect to outscoping relations than when it comes to *incomparability*.⁹ Last but not least, by looking at raw numbers, it may be noticed that adding prepositions to the feature vector which is fed to the classifier improves the performance of that classifier on all metrics. Adding preposition supersenses, instead of preposition lexicalizations, results in further improvement, both in the case of SS2 and SS \rightsquigarrow SS2 combinations. However, tests need to be performed in order to determine whether those differences are statistically significant. Since SUPER SS2 is characterised by both better performance and lower standard deviation than SUPER SS \rightsquigarrow SS2,¹⁰ only the former is selected for statistical testing.

Since, when it comes to the similarity metric, homogeneity of variance is present in the three compared groups (Bartlett test; $p = 0.31$) and each group’s results are normally distributed (Shapiro-Wilk test; $p = 0.82, 0.29$ and 0.32 for BASELINE, PREP and SUPER SS2, respectively), one-way ANOVA is performed to determine whether there are any statistically significant differences between compared systems. The test is statistically signif-

⁸In each of 30 different data splits, *wide scope* was the most frequent relation in the training set.

⁹One possible explanation of that fact might be that no dependency parser was used in the experiment. Hence, no information was extracted about, for instance, conjuncts occurring between quantifiers, which is a strong predictor of the incomparability relation.

¹⁰One possible explanation for the lower performance of SUPER SS \rightsquigarrow SS2 is that the SS \rightsquigarrow SS2 combinations are too fine-grained for the size of the data.

| Model | Similarity | | Precision | | Recall | |
|---------------------------|--------------|------|--------------|------|--------------|------|
| | Mean | SD | Mean | SD | Mean | SD |
| WIDE | 64.82 | 3.46 | 73.73 | 3.29 | 69.00 | 3.39 |
| BASE | 80.53 | 3.13 | 84.96 | 2.91 | 86.24 | 2.78 |
| PREP | 81.99 | 2.45 | 86.42 | 2.43 | 88.05 | 2.14 |
| SS2 | 83.57 | 2.45 | 88.50 | 1.88 | 89.88 | 1.76 |
| SS \rightsquigarrow SS2 | 83.45 | 2.69 | 88.23 | 2.15 | 89.48 | 1.92 |

Table 1: Mean results and standard deviation of each of the four systems — BASELINE, PREP, SUPER SS2 and SUPER SS \rightsquigarrow SS2 — and the WIDE measure.

icant with $F(2, 87) = 9.21$ and $p = 0.000$. Table 2 presents the p-values of the post hoc t-tests performed in order to inspect particular differences.

| Model | BASELINE | PREP |
|-----------|--------------|-------|
| PREP | 0.056 | - |
| SUPER SS2 | 0.000 | 0.056 |

Table 2: P-values for pairwise t-tests with Holm correction for multiple comparisons; similarity metric.

It is not the case that providing information about prepositions to the models significantly improves the performance of those models — the p-value for the comparison between BASELINE and PREP is bigger than 0.05. However, as predicted, providing the model with information about preposition supersenses, only the SS2 part, significantly improves the performance with respect to BASELINE — p-value < 0.05 but not with respect to PREP — p-value > 0.05 .¹¹

6 Summary and Future Work

This study dealt with the QSD task following the methodology established by Manshadi and Allen (Manshadi and Allen, 2011; Manshadi et al., 2013; Manshadi, 2014) which allows to consider any sentence, with no restriction on the number of quantifiers involved, in a ternary classification task. Applying this method to the scope-disambiguated corpus (AnderBois et al., 2012), additionally tagged

¹¹We also experimented with a model that includes both preposition lexicalizations and preposition supersenses (just the SS2 part). Performance of this system is as follows (mean, SD): (83.01, 2.29), (87.52, 2.17), (89.12, 1.88) for similarity, precision and recall, respectively. Statistical analysis of this system’s results led to exactly the same conclusions as when the model including only supersenses was used. That is, it performed significantly better with respect to the BASELINE but not with respect to the system involving only preposition lexicalizations. Thus, there is no theoretical reason to believe that preposition lexicalizations would encode any relevant information that is not already captured by preposition supersenses.

with the SNACS scheme (Schneider et al., 2020), allowed us to investigate the question of whether information encoded by prepositions, or preposition senses to be exact, proves useful in the QSD task, as inspected with SVM.

