European Public & Social Innovation Review

ISSN 2529-9824

Research Article

Predictive power of semantic information in the use context of multiword terms for their structural disambiguation

Potencia predictiva de la información semántica en el contexto de uso de unidades terminológicas poliléxicas para su desambiguación estructural

Juan Rojas-García: University of Granada, Spain.
juanrojas@ugr.es

Date of Receipt: 24/03/2025

Acceptance Date: 16/04/2025

Date of Publication: 21/04/2025

How to cite this article (APA 7ª):

Rojas-García, J. (2025). Predictive power of semantic information in the use context of multiword terms for their structural disambiguation [Potencia predictiva de la información semántica en el contexto de uso de unidades terminológicas poliléxicas para su desambiguación estructural]. *European Public & Social Innovation Review*, 10, 1-22. https://doi.org/10.31637/epsir-2025-2070

Abstract:
Introduction: The structural disambiguation of English multiword terms (MWT) of three or more constituents (e.g., *coastal sediment transport*), often known as *bracketing*, involves the grouping of the dependent components so that the MWT is reduced to its basic form of modifier+head, as in *coastal* [*sediment transport*], which is a right-bracketed ternary compound. This work presents a study that explored whether the bracketing of a ternary compound, when used as an argument in a sentence, can be predicted from the semantic information encoded in that sentence. **Methodology**: A set of 1.694 sentences were analyzed semantically and annotated with the lexical domain of the verbs, the semantic role and category of the arguments, and the semantic relation between the arguments. These semantic variables were then analyzed statistically to determine whether they are able to predict the bracketing of a ternary compound. **Results:** A random forest model, with the lexical domain of the verb, and the semantic role and category of the MWT, was able to predict the bracketing of the ternary compounds used as arguments in a sample of 380 MWTs (100% F_1-score). A decision tree, with solely the semantic relation of the MWT to another argument in the same sentence, was also able to predict the bracketing of the ternary compounds in the sample (94,12% F_1-score). **Discussion**: Only a subset of three variables was necessary for bracketing prediction with an error free performance, whereas previous research employed a minimum of 12 variables.

Conclusion: The semantic information in a sentence contributed substantially to compound parsing. This suggests a novel research direction in the integration of semantic variables into syntactic parsers and machine-translation applications.

Keywords: multiword-term bracketing; bracketing prediction; random forest model; decision-tree model; verb lexical domain; semantic role; semantic category; semantic relation.

Resumen:
Introducción: La desambiguación estructural de unidades terminológicas poliléxicas (UTP) de tres o más formantes en inglés (p.e., *coastal sediment transport*), también denominado *bracketing*, consiste en reducir la UTP a su forma básica de modificador+núcleo. Así, se puede indicar que el *bracketing* de *coastal* [*sediment transport*] es derecho. En este trabajo se explora si el *bracketing* de una UTP ternaria, usada como argumento en una oración, puede predecirse a partir de la información semántica codificada en dicha oración. **Metodología**: Se analizan semánticamente 1.694 oraciones, en las que se anotan el dominio léxico del verbo, la categoría y el rol semánticos de los argumentos verbales, y la relación semántica entre los argumentos. Estas variables semánticas se emplean para predecir el *bracketing* de una UTP ternaria. **Resultados:** El modelo predictivo *random forest*, con el dominio léxico del verbo, y el rol y la categoría de la UTP, predice el *bracketing* de las UTP ternarias, usadas como argumentos en una muestra de 380 UTP ternarias (100 % F_1-score). El modelo árbol de decisión, únicamente con la relación semántica entre la UTP y otro argumento en la misma oración, también predice el *bracketing* de las UTP ternarias de la muestra (94,12 % F_1-score). **Discusión:** Solo tres variables son necesarias para predecir el *bracketing* sin errores, mientras que las investigaciones previas emplearon un mínimo de 12 variables. **Conclusiones**: La información semántica, codificada en una oración, contribuye sustancialmente a la predicción del *bracketing*. Esto abre una nueva línea de investigación, que sugiere la incorporación de variables semánticas en analizadores sintácticos y en sistemas de traducción automática.

Palabras clave: *bracketing* de unidades terminológicas poliléxicas; predicción del *bracketing*; modelo *random forest*; modelo árbol de decisión; dominio léxico del verbo; rol semántico; categoría semántica; relación semántica.

1. Introduction

When multiword expressions are used in specialized knowledge domains, they are known as multiword terms (MWTs). MWTs often have more than two components. For MWTs of three or more constituents (e.g., the three-component MWT *sea level rise*), a semantic analysis, based on linguistic and domain knowledge, is necessary to resolve the dependency between components. This structural disambiguation, often known as *bracketing* or *parsing*, involves the grouping of the dependent components so that the MWT is reduced to its basic form of modifier+head, as in [*sea level*] *rise*. Knowledge of these dependencies facilitates the comprehension of an MWT and, consequently, its accurate translation into other languages. A case in point is *natural sediment supply*. Since we know that this is a right-branched MWT (i.e., it is parsed as *natural* [*sediment supply*]), its correct translation into Spanish would be *suministro natural de sedimentos*, and not *suministro de sedimentos naturales*. Another example is *sea level rise*, whose bracketing is left (i.e., it is parsed as [*sea level*] *rise*). This means that its correct translation into Spanish would be *incremento del nivel del mar*, and not *incremento marino del nivel*.

MWTs can be included in Terminological Knowledge Bases (TKBs). For that, it is often necessary to structurally disambiguate MWTs to make their relational structure explicit and thus favor knowledge acquisition in communicative situations such as specialized translation (León-Araúz et al., 2021). For this reason, a specific module for MWT representation has been

designed in the TKB EcoLexicon[1] (San Martín et al., 2020) to include bracketing-related information. This was considered useful because the resolution of MWT bracketing provides a higher overall accuracy in machine-translation systems (Garg et al., 2022) and sentence parsers (Vadas & Curran, 2008). Thus, given the importance of the structural disambiguation of MWTs, this study focuses on the bracketing prediction of three-component MWTs, henceforth also referred to as *ternary compounds*. As predictors for the predictive models (decision tree and random forest), semantic variables are used. These variables come from the semantic annotation of the predicate-argument structure of sentences that, at the same time, mention potamonyms (i.e., named rivers such as the Salinas River, in California) and include ternary compounds in an English specialized corpus on Coastal Engineering (7 million tokens).

The two reasons why the analyzed sentences cite potamonyms are to be clarified. Firstly, as shall be seen, the semantic relation that a ternary compound maintains with another argument in the same sentence is essential to the prediction of its bracketing. However, the semantic relations that are activated in each subject field are different and they remain unknown unless an in-depth semantic analysis is performed. Consequently, for the establishment of a closed set of semantic relations to use during the annotation process, a previous semantic analysis was carried out. This analysis limited to a sample of 1.694 sentences, extracted from the corpus on Coastal Engineering, which mentioned potamonyms. In doing so, we could determine which specific semantic relations potamonyms kept to other specialized concepts in Coastal Engineering, which were reported by Rojas-Garcia (2022a). Secondly, the ultimate goal of the semantic analysis of the 1.694 sentences was the semantic representation of named rivers in EcoLexicon. This is a digital, multilingual TKB on environmental sciences, designed according to Frame-based Terminology (Faber, 2015), the theoretical framework of this research, on which the semantic variables for bracketing prediction are drawn. Frame-based Terminology is a cognitive theory of Terminology – a discipline that deals with the theoretical and methodological description of the specialized language –, which contextualizes concepts in frames, also called semantic networks or knowledge structures, and is based on corpus analysis.

1.1. Hypothesys

Since the semantic information in a sentence firmly guides its syntactic parsing (Lazaridou et al., 2013), as hypothesis for this experiment, we assumed that the correct bracketing of a ternary compound, when used as an argument in a sentence, can be predicted from the semantic information encoded in that sentence. This assumption comes from the daily experience of a translator who must deal with ternary compounds in specialized texts. Although the compounds are somewhat familiar, it is useful to craft definitions for them to facilitate their translation into another language based on their context of use.

