

**ESTUDIO DEL ORDEN DE COMPOSICIÓN EN EL
MARCO TEÓRICO ICDS:
PROPUESTA DE ORDEN PARA ORACIONES SIMPLES**



Trabajo de Fin de Máster

Carmen Vázquez García

Trabajo de investigación para el
Máster en Tecnologías del Lenguaje

Universidad Nacional de Educación a Distancia

Dirigido por:

Prof. Dr. D. Víctor Fresno Fernández

Junio 2024

Resumen

En el ámbito del procesamiento del lenguaje natural (PLN), la representación del significado textual es un problema fundamental que implica codificar el lenguaje natural de manera que pueda ser manejado eficazmente por los sistemas de gestión de información. Esta investigación presenta una exploración integral de la semántica distribucional composicional basada en la teoría de la información (ICDS), un marco que tiene como objetivo integrar los principios de la hipótesis distribucional y el principio de composicionalidad. El objetivo es cerrar la brecha entre el espacio de representación y la teoría de la información, proporcionando restricciones formales para las funciones de *embedding*, composición y similitud.

Este trabajo plantea la incorporación del orden sintáctico en las representaciones semánticas dentro del marco ICDS. Esto se debe a que en este marco ya se ha expuesto que el orden influye en la representación semántica, por lo que se ha planteado una propuesta de orden basada en el orden sintáctico. A través de una revisión del estado del arte dentro del marco ICDS, pero desde una perspectiva lingüística, se ha planteado que el orden y la estructura inherentes en los elementos lingüísticos podrían influir significativamente en la representación semántica.

Inicialmente, esta investigación se centra en oraciones simples para establecer un enfoque fundamental de representación semántica. Sin embargo, también se plantea la posibilidad de extender el marco para incluir oraciones subordinadas, que funcionan como sustantivos, adjetivos o adverbios dentro de las oraciones principales. Cada oración subordinada posee su propia estructura interna e interactúa jerárquicamente con la oración principal, por lo que el enfoque es extrapolable a otro tipo de oraciones más complejas.

El propuesto basado en el orden enfatiza la importancia de la estructuración sintáctica en la mejora de la coherencia semántica. Al basar el método en cómo se codifican y representan los elementos lingüísticos, se busca desarrollar un modelo que capture las complejas relaciones dentro y entre las oraciones. Este enfoque no solo se alinea con la teoría lingüística, sino que también aborda los desafíos prácticos en el PLN. Los hallazgos de esta investigación destacan la naturaleza esencial del orden sintáctico en la representación semántica y proponen una metodología estructurada para integrar estos elementos dentro del marco ICDS.

Palabras clave: marco ICDS, representación semántica, word embeddings

Abstract

In the field of Natural Language Processing (NLP), the representation of text meaning is a fundamental issue that involves encoding natural language in a way that can be effectively handled by information management systems. This research presents a comprehensive exploration of Information Theory-based Compositional Distributional Semantics (ICDS), a framework that aims to integrate the principles of the Distributional Hypothesis and the Principle of Compositionality. The goal is to bridge the gap between the representation space and Information Theory, providing formal constraints for embedding, composition, and similarity functions.

This work proposes the incorporation of syntactic order in semantic representations within the ICDS framework. This is because it has already been shown within this framework that order influences semantic representation, leading to a proposed order based on syntactic order. Through a review of the state of the art within the ICDS framework but from a linguistic perspective, it has been suggested that the inherent order and structure in linguistic elements could significantly influence semantic representation.

Initially, this research focuses on simple sentences to establish a fundamental approach to semantic representation. However, it also considers the possibility of extending the framework to include subordinate clauses, which function as nouns, adjectives, or adverbs within main clauses. Each subordinate clause has its own internal structure and interacts hierarchically with the main clause, making the approach able to be extrapolated to more complex types of sentences.

The proposed order-based approach emphasizes the importance of syntactic structuring in improving semantic coherence. By basing the method on how linguistic elements are encoded and represented, the goal is to develop a model that captures the complex relationships within and between sentences. This approach not only aligns with linguistic theory but also addresses practical challenges in NLP, where representations are expected to be both accurate and contextually relevant. The findings of this research highlight the essential nature of syntactic order in semantic representation and propose a structured methodology to integrate these elements within the ICDS framework.