Our results confirm the formulated hypothesis — preposition senses, but not preposition lexicalizations, positively impact the learnability of the models and, hence, it may be inferred that they do convey *world knowledge* in a manner beneficial for the algorithm. Note that, out of 1679 tagged quantified NPs in the corpus, only around a third (581 to be exact) are nested in prepositional phrases; this fact further strengthens our conclusions.

The fact that the methodology followed in this paper reduces the QSD problem to a pairwise comparisons task has its benefits. For instance, it significantly expands our sample from 680 sentences into 1451 pairwise comparisons. However, it comes at the price of a simplification, which might lead to information loss since each pair of quantified NPs in a sentence is treated independently of other NPs in that sentence, which in reality is not the case. Another way to approach the QSD problem would be to treat it on a sentence level, making use of modern deep learning techniques such as pre-trained transformer neural networks. However, that might require larger quantifier scope-disambiguated corpora which do not yet exist.¹²

Acknowledgements

First of all, we would like to thank Scott AnderBois, Adrian Brasoveanu and Robert Henderson for sharing their scope-disambiguated corpus with us. We would also like to express our great thanks to the Law School Admission Council for allowing us to use the materials in the corpus for our research purposes. We are also grateful to the three anonymous reviewers for helpful comments and suggestions. The research of the first two authors and the fifth one is funded by the National Science Center, Poland (Grant No. DEC-2019/35/B/HS1/01541).

¹²We tested a solution based on the Universal Sentence Encoder (Cer et al., 2018). For each quantifier, we concatenated vectors: embedding of the quantifier expression and embedding of the entire sentence (2 x 512). The vectors were then classified using the SVM classifier with a radial kernel. This solution achieved an accuracy of 0.45 (the most frequent baseline was at 0.47) when predicting relative quantifier scope. The results are averages in a 10-fold cross-validation.

References

- Lasha Abzianidze and Johan Bos. 2017. Towards universal semantic tagging. In *Proceedings of IWCS*.
- Scott AnderBois, Adrian Brasoveanu, and Robert Henderson. 2012. The pragmatics of quantifier scope: A corpus study. In *Proceedings of Sinn und Bedeutung*, pages 15–28.
- Catherine Anderson. 2004. *The Structure and Real-time Comprehension of Quantifier Scope Ambiguity*. Ph.D. thesis, Northwestern University.
- Chris Barker and Chung-chieh Shan. 2014. *Continuations and natural language*, volume 53. Oxford Studies in Theoretical Linguistics.
- Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. *English Web Treebank*, volume 53. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, PA.
- Johan Bos. 1996. Predicate logic unplugged. In *Proceedings of the 10th Amsterdam Colloquium*, pages 133–143.
- Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. 2004. Wide-coverage semantic representations from a CCG parser. In *Proceedings of COLING 2004*, pages 1240–1246.
- Arthur Capelier-Mourguy, Philippe Blache, Christian Retoré, and Laurent Prévot. 2015. Quantifier scope: A formal and experimental study. In *CJC-SC: Colloque des Jeunes Chercheurs en Sciences Cognitives*, pages 67–69.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium. Association for Computational Linguistics.
- W. W. Cohen, R. E. Schapire, and Y. Singer. 1999. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270.
- Robin Cooper. 1983. *Quantification and semantic theory*. Dordrecht: Reidel.
- Ann Copestake, Alex Lascarides, and Dan Flickinger. 2001. An algebra for semantic construction in constraint-based grammars. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 140–147.
- Jakub Dotlačil and Adrian Brasoveanu. 2015. The manner and time course of updating quantifier scope representations in discourse. *Language, Cognition and Neuroscience*, 30(3):305–323.