1.2. Objectives

This article presents the statistical analysis of the semantic variables annotated with a view to finding evidence that the structural disambiguation, or bracketing, of a three-component MWT can be predicted from the semantic information encoded in the sentence where the ternary compound is used as an argument. For that purpose, a set of specific objectives were established:

1. To create a sample of 378 sentences from the corpus in which 380 ternary compounds are used as arguments. These 378 sentences are drawn from a larger sample of 1.694 sentences in the corpus in which a potamonym is cited. The 378 sentences are the total number of sentences that include ternary compounds in the sample of 1.694 sentences.

[1] https://ecolexicon.ugr.es

2. To annotate the following five semantic variables in the sample of 378 sentences: (a) semantic category of the arguments; (b) semantic role of the arguments; (c) semantic relation held between the potamonym and the other argument, which is a ternary compound, in the same sentence; (d) lexical domain of the verb; and (e) bracketing (left or right) of the ternary compounds used as arguments in the sentences.
3. To select the predictor variables, abovementioned, that best predict the bracketing of a ternary compound.
4. To build two predictive models, namely, random forest and decision tree, to predict the bracketing of a ternary compound.
5. To compare the performance of the two predictive models with each other.

1.3. Novelty of this Substantially Extended Version of a Conference Paper

It is to be mentioned that this article is a substantially extended version of the conference paper Rojas-Garcia (2022b), which presented that, in a sample of 190 ternary compounds, the semantic relation that a ternary compound maintains with another argument in the same sentence can predict its correct bracketing. Notwithstanding, the work in this article goes beyond the earlier publication and establish some novelty and new results:

1. The sample is expanded from 190 to 380 ternary compounds.
2. The contribution of the verb lexical domain, and semantic role and category of the arguments in a sentence to bracketing prediction is offered.
3. Linguistic insights are provided as to how the verb lexical domains, semantic roles, categories, and relations are intertwined with the bracketing of ternary compounds.
4. The study of the pairwise associations between variables with Fisher's exact tests and Cramer's V values is carried out. The analysis of the predictor variables that most contributed to bracketing prediction and thus to higher model performance is accomplished.
5. Also described is the comparison of the performance of the predictive models.

1.4. Related Work: Methods for Bracketing Prediction in the Literature

Most work on compound bracketing (or parsing) exploits either unsupervised methods (e.g., based on bigram corpus frequency) or supervised ones (i.e., based on training data, containing manually bracketed compounds, which are used to train an algorithm for predicting compound bracketing). We refer readers to Rojas-Garcia (2022b, pp. 142-145) for a detailed description of previous methods on bracketing prediction.

Previous research focused on semantic information provided by the components of an MWT. The number of variables used for prediction ranged from 12 to 517.254 features. These variables were mostly based on n-gram statistics, and semantic information of the MWT components stored in linguistic resources such as WordNet. The overall accuracy of the prediction models ranged from 72,60% to 95,40%. Our approach, however, was based on semantic information that previous research has not as yet considered. This semantic information was encoded in both the co-text of a ternary compound (i.e., the sentence where the ternary compound was used as an argument) and the ternary compound seen as a unit (i.e., its semantic role and category). The initial set of variables was only five, whereas previous research employed a minimum of 12 variables (León-Araúz et al., 2021).

1.5. Organization of the Article

The rest of the article is organized as follows. Section 2 explains the materials used in this study, and covers our semantic approach to predicting ternary compound bracketing based on two

supervised models (decision tree and random forest). Also described are the sample of ternary compounds, the pairwise associations between variables, and the predictor variable selection. Section 3 focuses on the training and testing phases for the predictive models, and presents the results, which provide linguistic insights as to how semantic relations, verb lexical domains, and semantic roles and categories are intertwined with the bracketing of ternary compounds. Section 4 discusses the results and compares them to those outlined in the literature review. Finally, Section 5 presents the conclusions derived from this work along with plans for future research.

2. Methodology

2.1. Materials

A set of 1.694 sentences, in which a named river (e.g., Mississippi River) was an argument of the predicate of the sentences, were semantically analyzed and annotated. These sentences were extracted from a corpus of English texts on Coastal Engineering, comprising roughly 7 million tokens. This subcorpus was composed of specialized texts (scientific articles, technical reports, and Ph.D. dissertations), which amounted to 73,17% of the corpus size; and semi-specialized texts (textbooks and encyclopedias on Coastal Engineering), which constituted 26,83% of the corpus size. The total number of texts of the subcorpus was 2.249, whose publication data ranged from 1996 to 2018.

2.2. Semantic Variables

As semantic information in a sentence, this study explored the contribution of five semantic variables to bracketing prediction in ternary compounds. From the 1.694 sentences semantically analyzed and annotated, a total number of 378 sentences contained 380 ternary compounds as arguments. This sample of 380 ternary compounds, along with the values of the five semantic variables annotated in their corresponding sentences, were employed for the training and testing of two supervised models to predict whether a ternary compound was right-branched or left-branched.

2.3. Annotation of the Semantic Variables

A set of 1.694 sentences from the corpus, where 294 different rivers are mentioned, were manually annotated by three terminologists from the LexiCon research group (University of Granada),[2] who have wide experience in environmental knowledge representation. They performed the semantic annotation of the predicate-argument structure of a sentence by assigning, in this order, a: (1) lexical domain to the verb; (2) semantic role to the potamonym; (3) semantic role to the other argument, which was a ternary compound in this study; (4) semantic category to the ternary compound; (5) semantic relation to the link between the potamonym and the other argument, which was a ternary compound in this study; and (6) bracketing (left or right) to the ternary compounds used as arguments in the sentence. The values of these six variables are shown in Table 1.

Table 1.

Semantic variables annotated in the set of 378 sentences, and their values.

Semantic variables annotated	Values
Lexical domain of the verbs (8 values)	CHANGE, MOVEMENT, EXISTENCE, POSSESSION, POSITION, MANIPULATION, ACTION, COGNITION

[2] http://lexicon.ugr.es/

Semantic roles of the arguments (13 values)	AGENT, RESULT, PATIENT, THEME, LOCATION, RECIPIENT, INSTRUMENT, TIME, RATE, MANNER, DESCRIPTION, CONDITION, PURPOSE
Semantic categories of the arguments (3 values)	PROCESS, ENTITY, ATTRIBUTE
Semantic relation between the ternary compound and the named river (30 values)	*type_of*, *part_of*, *made_of*, *delimited_by*, *located_at*, *takes_place_in*, *phase_of*, *affects*, *causes*, *result_of*, *attribute_of*, *has_function*, *studies*, *measures*, *effected_by*, *improves*, *worsens*, *creates*, *becomes*, *gives*, *gives_to*, *receives*, *receives_from*, *drains*, *has_path*, *transfers*, *discharges_into*, *places*, *controls*, *applied_to*
Bracketing of the ternary compounds in the sentences (2 values)	RIGHT, LEFT

Source: Self-made.

The annotation process was carried out with the INCEpTION annotation tool (Klie et al., 2018) and lasted five months. The INCEpTION annotation scheme organizes annotations in layers, which represent the features to be annotated in a project and their labels. We set five layers for five annotation features: (1) the first layer for the *verb lexical-domain* feature; (2) the second layer for the *semantic role*; (3) the third layer for the *semantic category*; (4) the fourth layer for the *semantic relation*; and (5) the fifth layer for the *bracketing* of ternary compounds.

The most frequent verbs in the corpus are general language verbs (e.g., *accumulate*, *pollute*, *increase*, *discharge*, *supply*, *drain*), which are also used in specialized texts and thus reflect how environmental entities interact. In this sense, such verbs are susceptible to classification in the eight lexical domains proposed by Faber and Mairal (1999), within the Functional Lexematic Model. In this regard, the majority of verbs in the English EcoLexicon Corpus were found to belong to the lexical domains shown in Table 1, which were used to annotate the verbs of our set of sentences.