Keywords: ICDS framework, semantic representation, word embeddings

Índice

1. Introducción	1
1.2. Hipótesis y objetivos	5
1.3. Contribuciones	6
1.4. Estructura del documento.....	7
2. Estado del arte/Related work	9
2.1. Marco de la función ICDS	10
2.2 Semántica	16
2.3 Composicionalidad.....	20
2.4 Distribucionalidad y síntesis kantiana.....	27
2.5 Fundamentos lógico-lingüísticos de las oraciones	29
2.6 Estructura y tipología sintáctica: oraciones y componentes lingüísticos	36
2.7 Conclusiones	45
3. Propuesta de modelo de orden	47
3.1. Introducción	47
3.2. Propuesta de orden jerárquico para oraciones simples.....	48
3.2. Aplicabilidad y escalabilidad de la propuesta	53
3.3. Aplicabilidad a otras lenguas	54
4. Metodología	56
4.1. <i>Dataset</i>	57
4.2. GloVe	59
4.3. <i>Baselines</i>	60
4.3. Funciones de composición	61
4.4. Métrica de evaluación	63
4.5. Descripción del experimento.....	64
4.5.1 Diseño del experimento.....	64
5. Experimentos	67
5.1. Experimento 1	67
5.1.1 Resultados	68
5.2. Experimento 2	75
5.2.1 Resultados	75
6. Conclusiones y futuras vías de estudio	81
6.1. Conclusiones	81
6.2. Futuras vías de estudio	85
Bibliografía	87

1. Introducción

En el ámbito del procesamiento del lenguaje natural (PLN), representar el significado del texto requiere codificar el lenguaje natural de tal manera que los sistemas de información puedan gestionarlo eficazmente. En este contexto, la semántica distribucional composicional basada en la teoría de la información (Information Theory-based Compositional Distributional Semantics, ICDS) emerge como un marco teórico innovador, combinando los principios de la teoría de la información de Shannon con la semántica distribucional y composicional para abordar estos desafíos.

El marco ICDS, propuesto por Amigó et al. en su artículo Information Theory-based Compositional Distributional Semantics (2022), se basa en la integración de los principios de la hipótesis distribucional y el principio de composicionalidad. La hipótesis distribucional sostiene que el significado de una palabra está determinado por su contexto de uso, mientras que el principio de composicionalidad establece que el significado de una expresión compleja se deriva del significado de sus partes y de la forma en que estas se combinan sintácticamente. Al fusionar estos principios, ICDS busca capturar tanto la contextualidad como la sistematicidad del lenguaje (Amigó et al., 2022).

Para abordar la complejidad de combinar representaciones semánticas de unidades lingüísticas en el marco ICDS, Amigó et al. (2022) desarrollaron una función de composición específica. Esta función es fundamental no solo para capturar el significado contextual de las palabras, sino también para preservar la estructura sintáctica y la integridad semántica del texto. Esto se debe a que se construye sobre una serie de propiedades formales deseables que debería cumplir todo espacio de representación semántico. La función de composición generalizada $F_{\lambda,\mu}$ se diseñó para combinar vectores de manera que reflejen adecuadamente tanto la contextualidad como la composicionalidad. Por tanto, para poder reflejar el espacio semántico, en el marco ICDS se trabaja con vectores estáticos. Esto es esencial para el funcionamiento efectivo del marco ICDS, porque, en principio, sus espacios de representación cumplen más las propiedades. La relación de los vectores de palabras estáticos dentro de su espacio de representación es isométrica con la similitud semántica, es decir, que las distancias relativas entre los vectores se conservan entre el espacio de embedding y los significados. Los vectores estáticos elegidos para el estudio de Amigó et al. (2022) fueron Word2Vec,

ya que permiten una representación semántica más estable y coherente en comparación con los vectores dinámicos de modelos como BERT. Como ya se mencionaba, esto se debe a que ICDS opera mejor con vectores estáticos, ya que estos reflejan de manera más precisa el espacio semántico necesario para capturar las relaciones contextuales y composicionales del lenguaje.