- Markus Egg, Alexander Koller, and Joachim Niehren. 2001. The constraint language for lambda structures. *Journal of Logic, Language, and Information*, 10(4):457–485.
- Kilian Evang and Johan Bos. 2013. Scope disambiguation as a tagging task. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 314–320.
- Kathryn Gillen. 1991. *The Comprehension of Doubly Quantified Sentences*. Ph.D. thesis, University of Durham.
- Justyna Grudzińska and Marek Zawadowski. 2019. Inverse linking, possessive weak definites and Haddock descriptions: A unified dependent type account. *Journal of Logic, Language and Information*, 28(2):239–260.
- Justyna Grudzińska and Marek Zawadowski. 2020. A scope-taking system with dependent types and continuations. In *Logic and Algorithms in Computational Linguistics 2018 (LACompLing2018)*, pages 155–176. Springer.
- Herman Hendriks. 1993. *Studied flexibility: Categories and types in syntax and semantics*. Institute for Logic, Language and Computation.
- Derrick Higgins and Jerrold M. Sadock. 2003. A machine learning approach to modeling scope preferences. *Computational Linguistics*, 29(1):73–96.
- Jerry R. Hobbs and Stuart M. Shieber. 1987. An algorithm for generating quantifier scopings. *Computational Linguistics*, 13:47–63.
- Jena D. Hwang, Archana Bhatia, Na-Rae Han, Tim O’Gorman, Vivek Srikumar, and Nathan Schneider. 2017. Double trouble: The problem of construal in semantic annotation of adpositions. In *Proceedings of *SEM*, pages 178–188.
- Georgette Ioup. 1975. Some universals for quantifier scope. In *Syntax and Semantics*, volume 4, pages 37–58. Academic Press, New York.
- Alexander Koller, Michaela Regneri, and Stefan Thater. 2008. Regular tree grammars as a formalism for scope underspecification. In *Proceedings of ACL-08: HLT*, pages 218–226.
- Howard S. Kurtzman and Maryellen C. MacDonald. 1993. Resolution of quantifier scope ambiguities. *Cognition*, 48(3):243–279.
- Ken Litkowski and Orin Hargraves. 2007. Word-sense disambiguation of prepositions. In *Proceedings of SemEval*, pages 24–29.
- Mehdi Manshadi and James Allen. 2011. Unrestricted quantifier scope disambiguation. In *Proceedings of Association for Computational Linguistics’11, Workshop on Graph-based Methods for NLP (TextGraph-6)*, pages 51–59.
- Mehdi Manshadi, Daniel Gildea, and James Allen. 2013. Plurality, negation, and quantification: Towards comprehensive quantifier scope disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 64–72.
- Mohammad Hafezi Manshadi. 2014. *Dealing with quantifier scope ambiguity in natural language understanding*. University of Rochester.
- Robert May. 1978. *The grammar of quantification*. Ph.D. thesis, Massachusetts Institute of Technology.
- Robert May. 1985. *Logical Form: Its structure and derivation*, volume 12. MIT press.
- Dennis L. Micham, Jack Catlin, Nancy J. VanDerveer, and Katherine A. Loveland. 1980. Lexical and structural cues in quantifier scope relations. *Journal of Psycholinguistics Research*, 48(9):367–377.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- James Pustejovsky. 1991. The syntax of event structure. *Cognition*, 41(1–3):47–81.
- Janina Radó and Oliver Bott. 2011. Underspecified representations of scope ambiguity? In *Proceedings of the 18th Amsterdam Colloquium*, pages 180–189.
- Walid S. Saba and Jean-Pierre Corriveau. 2001. Plausible reasoning and the resolution of quantifier scope ambiguities. *Studia Logica*, 67(2):271–289.
- Asad Sayeed. 2016. Representing the effort in resolving ambiguous scope. In *Proceedings of Sinn und Bedeutung*, pages 604–621.
- Asad Sayeed, Matthias Lindemann, and Vera Demberg. 2019. Verb-second effect on quantifier scope interpretation. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 134–139.
- Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Adi Shalev, Omri Abend, Archana Bhatia, Na-Rae Han, and Tim O’Gorman. 2020. Adposition and case supersenses v2.5: Guidelines for english. arxiv.org/pdf/1704.02134v6.
- Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 185–196. Association for Computational Linguistics.

- Vivek Srikumar and Dan Roth. 2013. Modeling semantic relations expressed by prepositions. *Transactions of the Association for Computational Linguistics*, 1:231–242.
- Prakash Srinivasan and Alexander Yates. 2009. Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1465–1474.
- Mark Steedman. 2012. *Taking scope: The natural semantics of quantifiers*. Mit Press.
- KC Tsiolis. 2020. Quantifier scope disambiguation. Summary of research conducted in summer 2020 under the supervision of Jackie Cheung.
- Susanne Lynn Tunstall. 1998. *The interpretation of quantifiers: semantics and processing*. Ph.D. thesis, University of Massachusetts Amherst.
- William A. Woods. 1987. Semantics and quantification in natural language question answering. *Advances in Computers*, 17:1–87.