Specialized knowledge representation includes semantic properties that help to describe the nature of entities and processes. These semantic properties are reflected as the relations between a verb and its arguments, which are typical semantic roles. The semantic roles used to annotate the arguments in our set of sentences were based on those specified by Kroeger (2005, pp. 54-55) and Thompson et al. (2009), which are summarized in Table 1.

Depending on the ontological nature of concepts, they can be classified in three basic categories, namely PROCESS (i.e., events extending over time and involving different participants), ENTITY (i.e., physical and mental objects), and ATTRIBUTE (i.e., properties of entities and processes). These three values were all used to annotate the semantic category of the arguments in our set of sentences, and collected in Table 1.

Conceptual description of specialized concepts includes their relational behavior. These relations for environmental concepts, with additional non-hierarchical relations specific to named rivers (Rojas-Garcia, 2022a), were all used to annotate the semantic relation between the arguments in our set of sentences, as shown in Table 1.

A selection of 19 sentences from the sample, which incorporated ternary compounds as arguments, is provided in Table 2. This selection of 19 annotated sentences represents the main characteristics of the sample of 380 ternary compounds. For each of those 19 sentences, Table 3 shows the values of the following six annotated variables:

1. Lexical domain of the verb (*LexDom*).
2. Semantic category of the ternary compound in the first categorization level, that is, its ontological category (*SemCat_Level_1*).
3. Semantic role of the ternary compound (*SemRol_mwt*).
4. Semantic role of the named river (*SemRol_river*).

5. Semantic relation between the ternary compound and the named river (*SemRel*).
6. Bracketing of the ternary compound (*Bracketing*), the variable to be predicted.

Table 2.

Representative selection of 19 sentences (from the sample of 378 sentences), which included 19 ternary compounds as arguments.

Sentences from the Sample with Ternary Compounds as Arguments
(1) **Sea level rise** *is occurring* in the region of the **Mississippi River Delta** on the order of 1 cm per year, eight times the worldwide rate.
(2) After the dam construction, the **sediment supply decrease** in the **Tenryu River** *resulted* in a delta coastline recession.
(3) After the dam construction, the sediment supply decrease in the **Tenryu River** *resulted* in a **delta coastline recession**.
(4) The dramatical **sediment load variation** in the **Pearl River**, with the almost unchanged water discharge level, *represents* an example of such effect that human activities can have on river deltas.
(5) The **Changjiang River** *provides* most of the **fresh water input**.
(6) Natural and anthropogenic effects combine to *result* in the **maximum erosional stress** on barrier islands, located near the **Mississippi River mouth**.
(7) Blackstone River draining into Narragansett Bay has been extensively dammed, and although not well quantified, models *show* **decreasing sediment load** in the **Blackstone River**.
(8) The dramatical sediment load variation in the **Pearl River**, with the almost unchanged **water discharge level**, *represents* an example of such effect that human activities can have on river deltas.
(9) **Muddy silt deposition** in the **Clyne River** discharging into the Swansea Bay *would increase*.
(10) The **vegetation removal effect** over the entire study reach *changed* the **Gila River** from a continually losing river for most years before clearing to a gaining stream during some months for most years following clearing.
(11) **Rising sea levels** *change* **Salinas River Estuary** and could thus potentially alter sediment supplies and process patterns.
(12) However, in this instance, the anthropogenic effects probably dominate and *include* additional subsidence resulting from withdrawal of hydrocarbons, and the **sediment supply reduction** in the **Mississippi River** by the construction of upstream impoundments and jetties that direct the riverine sediment offshore to deepwater.
(13) The rest of the gravel for beach nourishment (5%) *comes* from **flood prevention operations** upstream on the gravelly bed of the **Paillon River**, a small (catchment size: 236 km^2), completely urbanised river debouching in the eastern part of Nice Bay.
(14) The **Salinas River** no longer *contributes* substantial **beach size sand** to the Littoral Cell because the river gradient has greatly decreased with sea level rise, reducing the flow rate.
(15) The **River Murray** flows across Tertiary formations to *enter* coastal lagoons behind the **dune calcarenite barriers** of Encounter Bay.
(16) Not all the sediments drained by the **Dee River** *participate* to **coastal sediment transport**.
(17) The field site for this study is the **Zuidgors salt marsh**, *located* in the **Western Scheldt estuary** in The Netherlands.
(18) **Natural sediment supply** within this region *is defined* by the **Ventura River** that drains large watersheds.
(19) The **average discharge rate** of beach size sand in the **Salinas River** *is estimated* at approximately 65.000 cubic yards per year.

Source: Self-made.

Table 3.

Semantic annotations and variables for a set of 19 MWTs out of the 380 MWTs that comprised the sample. The semantic information in the rows corresponds to the respective sentences in Table 2.

Sent.	MWT	LexDom	SemCat_ Level_1	SemRol_ mwt	SemRol_ river	SemRel	Bracketing
1	[sea level] rise	EXISTENCE	PROCESS	AGENT	LOCATION	*takes_place_in*	LEFT
2	[sediment supply] decrease	EXISTENCE	PROCESS	AGENT	LOCATION	*takes_place_in*	LEFT
3	[delta coastline] recession	EXISTENCE	PROCESS	RESULT	LOCATION	*takes_place_in*	LEFT
4	[sediment load] variation	EXISTENCE	PROCESS	THEME	LOCATION	*takes_place_in*	LEFT
5	[fresh water] input	EXISTENCE	PROCESS	THEME	AGENT	*causes*	LEFT
6	maximum [erosional stress]	EXISTENCE	ENTITY	RESULT	PATIENT	*affects*	RIGHT
7	decreasing [sediment load]	EXISTENCE	ATTRIBUTE	DESCRIPTION	LOCATION	*attribute_of*	RIGHT
8	[water discharge] level	EXISTENCE	ATTRIBUTE	THEME	LOCATION	*attribute_of*	LEFT
9	[muddy silt] deposition	CHANGE	PROCESS	PATIENT	LOCATION	*takes_place_in*	LEFT
10	[vegetation removal] effect	CHANGE	ENTITY	AGENT	PATIENT	*improves*	LEFT
11	rising [sea level]	CHANGE	ATTRIBUTE	AGENT	PATIENT	*worsens*	RIGHT
12	[sediment supply] reduction	POSSESSION	PROCESS	THEME	LOCATION	*takes_place_in*	LEFT
13	[flood prevention] operation	POSSESSION	PROCESS	AGENT	LOCATION	*takes_place_in*	LEFT
14	[beach size] sand	POSSESSION	ENTITY	THEME	AGENT	*gives*	LEFT
15	dune [calcarenite barrier]	MOVEMENT	ENTITY	AGENT	LOCATION	*has_path*	RIGHT
16	coastal [sediment transport]	ACTION	PROCESS	DESCRIPTION	AGENT	*affects*	RIGHT
17	Zuidgors [salt marsh]	POSITION	ENTITY	THEME	LOCATION	*located_at*	RIGHT
18	natural [sediment supply]	MANIPUL.	PROCESS	PATIENT	AGENT	*controls*	RIGHT
19	average [discharge rate]	COGNITION	ATTRIBUTE	THEME	LOCATION	*attribute_of*	RIGHT

Source: Self-made.

2.4. Inter-Annotation Agreement

As for the inter-annotation agreement (Brezina, 2018, pp. 87-92), Cohen's kappa coefficient (κ) was used. Table 4 shows the values of κ between the pairs of annotators for semantic categories, semantic relations, semantic roles, lexical domains, and bracketing.

Table 4.

Inter-annotator agreement for semantic categories, semantic relations, semantic roles, lexical domains, and bracketing.