A raíz de estos planteamientos, el artículo se propone crear una función de composición que combine los vectores de las palabras de manera iterativa, asegurando que cada paso de la composición tenga en cuenta las propiedades semánticas de los pares de vectores involucrados. La función de composición generalizada $F_{\lambda,\mu}$ se define como:

$$F_{\lambda,\mu}(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|} \cdot \sqrt{\lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle}$$

Esta función combina dos vectores la siguiente manera:

1. **Vector de dirección:** La primera parte de la expresión normaliza la suma de los vectores, determinando la dirección del vector resultante.

$$\frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|}$$

2. **Vector de magnitud:** La segunda parte ajusta la magnitud del vector en función de las normas de los vectores individuales y su producto interno.

$$\sqrt{\lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle}$$

Esta función generalizada permite derivar varias funciones de composición comunes en la literatura, como la suma de vectores o el promedio por pares, que se comentarán más adelante.

En el contexto del marco ICDS, la función de composición opera combinando pares de vectores de manera iterativa. Al combinar pares de vectores, se puede controlar mejor el proceso de composición, asegurando que se mantengan las relaciones semánticas y contextuales en cada paso de la iteración.

Por ejemplo, dada una secuencia de tres vectores:

1. Primero se combinan el primer y segundo vector utilizando la función $F_{\lambda, \mu}$, lo que resulta en un nuevo vector compuesto por estos dos vectores.
2. Luego, el vector compuesto por el primer y segundo vector se combina con tercero, para obtener un único vector final compuesto por ambos vectores

Este enfoque iterativo por pares garantiza que cada paso de la composición tenga en cuenta las propiedades semánticas de los pares de vectores involucrados, asegurando así una representación semántica coherente y precisa. Por tanto, en este marco, la función de composición es fundamental para componer representaciones semánticas de unidades lingüísticas. Para ilustrar la importancia del orden de composición, se va a considerar a continuación cómo las funciones de composición operan en el marco ICDS. Estas funciones combinan vectores de palabras para formar nuevas representaciones compuestas. La dirección y la magnitud del vector compuesto dependen del orden en que se combinan los vectores. Por ejemplo, en una oración como «El perro negro corre rápidamente», la composición de los vectores correspondientes a «perro» y «negro» antes de «corre» y «rápidamente» podría producir una representación diferente a la obtenida si se combinan en un orden distinto. Esta variabilidad, por tanto, afecta la precisión y coherencia de las representaciones semánticas resultantes.

Con la finalidad de llevar a cabo distintas pruebas, la función de composición generalizada $F_{\lambda, \mu}$ se utiliza con distintos valores y configuraciones para evaluar su efectividad en mantener la integridad semántica y estructural en las representaciones resultantes. Este enfoque permite explorar cómo diferentes parametrizaciones de la función afectan la precisión y coherencia de las representaciones semánticas. Por ejemplo, variaciones en los parámetros λ y μ pueden ajustar la influencia relativa de la magnitud y dirección de los vectores compuestos, lo que a su vez impacta la forma en que se preserva la información contextual y sintáctica de las palabras originales.

Por tanto, Amigó et al. (2022) establecen que el orden de composición tiene un impacto significativo en la representación semántica. El análisis detallado en este trabajo ha demostrado que distintos órdenes de composición pueden producir vectores compuestos con propiedades diversas, lo que resalta la necesidad de definir un orden claro y consistente en el proceso de composición. Este hallazgo es esencial dentro de este marco

de trabajo, ya que evidencia que la manera en que se estructuran y combinan las palabras en una oración influye directamente en la representación final dentro del espacio semántico.

De hecho, es por esto que en el enfoque ICDS se reconoce que la función de composición debe ser sensible a la estructura lingüística. Esto significa que el orden en el que se combinan las unidades lingüísticas tiene un impacto significativo en la representación semántica resultante. Sin embargo, uno de los problemas abiertos en este marco es que, aunque se es consciente de que el orden marca una diferencia a la hora de componer, aún no se ha establecido una propuesta definitiva sobre cómo determinar dicho orden. Esta falta de claridad crea una necesidad crítica de desarrollar y establecer un método estándar para manejar el orden en la composición semántica dentro del marco ICDS. La ausencia de una propuesta clara sobre el orden en el marco ICDS es un problema significativo, ya que, sin un orden establecido, la representación semántica no es sistemática (Cummins, 1996). Por tanto, establecer un orden claro debería permitir que las funciones de composición mantengan las relaciones sintácticas y la estructura del texto, lo que resultará en una representación semántica más coherente y precisa.