Variable	Annotator 1 – Annotator 2	Annotator 1 – Annotator 3	Annotator 2 – Annotator 3
Semantic Category (Level 1)	97,4% (*p*-value<0,05)	96,2% (*p*-value<0,05)	97,2% (*p*-value<0,05)
Semantic Relation	96,3% (*p*-value<0,05)	95,7% (*p*-value<0,05)	96,1% (*p*-value<0,05)
Semantic Role	94,8% (*p*-value<0,05)	92,6% (*p*-value<0,05)	90,8% (*p*-value<0,05)
Lexical Domain	87,9% (*p*-value<0,05)	86,7% (*p*-value<0,05)	84,6% (*p*-value<0,05)
Bracketing	98,8% (*p*-value<0,05)	98,1% (*p*-value<0,05)	98,5% (*p*-value<0,05)

Source: Original analysis by the author.

For the initial annotations of semantic roles, categories, and relations, thanks to the extensive experience of the annotators in semantic analysis of environmental knowledge, Cohen's kappa coefficient showed a very good agreement for all annotator pairs (90%<κ<98%, *p*-values<0,05), according to Krippendorff's (2012) recommendations for text content analysis. A review of the differences between annotators showed no systematic pattern of disagreement. Given the nature of the judgement variables, the level of agreement was deemed acceptable. Notwithstanding, the disagreements in the original annotations were resolved based on discussion between the annotators to reach a consensus on the definitive annotations of semantic roles, categories, and relations.

For the initial annotation of verbs with lexical domains, the inter-annotation agreement was lower for all the annotator pairs (84%<κ<88%, *p*-values<0,05), indicating that this variable lent itself to alternative, though plausible, interpretations. A review of the differences between annotators showed that the lexical domains of movement and possession were more prone to confusion. The issues fundamentally arose from verbs that could potentially belong to more than one lexical domain (e.g., *drain* and *discharge* could belong to the movement and possession domains), as Faber and Mairal (1999) already proved. To arrive at a consensus on the definitive annotations of lexical domains, the factorization of meaning from the Functional Lexematic Model framework was applied to confusion-prone verbs.

Finally, for the initial annotations of ternary compound bracketing, the inter-annotation agreement coefficient showed a very good agreement for all the annotator pairs (κ>98%, *p*-values<0,05) thanks to the considerable experience of the annotators in semantic analysis of environmental knowledge. A review of the differences between the annotators showed disagreement on the bracketing of certain MWTs, such as *dune calcarenite barrier* (row 15 in Table 3), because of their *semantic indeterminacy*. A ternary compound is semantically indeterminate when the meanings associated with the two bracketing structures cannot be distinguished in the context (Hindle & Rooth, 1993, p. 113), or when the meanings corresponding to each bracketing structure are the same (Lauer, 1995, p. 21). The disagreements in the original annotations were resolved by discussion between the annotators. For instance, in the case of *dune calcarenite barrier*, the decision was thus to regard the MWT as right-branched since the values of its semantic variables (*LexDom*=MOVEMENT, *SemCat_Level_1*=ENTITY, *SemRol_mwt*=AGENT) were the same as the values of other MWTs annotated as right-branched. Hence, given the nature of the judgement variable (*Bracketing*), the level of agreement in the initial annotations was deemed almost perfect.

2.5. Description of the Sample of Ternary Compounds

The distribution of bracketing structures within the MWT sample was reasonably balanced between left-branching (220 MWTs, 58% of the sample), and right-branching (160 MWTs, 42% of the sample). Table 5 summarizes the counts for the sample data, disaggregated by lexical domain, semantic category of the MWTs, and their bracketing structure, and describes the distribution of the 380 MWTs across these variables.

Table 5.

Description of the sample of ternary compounds.

Lexical Domain	LEFT-branched MWTs				RIGHT-branched MWTs				Total
	Process	Entity	Attrib.	Total	Process	Entity	Attrib.	Total	
MOVEMENT	0	0	0	0	0	20	0	20	20 (5,3%)
POSSESSION	40	20	0	60	0	0	0	0	60 (15,8%)
CHANGE	20	20	0	40	0	0	20	20	60 (15,8%)
EXISTENCE	100	0	20	120	0	20	20	40	160 (42,0%)
ACTION	0	0	0	0	20	0	0	20	20 (5,3%)
POSITION	0	0	0	0	0	20	0	20	20 (5,3%)
MANIPUL.	0	0	0	0	20	0	0	20	20 (5,3%)
COGNITION	0	0	0	0	0	0	20	20	20 (5,3%)
Total	160 (42%)	40 (10,5%)	20 (5,4%)	**220 (58%)**	40 (10,5%)	60 (15,8%)	60 (15,8%)	**160 (42%)**	**380 (100%)**

Source: Original analysis by the author.

Some conclusions could be drawn from the characteristics of the sample:

1. Most ternary compounds designating processes were left-branched.
2. Ternary compounds designating entities remained fairly well distributed among the two bracketing structures.
3. Most ternary compounds designating attributes were right-branched.
4. Sentences whose predicate belonged to the lexical domains of MOVEMENT, ACTION, POSITION, MANIPULATION, AND COGNITION included ternary compounds which were only right-branched.
5. Sentences whose predicate belonged to the lexical domain of POSSESSION incorporated ternary compounds which were only left-branched.

2.6. Pairwise Association between Variables

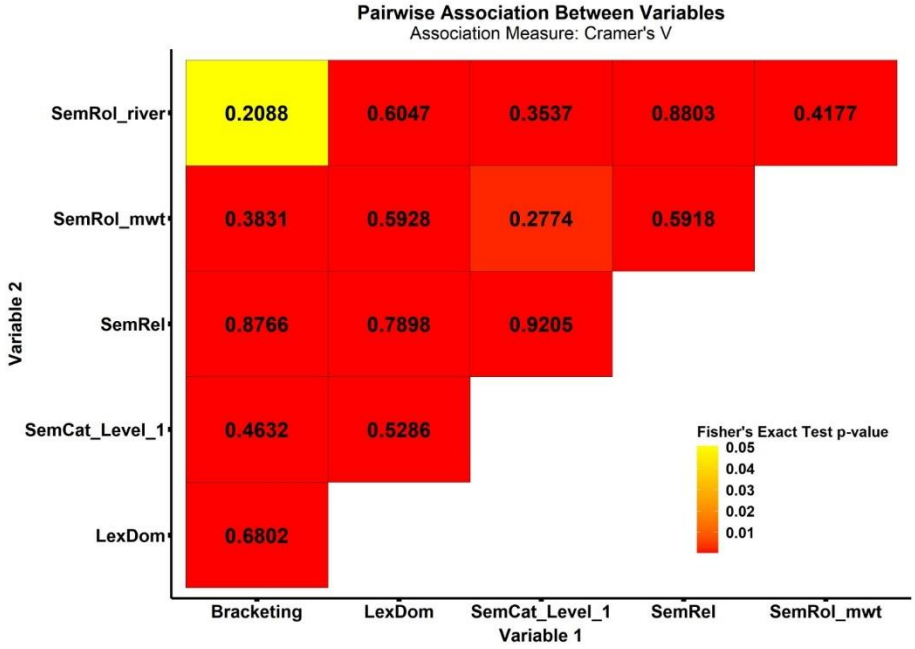
The pairwise associations between the categorical variables were statistically analyzed with Fisher's exact test, which is more appropriate for small-sized samples, such as ours. Two categorical variables are associated if Fisher's exact test finds evidence in the sample to establish that both variables are statistically related to each other (i.e., two variables are statistically related if the *p*-value of Fisher's exact test is less than or equal to 0,05). Then, if two variables were found to be associated, we measured the strength of their association by means of Cramer's V. This measure ranges from 0 to 1. The higher the Cramer's V value, the stronger the association between two variables. For the sake of simplicity, Cramer's V values are usually interpreted as follows: $V \in [0,1, 0,3]$ means weak association; $V \in [0,4, 0,5]$ indicates medium association; and $V > 0,5$

reflects strong association.

Figure 1 shows a pairwise association plot between the categorical variables determined by Fisher's exact tests, where the color-coded squares represents *p*-values (the lower the *p*-values, the redder the squares), and the figures in the middle of the squares are the values of Cramer's V.

Figure 1.

Association plot for the categorical variables of the sample.



Source: Original analysis by the author.

The analysis of the pairwise associations indicated that all variables were statistically associated, except for *Bracketing* and *SemRol_river* (*p*-value=0,051>0,05), whose association strength was thus the lowest (V=0,2088). This signified that the semantic role of the argument slot filled with a named river in our case would not be sufficient to predict the bracketing of the ternary compound that filled the other argument in the same sentence. Nevertheless, the *Bracketing* variable was statistically associated with the other four variables. More specifically, *Bracketing* was strongly associated with both *SemRel* (V=0,8766) and *LexDom* (V=0,6802), moderately associated with *SemCat_Level_1* (V=0,4632), and weakly associated with *SemRol_mwt* (V=0,3831). As a result, both the semantic relation of the MWT to the named river, the lexical domain of the verb, the ontological category of the MWT, and its semantic role would potentially predict bracketing in our sample of ternary compounds.

Another result that should be highlighted is the particularly strong association which the *SemRel* variable maintained with all the other predictor variables, such as *SemCat_Level_1* (V=0,9205), *SemRol_river* (V=0,8803), *LexDom* (V=0,7898), and *SemRol_mwt* (V=0,5918). In our opinion, this finding suggests that the semantic relation between two arguments in a sentence condenses information about the lexical domain of the verb, the semantic categories of both arguments, and their semantic roles. In other words, *SemRel* is a dependent variable that may be predicted from *LexDom*, *SemRol_mwt*, *SemCat_Level_1*, and *SemRol_river*. However, this fact is not studied in this work, and is thus deferred for further investigation in the future.

2.7. Supervised Models

For classification using supervised models, binary decision tree and random forest techniques were tested to predict ternary compound bracketing. Because our dataset consisted entirely of categorical variables, we chose these tree-based models as they handle qualitative data effectively.

A decision-tree model is straightforward and easy to understand because it visually summarizes prediction rules in a tree format, with the terminal nodes or leaves, which represent the predictions, located at the bottom. However, it often does not perform as well as other predictive models. Therefore, we also tried using a random forest model. This model creates many decision trees and combines their predictions to form a single consensus. Specifically, each tree in the forest votes on the bracketing of an MWT, and the final classification is based on the majority vote. While random forest models significantly improve prediction accuracy, they are less interpretable, making it harder to understand how the predictions are made.

2.8. Data Splitting

To build and evaluate the models, the dataset containing 380 MWTs was split into two parts: (1) the training dataset, which consisted of 266 MWTs (70% of the original dataset) and was used to develop the models; and (2) the test dataset, which included 114 MWTs (30% of the original dataset) and was used to assess model performance. To ensure that both the training and test datasets had the same distribution of the outcome variable (i.e., *Bracketing*) as the original dataset (with 58% left-branched MWTs and 42% right-branched MWTs), stratified random sampling was employed. This method randomly selected observations within the LEFT and RIGHT classes of the *Bracketing* variable from the original dataset.

2.9. Model Performance Measures

The effectiveness of the two models was evaluated by examining their performance on the test dataset, which was kept separate from the model-development process to ensure an unbiased assessment. The models' predictions were compared with the actual classes in the test dataset (i.e., the true bracketing structures LEFT and RIGHT, as specified in the *Bracketing* variable). Performance metrics were then calculated based on this comparison.

A common way to measure performance is overall accuracy, which indicates the percentage of correctly classified instances. However, this metric can be misleading in imbalanced datasets, where the outcome variable has a significant disproportion in the number of instances for each class. Although our dataset was only slightly imbalanced in terms of bracketing, we opted to use additional measures that are not affected by class proportion disparities to evaluate classification performance. These measures included the area under the ROC curve (AUC) and the F_1-score (Fernández et al., 2018, pp. 52-55). Both AUC and F_1-score range from 0 to 1, with higher values indicating better model performance in distinguishing between the two classes.

2.10. Predictor Variable Selection

Before the construction of the models, the predictor variables that most contributed to inter-class discrimination and thus to higher model performance were selected. This pre-processing task is essential because a predictive model with fewer variables may be more interpretable, and non-informative variables may negatively affect the model performance.

The *recursive feature elimination with random forest* (RFE-RF) method, implemented in the *caret*

package (Kuhn, 2021) for the R programing language (R Core Team, 2022), was applied to search for the best subset of predictors. According to Kuhn and Johnson (2016, p. 494-495), in RFE-RF, an initial model contains all predictors, which are then iteratively removed to determine which are not significantly contributing to performance. The elimination of the predictors is based on the variable importance criterion that provides the random forest model. This criterion ranks the predictors from the most important to least. Hence, at each stage of the search, the least important variables are iteratively disregarded before building the next model in the following stage.

On the one hand, we selected the best predictors from a set of four variables, namely, *LexDom*, *SemCat_Level_1*, *SemRol_mwt*, and *SemRol_river*. To preclude correlations between the predictors, *SemRel* was initially not included because it was a dependent variable that might be predicted from the other four variables, as previously seen in the analysis of pairwise associations between variables (Section 2.6). The presence of correlated predictors – situation called *collinearity* – poses problems for predictive models in determining how each one separately is associated with the outcome. Notwithstanding, the contribution of the *SemRel* variable to bracketing prediction will be examined later in this Section.

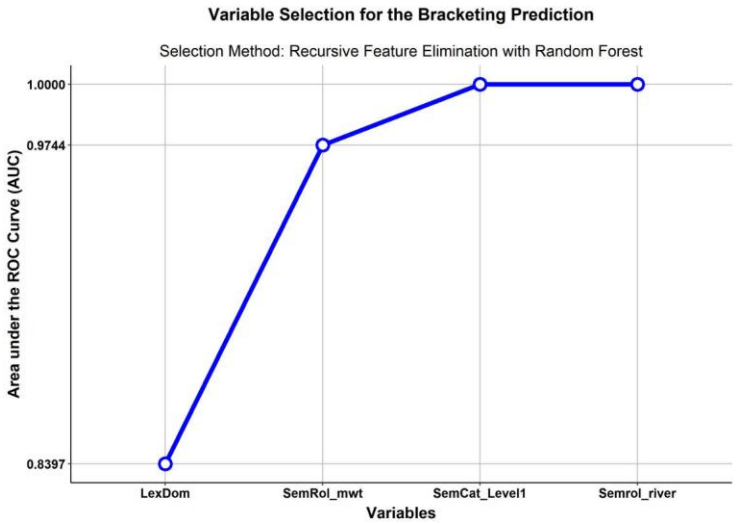
In the training dataset (with 266 instances, 70% of the original dataset), 7-fold cross-validation was used to evaluate the RFE-RF variable selection method. Although 10 folds are conventionally employed, we chose 7 folds, a divisor of 266, so that the number of instances in all folds would be the same (i.e., 38 instances). During the process, the AUC performance measure was chosen to be maximized. In other words, the RFE-RF selected the subset of predictors that reached the greatest AUC in the model performance. Two parameters were also set: (1) the splits in the trees were allowed to use one predictor of a subset of two predictors; and (2) the number of trees used in the forest was set to the default value of 500 trees.

The 7-fold cross-validation randomly divided the training dataset into 7 groups of equal size (38 instances). The first group was treated as a test dataset, and the model was fit to the other 6 groups. The AUC was then calculated on the instances in the held-out group. This procedure was repeated 7 times, but each time, a different group was treated as a test dataset. The process resulted in 7 estimates of the AUC, and the final AUC was calculated by averaging these values.