Siguiendo con el contexto del trabajo desarrollado por Amigó et al. en su artículo *Information Theory-based Compositional Distributional Semantics* (2022), el enfoque inicial se ha centrado en la composición de vectores a nivel de palabras. Este marco teórico ha demostrado ser eficaz para capturar tanto la contextualidad como la composicionalidad del significado de las palabras individuales. Sin embargo, para avanzar en la representación semántica y abordar problemas más complejos en el procesamiento del lenguaje natural, se plantea extender esta metodología a la composición de oraciones en el mismo espacio de representación semántica de word embeddings.

La motivación principal de este trabajo es precisamente establecer un orden claro y consistente para la composición de oraciones en el marco ICDS. La integración de oraciones en el espacio semántico presenta desafíos adicionales en comparación con la composición de palabras, ya que las oraciones tienen una estructura sintáctica más compleja y una mayor cantidad de información contextual que debe ser preservada para mantener la coherencia y precisión semántica. En consecuencia, para asegurar que las representaciones semánticas de las oraciones sean precisas y útiles, es importante

determinar un método adecuado para componer los vectores de las palabras que conforman la oración, de forma en la que se plantee un modelo sistemático.

Como resultado de estos planteamientos, en este trabajo se busca desarrollar una propuesta de orden aplicable a las oraciones simples (y adaptable a otro tipo de oraciones), que permita pasar del nivel de palabra al nivel de oración de manera efectiva y sistemática. Esta propuesta se basa en la necesidad de una función de composición que pueda manejar eficazmente el orden y la estructura de las palabras dentro de las oraciones. Al establecer un método claro para determinar el orden de composición, se espera lograr una representación semántica más consistente y coherente, lo que a su vez facilitará tareas dentro del ámbito del procesamiento del lenguaje natural.

1.2. Hipótesis y objetivos

En el marco de la representación textual mediante vectores de ICDS, es fundamental explorar y profundizar en la composicionalidad semántica y su relación con el orden de composición. A partir de esta premisa, se plantean diversas hipótesis y enfoques que servirán como punto de partida para abordar preguntas clave de investigación. Estas guiarán el análisis y evaluación de las técnicas de composición, así como su impacto en la precisión y coherencia de las representaciones semánticas.

Hipótesis:

1. H1: ¿Existe un orden intrínseco en la composición semántica de oraciones que impacta significativamente en la representación textual mediante vectores dentro del marco teórico ICDS (cambiarlo en capítulo 6)?
2. H2: ¿La comprensión de este orden inherente contribuye de manera sustancial a la mejora de la representación de mensajes en espacios vectoriales semánticos dentro del marco teórico ICDS?
3. H3: Según el marco presentado por Amigó et al., el orden en la composición de vectores es esencial para la estructuración y representación de palabras en espacios vectoriales. ¿Sería por tanto posible plantear una propuesta de orden de oraciones que sea capaz de capturar la semántica compuesta de la oración?

Como respuesta a las hipótesis planteadas, se establecen metas específicas que incluyen la exploración del límite de la composicionalidad, la definición de un enfoque de composición semántica para oraciones y palabras y la evaluación de diferentes estrategias en la mejora de la representación textual.

1. Explorar el límite de la composicionalidad semántica desde una perspectiva lingüística, identificando patrones de orden que afectan la representación de palabras y oraciones.
2. Definir un enfoque de composición semántica que permita la representación de oraciones simples, considerando el orden como un factor esencial.
3. Evaluar la efectividad de diferentes enfoques de composición semántica en la mejora de la representación textual, utilizando métricas específicas para medir la coherencia y la utilidad de las representaciones generadas.
4. Contribuir al avance en la comprensión de la composicionalidad semántica no supervisada dentro del modelo ICDS y su aplicabilidad en la representación textual, ofreciendo *insights* para futuras investigaciones en el campo del procesamiento del lenguaje natural.
5. Evaluar la viabilidad y el valor de la propuesta en oraciones simples, con el objetivo de determinar su aplicabilidad futura y eficacia en oraciones más complejas.