The RFE-RF method selected the predictors *LexDom*, *SemRol_mwt*, and *SemCat_Level_1*, and disregarded *SemRol_river*. This meant that the semantic role of the other argument, filled with a named river in our case, apparently exerted no influence on the prediction of ternary compound bracketing. As shown in Figure 2, the most important predictor was the lexical domain of the verb (*LexDom*), which would yield an AUC of 0,8397 if this were the only variable used for predictions. The second most important predictor was the semantic role of the MWT (*SemRol_mwt*). Hence, if both *LexDom* and *SemRol_mwt* were employed, the model would obtain a higher AUC, equal to 0,9744. Finally, the third most important predictor was the ontological category of the MWT (*SemCat_Level_1*). If *LexDom*, *SemRol_mwt*, and *SemCat_Level_1* were considered, the random forest model would not commit any prediction errors.

Figure 2.

Results of the variable selection method, RFE-RF, for bracketing prediction. The predictors LexDom, SemRol_mwt, *and* SemCat_Level_1 *were selected.*



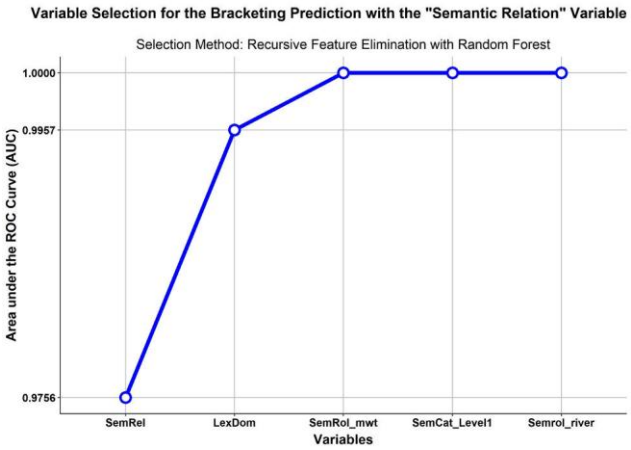
Source: Original analysis by the author.

On the other hand, we then selected the best predictors from a set of five variables, namely, *LexDom*, *SemCat_Level_1*, *SemRol_mwt*, *SemRol_river*, and *SemRel*. On this occasion, we decided to include the *SemRel* predictor to examine its contribution to bracketing prediction. Although *SemRel* was a correlated predictor, the RFE-RF method has been found to mitigate this problem in small datasets as ours (Gregorutti et al., 2017).

As such, the RFE-RF method selected the predictors *SemRel*, *LexDom*, and *SemRol_mwt*, and disregarded *SemCat_Level_1* and *SemRol_river*. This implied that, when the *SemRel* predictor received consideration, neither the ontological category of the MWT, nor the semantic role of the other argument, filled with a named river in our case, was significant for the prediction of ternary compound bracketing. Figure 3 illustrates that the most important predictor was the semantic relation between the MWT and the named river (*SemRel*), which would yield an AUC of 0,9756 if this were the only variable used for predictions. In our opinion, the enormous predictive power of the semantic relation was due to the fact that this predictor retains, to some extent, valuable information concerning the lexical domain of the verb and the semantic role of the MWT, as revealed in the previous analysis of pairwise association between variables. In other words, given that *SemRel* was correlated with *LexDom* and *SemRol_mwt*, it may be possible to predict *SemRel* from *LexDom* and *SemRol_mwt*. The second most important predictor was the lexical domain of the verb (*LexDom*). As a result, if both *SemRel* and *LexDom* were employed, the model would achieve a higher AUC, equal to 0,9957. Finally, the third most important predictor was the semantic role of the MWT (*SemRol_mwt*). If *SemRel*, *LexDom*, and *SemRol_mwt* were considered, the random forest model would preclude prediction errors.

Figure 3.

Results of the variable selection method, RFE-RF, for bracketing prediction when the SemRel *predictor was also included.*



Source: Original analysis by the author.

3. Results

3.1. Construction of the Models with the Predictors LexDom, SemRol_mwt, *and* SemCat_Level_1

The variable selection task established that the best predictors of the *Bracketing* variable could be grouped into two subsets. The first subset of the best predictors was composed of *LexDom*, *SemRol_mwt*, and *SemCat_Level_1*, which were used to construct two predictive models. For the random forest, 7-fold cross-validation in the training dataset (with 266 instances, 70% of the original dataset) was used to evaluate its performance in training. Although 10 folds are conventionally employed, we chose 7 folds, a divisor of 266, so that the number of instances in all folds would be the same (i.e., 38 instances). During the process of tuning parameters, the AUC performance measure was chosen to be maximized. Accordingly, the random forest model reached in training an AUC value equal to 1,0 when: (1) the splits in the trees were allowed to use one predictor of a subset of one predictor; and (2) the number of trees in the forest was, surprisingly, only three trees. In the test dataset (with 114 instances, 30% of the original dataset), the random forest also achieved an AUC value equal to 1,0.

Similarly, for the decision tree, 7-fold cross-validation in the training dataset was employed to evaluate its performance in training. During the process of tuning parameters, the AUC performance measure was also chosen to be maximized. Therefore, the decision-tree model yielded in training the greatest AUC, equal to 0,9781, when: (1) the cost-complexity parameter (*cp*), which controls a trade-off between tree complexity (i.e., number of terminal nodes) and its fit to the training data to avoid overfitting, was equal to *cp*=0,0625 (the higher the *cp*, the simpler the model); and (2) the splitting criterion for predictors was the information gain, and not the Gini index. In the test dataset, the decision-tree model achieved an outstanding AUC value equal to 0,9830. Table 6 provides further performance measures, in the training and test datasets, for the random forest and decision-tree models.

Table 6.

Performance measures of the models for bracketing prediction with the predictors lexical domain, semantic role, and semantic category.

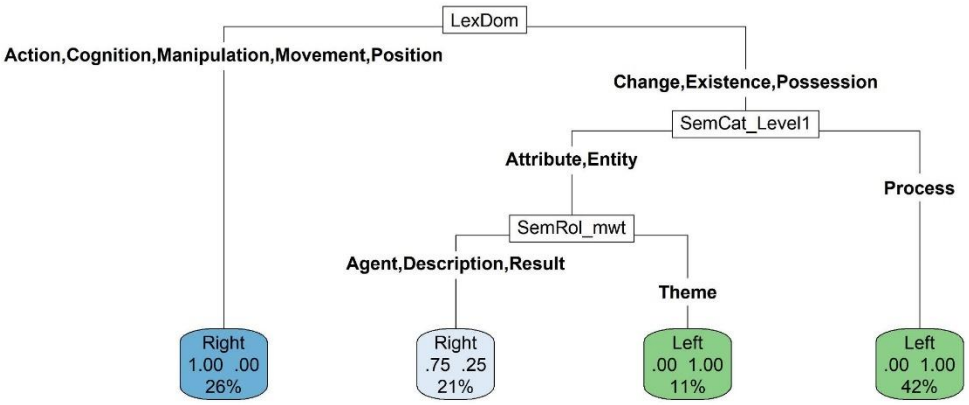
Predictors: *LexDom*, *SemRol_mwt*, and *SemCat_Level_1*							
	Decision Tree Model						
Dataset	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F_1**	**Overall Accuracy**
Training	0,9781	0,9464	0,9286	0,9165	0,9464	0,9249	0,9361
Test	**0,9830**	1,0000	0,9091	0,8889	1,0000	**0,9412**	**0,9474**
	Random Forest Model (3 ensembled decision trees)						
Training	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Test	**1,0000**	1,0000	1,0000	1,0000	1,0000	**1,0000**	**1,0000**

Source: Original analysis by the author.

Since the decision-tree model reached an extraordinarily high AUC in the test dataset (AUC=0,9830), its prediction rules, graphically summarized in Figure 4, deserved attention.

Figure 4.

Classification tree for bracketing prediction, inferred by the decision-tree model, trained with the predictors lexical domain, and semantic role and category of the MWTs.



Source: Original analysis by the author.

In our constrained context (i.e., specialized ternary compounds from Coastal Engineering, used in sentences where a named river was mentioned), the classification tree of the model (see Figure 4) can be interpreted as follows. *LexDom* was the most important factor in determining *Bracketing*. The ternary compounds used in sentences where the verb belonged to either ACTION, COGNITION, MANIPULATION, MOVEMENT, or POSITION embraced 26% of the sample; these MWTs were all right-branched and correctly classified. Thus, given those five lexical domains, the semantic category and role of the MWTs did not seem to influence the bracketing-structure prediction.

Nevertheless, the predictions became more complicated when MWTs were employed with verbs belonging to the domains of CHANGE, EXISTENCE, or POSSESSION. The MWTs that appeared with verbs from those lexical domains, and which were also processes, constituted 42% of the sample; these MWTs were all left-branched and correctly classified. However, when the MWTs were attributes or entities in combination with the previously mentioned lexical domains, their semantic roles were necessary to disambiguate bracketing structures. As such, the MWTs with the role of THEME, in this setting, comprised 11% of the sample; these MWTs

were all left-branched and correctly classified. In contrast, the MWTs with the role of either AGENT, DESCRIPTION, or RESULT constituted 21% of the sample and could be right- or left-branched. In these conditions, the model correctly classified all the right-branched MWTs (75%), but misclassified the true left-branched MWTs (25%) as right-branched.

An analysis of the errors of the decision-tree model revealed that, both in the training and test datasets, those left-branched MWTs with the values *LexDom*= CHANGE, *SemCat_Level_1*= ENTITY, and *SemRol_mwt*= AGENT (e.g., *vegetation removal effect*, in row 10 of Table 3), were all misclassified as right-branched.

3.2. Construction of the Models with the Predictors SemRel, LexDom, *and* SemRol_mwt

The variable selection task established that the best predictors of the *Bracketing* variable could be grouped into two subsets. The second subset of the best predictors was composed of *SemRel*, *LexDom*, and *SemRol_mwt*, which were used to construct two predictive models. For the random forest, 7-fold cross-validation in the training dataset was used to evaluate its performance in training. During the process of tuning parameters, the AUC performance measure was chosen to be maximized. Accordingly, the random forest model attained in training an AUC value equal to 1,0 when: (1) the splits in the trees were allowed to use one predictor of a subset of one predictor; and (2) the number of trees in the forest was, surprisingly, only three trees. In the test dataset, the random forest also achieved an AUC value equal to 1,0. Consequently, both the first subset of the best predictors (*LexDom*, *SemRol_mwt*, and *SemCat_Level_1*) and the second one (*SemRel*, *LexDom*, and *SemRel_mwt*) were capable of correctly predicting bracketing in the test dataset with a random forest model.

Similarly, for the decision tree, 7-fold cross-validation in the training dataset was employed to evaluate its performance in training. During the process of tuning parameters, the AUC performance measure was also chosen to be maximized. Therefore, the decision-tree model yielded in training the greatest AUC, equal to 0,9643, when: (1) the cost-complexity parameter (*cp*) was equal to *cp*=0,05357143; and (2) the splitting criterion for predictors was the information gain, and not the Gini index. In the test dataset, the decision-tree model achieved an AUC value equal to 0,9545, which indicated a very satisfactory performance, though not as good as that of the decision tree with the first subset of predictors (AUC=0,9830). Table 7 provides further performance measures, in the training and test datasets, for the random forest and decision-tree models.

Table 7.

Performance measures of the models for bracketing prediction with the predictors semantic relation, lexical domain, and semantic role.

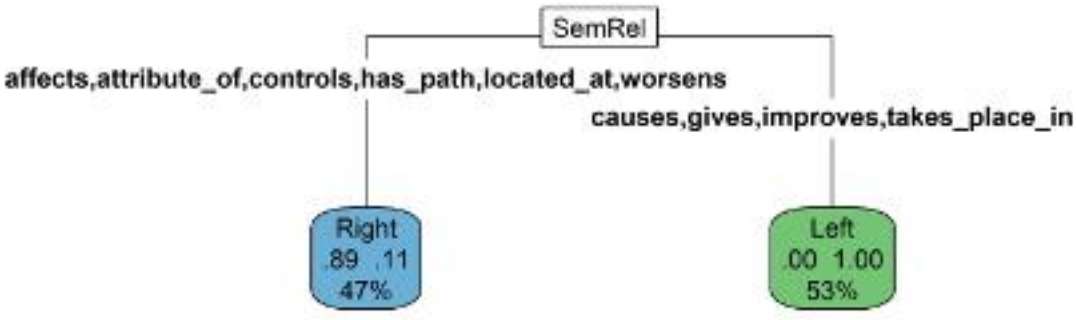
Predictors: *SemRel*, *LexDom*, and *SemRol_mwt*							
	Decision Tree Model						
Dataset	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**F_1**	**Overall Accuracy**
Training	0,9643	1,0000	0,9286	0,9206	1,0000	0,9562	0,9586
Test	**0,9545**	1,0000	0,9091	0,8889	1,0000	**0,9412**	**0,9474**
	Random Forest Model (3 ensembled decision trees)						
Training	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Test	**1,0000**	1,0000	1,0000	1,0000	1,0000	**1,0000**	**1,0000**

Source: Original analysis by the author.

Since the decision-tree model reached a significant AUC in the test dataset (AUC=0,9545), its only prediction rule, graphically summarized in Figure 5, is worth mentioning.

Figure 5.

Classification tree for bracketing prediction, inferred by the decision-tree model, trained with the predictors semantic relation, lexical domain, and semantic role of the MWTs.



Source: Original analysis by the author.

SemRel emerged as the most crucial factor in determining *Bracketing*, and it was the sole predictor chosen by the decision-tree model. We believe that the predictive strength of the semantic relation between an MWT and another argument within the same sentence is so significant that the model had to exclude the predictors *LexDom* and *SemRol_mwt* to prevent overfitting the training data. Consequently, ternary compounds whose semantic relation to the other argument (in our case, a named river) fell into the set of *causes*, *gives*, *improves*, and *takes_place_in* made up 53% of the sample. These MWTs were consistently left-branched and accurately classified. Thus, it appears that these four semantic relations necessitated the exclusive use of left-branched MWTs.

Conversely, ternary compounds with a semantic relation to the other argument falling into the group of *affects*, *attribute_of*, *controls*, *has_path*, *located_at*, and *worsens* accounted for 47% of the sample. These could be either right- or left-branched. Under these conditions, the model accurately classified all the right-branched MWTs (89%), but incorrectly classified the true left-branched MWTs (11%) as right-branched.

An analysis of the errors made by the decision-tree model revealed that, both in the training and test datasets, those left-branched MWTs with the values *SemRel= attribute_of*, *LexDom=* EXISTENCE, *SemRol_mwt=* THEME, and *SemCat_Level_1=* ATTRIBUTE (e.g., *water discharge level*, in row 8 of Table 3), were all misclassified as right-branched.

4. Discussion

For predicting bracketing, past research has focused on the semantic information derived from the components of an MWT. The number of variables used for prediction varied significantly, ranging from 12 to 517.254 features. These variables were primarily based on n-gram statistics, which are thought to capture some semantic information through the frequent co-occurrence of MWT components (Lazaridou et al., 2013, p. 1909). Other studies utilized semantic information from linguistic resources like WordNet. The overall accuracy of these prediction models varied between 72,60% and 95,40%.

Our semantic approach, however, utilized types of semantic information that had not been considered in previous research. This information was derived from both the context in which a ternary compound appeared (i.e., the sentence where the ternary compound was used) and the ternary compound itself as a unit (i.e., its semantic role and category). Initially, we

considered only five variables: (1) semantic relation, (2) lexical domain, (3) semantic role of the MWT, (4) semantic category of the MWT, and (5) semantic role of the named river. However, it turned out that only a subset of three variables was necessary for accurate bracketing prediction, whereas earlier studies used at least 12 variables (León-Araúz et al., 2021). This three-variable subset achieved perfect accuracy on the test dataset using a random forest model, outperforming previous research, which had achieved a maximum overall accuracy of 95,40% using a support-vector machine (Pitler et al., 2010), a less interpretable model.