1.3. Contribuciones

Este trabajo pretende ofrecer una perspectiva esclarecedora a la comunidad científica en el ámbito de la representación semántica textual, con un enfoque particular en la composicionalidad en la tarea de representar oraciones a través de vectores de palabras estáticos. La idea principal que se subraya es que el orden en la composición de estos vectores, basados en la composicionalidad, es esencial para la estructuración de las oraciones. Esto resalta la importancia de considerar la secuencia y disposición de las palabras al representar textos de manera semántica.

La contribución principal de este trabajo radica en la propuesta de un orden específico para la composición de oraciones en el marco ICDS (Semántica Distribucional

Composicional basada en la Teoría de la Información). La integración de oraciones requiere una función de composición que maneje eficazmente el orden y la estructura de las palabras para asegurar que las representaciones semánticas sean precisas y útiles. Por tanto, establecer un método claro para determinar este orden permitirá una representación más consistente de las oraciones, facilitando comparaciones y análisis más precisos.

1.4. Estructura del documento

El resto del trabajo se divide en varios capítulos que detallan desde la revisión bibliográfica hasta la propuesta de modelos, los experimentos realizados y las conclusiones obtenidas.

El capítulo 2 proporciona una revisión del estado del arte y *related work*, comenzando por la presentación del marco teórico de ICDS. Por otro lado, se exploran las teorías fundamentales y los modelos actuales en el estudio del significado lingüístico, profundizando en cómo se representan y procesan las palabras y las oraciones. Posteriormente, se revisa la literatura sobre la teoría de la composicionalidad, destacando cómo el significado de las oraciones se deriva de sus partes constitutivas y de las reglas sintácticas que las combinan. Asimismo, se examinan los principios de la distribucionalidad, que postulan que el significado de una palabra puede inferirse de los contextos en los que aparece. Este análisis se complementa con una discusión sobre la síntesis kantiana y cómo esta integra las ideas distribucionales. El capítulo también contextualiza la función lingüística en la representación de oraciones, explorando cómo se relaciona con las estructuras sintácticas y semánticas. Para finalizar, se lleva a cabo una presentación de los elementos principales de las oraciones para poder asentar las bases de la propuesta de orden basada en el árbol sintáctico-jerárquico.

El capítulo 3 desarrolla la propuesta de un modelo que integra el orden de las palabras en la composición semántica, centrado específicamente en oraciones simples Y basándose en las hipótesis planteadas inicialmente en el marco de ICDS. Se detalla la metodología adoptada para desarrollar este modelo, describiendo su arquitectura y cómo se integra el orden de los elementos de la oración para mejorar la representación semántica. Además, se analiza la aplicabilidad de este modelo a otras lenguas y tipos de oraciones, considerando las particularidades lingüísticas y cómo estas pueden influir en la eficacia del modelo propuesto.

El capítulo 4 describe la metodología empleada para la investigación, comenzando con una presentación del conjunto de datos utilizado. Se detallan las características del *dataset* y su relevancia para las tareas de representación semántica. Aunque en el trabajo de Amigó et al. se utilizó Word2Vec, en esta investigación se ha recurrido a GloVe de 300 dimensiones para probar otro tipo de vector que refleje el espacio semántico. Tras ello, se describen los *baselines* utilizados para comparar el rendimiento del modelo propuesto, proporcionando un marco de referencia para evaluar su efectividad. También se explican las funciones de composición utilizadas en el modelo y cómo estas contribuyen a la integración del orden de las palabras en la representación semántica. Finalmente, se definen las métricas de evaluación que se utilizarán para medir la coherencia y la utilidad de las representaciones generadas, asegurando una evaluación rigurosa y precisa del modelo.

El capítulo 5 se enfoca en el diseño del experimento, describiendo en detalle cómo se llevarán a cabo los ensayos para evaluar el modelo propuesto. Se explican los procedimientos y las condiciones experimentales, se presentan los diferentes experimentos diseñados para probar las hipótesis planteadas y se describe cómo se medirán los resultados. Para finalizar, se exponen los resultados de las pruebas realizadas, proporcionando un análisis detallado de los datos obtenidos. Asimismo, se comparan los resultados del modelo propuesto con los otros *baselines*, evaluando su rendimiento y efectividad.