5. Conclusion

This study provided support for our hypothesis that semantic information within a sentence — specifically, the lexical domain of the verb, the semantic role and category of the ternary compound, and its semantic relation to another argument in the same sentence — significantly aids in parsing compounds. Considering the positive impact of multiword-term bracketing on the overall accuracy of sentence parsers (Vadas & Curran, 2008) and machine-translation systems (Garg et al., 2022), this finding suggests a promising new direction for incorporating such semantic variables into syntactic parsers and machine-translation applications, aligning with the work of Agirre et al. (2008) and Girju et al. (2005).

The pairwise associations between the semantic variables annotated were statistically analyzed with Fisher's exact test. The analysis indicated that all variables were statistically associated, except for *bracketing* and *semantic role* of the river. This signified that the semantic role of other arguments in the same sentence was not necessary for the bracketing prediction of ternary compounds. Nevertheless, the bracketing variable was statistically associated with the other four variables (i.e., *lexical domain*, *MWT role*, *MWT category*, and *semantic relation*), which could predict bracketing in our sample of ternary compounds.

The *semantic relation* variable maintained a strong association with all the other predictor variables (i.e., *lexical domain*, *MWT role*, *MWT category*, and *river role*). This finding suggests that the semantic relation between two arguments in a sentence condenses information about the lexical domain of the verb, the semantic categories of both arguments, and their semantic roles. In other words, *semantic relation* is a dependent variable that may be predicted from *lexical domain*, *MWT role*, *MWT category*, and *river role*.

This study demonstrated that semantic information within a sentence significantly aids in compound bracketing. However, it remains uncertain whether the semantic variables examined here can also predict the bracketing of MWTs with four or more constituents, a topic for future research. Despite the promising results, they should be viewed with caution due to the small sample size (380 ternary compounds) and the limited scope of the analysis, which focused on specialized ternary compounds from Coastal Engineering used in sentences mentioning named rivers. Future research should adopt a broader framework to gain a deeper understanding of how these semantic variables influence multiword-term bracketing.

6. References

Agirre, E., Baldwin, T., & Martinez, D. (2008). Improving parsing and PP attachment performance with sense information. In *Proceedings of the 46th Annual Meeting of the ACL* (pp. 317-325). ACL. https://cutt.ly/uekWpF6t

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Faber, P. (2015). Frames as a framework for Terminology. In H. Kockaert, & F. Steurs (Eds.), *A Handbook of Terminology* (pp. 14-33). John Benjamins. https://doi.org/10.1075/hot.1.fra1

Faber, P., & Mairal, R. (1999). *Constructing a Lexicon of English Verbs*. Mouton de Gruyter.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer.

Garg, K.D., Shekhar, S., Kumar, A., Goyal, V., Sharma, B., Chengoden, R., & Srivastava, G. (2022). Framework for Handling Rare Word Problems in Neural Machine Translation System Using Multi-Word Expressions. *Applied Sciences*, *12*, 11038. https://cutt.ly/Jekhbfpg

Girju, R., Moldovan, D.I., Tatu, M., & Antohe, D. (2005). On the semantics of noun compounds. *Computer Speech and Language*, *19*(4), 479-496. https://doi.org/10.1016/j.csl.2005.02.006

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27, 659–678. https://cutt.ly/FYFYCbF

Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, *19*(1), 103-120. https://aclanthology.org/J93-1005.pdf

Klie, J.C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference COLING 2018* (pp. 5-9). ACL. https://cutt.ly/UekjrFRT

Krippendorff, K. (2012). *Content Analysis: An Introduction to its Methodology*. Sage.

Kroeger, P.R. (2005). *Analyzing Grammar: An Introduction*. Cambridge University Press.

Kuhn, M. (2021). *caret: Classification and Regression Training*. R package. https://cutt.ly/aekhnQuG

Kuhn, M., & Johnson, K. (2016). *Applied Predictive Modeling*. Springer.

Lauer, M. (1994). Conceptual Association for Compound Noun Analysis. In *Proceedings of the 32nd Annual Meeting of the ACL* (pp. 337-339). CoRR. https://cutt.ly/CekhxzZ5

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds*. (Ph.D. Thesis). Macquarie University, Sidney. https://arxiv.org/abs/cmp-lg/9609008

Lazaridou, A., Vecchi, E.M., & Baroni, M. (2013). Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proceedings of the 2013 Conference on EMNLP* (pp. 1908-1913). ACL. https://aclanthology.org/D13-1196

León-Araúz, P., Cabezas-García, M., & Faber, P. (2021). Multiword-term bracketing and representation in terminological knowledge bases. In *Proceedings of the eLex 2021 Conference* (pp. 139-163). Lexical Computing CZ. https://cutt.ly/oRbV3LW

León-Araúz, P., Reimerink, A., & Faber, P. (2019). EcoLexicon and by-products: Integrating and reusing terminological resources. *Terminology*, *25*(2), 222-258. https://cutt.ly/BekhlSbR

Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language*. MIT Press.

Pitler, E., Bergsma, S., Lin, D., & Church, K.W. (2010). Using web-scale n-grams to improve base NP parsing performance. In *Proceedings of the 23rd International Conference COLING 2010* (pp. 886-894). ACL. https://cutt.ly/PekhlpjN

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rojas-Garcia, J. (2022a). Semantic Representation of Context for Description of Named Rivers in a Terminological Knowledge Base. *Frontiers in Psychology*, *13*, 847024. https://cutt.ly/0ekjeN3q

Rojas-Garcia, J. (2022b). Semantic Relations Predict the Bracketing of Three-Component Multiword Terms. *Procesamiento del Lenguaje Natural*, *69*, 141-152. https://shorturl.at/bdEG6

San Martín, A., Cabezas-García, M., Buendía-Castro, M., Sánchez-Cárdenas, B., León-Araúz, P., Reimerink, A., & Faber, P. (2020). Presente y futuro de la base de conocimiento terminológica EcoLexicon. *Onomázein*, *49*, 174-202. https://doi.org/10.7764/onomazein.49.09

Thompson, P., Iqbal, S.A., McNaught, J., & Ananiadou, S. (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, *10*, 349. https://doi.org/10.1186/1471-2105-10-349

Vadas, D., & Curran, J.R. (2008). Parsing noun phrase structure with CCG. In *Proceedings of the 46th Annual Meeting of the ACL* (pp. 335-343). ACL. https://aclanthology.org/P08-1039.pdf

AUTHOR CONTRIBUTIONS, FUNDING, AND ACKNOWLEDGEMENTS

Author contributions: The author confirms being the sole contributor of this work and has approved it for publication.

Funding: This research was carried out as part of the project PID2020-118369GB-I00, "Transversal Integration of Culture in a Terminological Knowledge Base on Environment" (TRANSCULTURE), funded by the Spanish Ministry of Science and Innovation. Funding was also provided by an FPU grant given by the Spanish Ministry of Education to the author.

Acknowledgements: I am deeply grateful to reviewers and editors, whose perceptive comments helped to enhance this paper.

Conflict of interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR:

Juan Rojas-García
University of Granada.

Juan Rojas-García is Junior Lecturer at the University of Granada, holds a PhD in Translation and Interpreting from the University of Granada, and a Master's degree in Teaching Spanish as a Foreign Language, a Master's degree in Teaching Spanish Language and Literature in Secondary Education, and a Master's degree in Data Science. In addition, he holds a degree in Telecommunication Systems Engineering from the University of Malaga. He is a member of the research group LexiCon (University of Granada). He completed a doctoral thesis in the field of Terminology, for which he received a research grant (FPU) from the Spanish Ministry of Economy and Competitiveness. His research areas are terminology, representation of named entities in terminological knowledge bases, text mining, and employability of Translation and Interpreting students. He has presented and published research papers on these areas at international conferences and in journals on applied linguistics, natural language processing, and translatology.

juanrojas@ugr.es

Orcid ID: https://orcid.org/0000-0002-7611-1386