Finalmente, en el capítulo 6 se concluye el trabajo y se plantean futuras vías de investigación. Se resumen las conclusiones principales derivadas de los resultados experimentales y se discuten las contribuciones del trabajo al campo del procesamiento del lenguaje natural. Además, se proponen direcciones futuras para la investigación, sugiriendo áreas donde se pueden realizar mejoras y nuevas exploraciones basadas en los hallazgos obtenidos.

2. Estado del arte/*Related work*

En el marco de la representación textual mediante vectores dentro del ICDS, es esencial presentar un sólido marco teórico que sustente la investigación y desarrollo de técnicas avanzadas en procesamiento del lenguaje natural (PLN). Este capítulo se centrará primero en detallar el marco teórico de la función ICDS, proporcionando una comprensión profunda de sus fundamentos y cómo se aplica en la semántica distribucional composicional. A continuación, se abordará la semántica en general, estableciendo las bases necesarias para explorar la semántica composicional desde una perspectiva lingüística. Esto incluye un análisis de cómo se construye el significado en el lenguaje natural y cómo se puede modelar computacionalmente para tareas de PLN.

Seguidamente, se discutirá la distribucionalidad y la síntesis kantiana, vinculándolas con lo expuesto en el marco teórico de Amigó et al. (2022) con la lingüística, con la finalidad de entender su relevancia en la representación semántica. Estos conceptos permitirán contextualizar la importancia de las relaciones entre palabras y la estructura de la información en la creación de modelos semánticos robustos. El capítulo también incluirá una sección sobre los fundamentos lógico-lingüísticos de las oraciones. Aquí se establecerá la importancia de la jerarquía en las oraciones, argumentando cómo la estructura sintáctica influye en la composición semántica y la representación del significado.

Finalmente, se llevará a cabo una revisión de la estructura y tipología sintáctica, desde un nivel macro de oraciones hasta un nivel micro de componentes lingüísticos, donde se engloban sintagmas que desempeñan distintas tareas. En esta revisión, se expondrán los elementos y la jerarquía que estos tienen dentro de las oraciones. Esto proporcionará la base principal para poder plantear, en el siguiente capítulo, toda la propuesta de orden. La jerarquía de los componentes lingüísticos será esencial para entender cómo se debe estructurar la composición semántica para mejorar la precisión y coherencia de las representaciones textuales.

En resumen, este capítulo establecerá un marco teórico robusto, abordando la semántica composicional desde una perspectiva lingüística y preparando el terreno para la propuesta de orden que se desarrollará en el capítulo siguiente.

2.1. Marco de la función ICDS

Como ya se ha presentado anteriormente, la representación del significado del texto es un desafío fundamental en el procesamiento del lenguaje natural (PLN) que implica codificar el lenguaje de manera que pueda ser manejada eficientemente por sistemas de gestión de información. El marco de la función ICDS aborda este problema combinando principios de la teoría de la información, la distribucionalidad, la síntesis kantiana y la semántica composicional, proponiendo una solución que mejora el acceso a la información textual, la minería de texto y los sistemas de diálogo, entre otros (Maruyama, 2019).

Para alcanzar estos objetivos, la función ICDS se basa en dos principios clave: la composicionalidad y la contextualidad. Por un lado, el principio de composicionalidad establece que el significado de una expresión es una función del significado de sus partes y de la manera en que se combinan sintácticamente. Este principio es la base del paradigma simbólico, que asocia el lenguaje con la lógica proposicional a través de gramáticas que capturan las estructuras del lenguaje. Por ejemplo, en este paradigma, las palabras «mesa» y «sentado» se combinan sintácticamente para formar la expresión «sentado en una mesa», cuyo significado se deriva de sus partes constituyentes y su estructura (Amigó et al., 2022). Por otro lado, el principio de contextualidad sostiene que el significado de las palabras y expresiones está determinado por su contexto de uso. En el paradigma distribucional, el significado se infiere del contexto en que aparecen las palabras. Las palabras y expresiones se representan como puntos en un espacio vectorial continuo. Por ejemplo, las palabras «mesa» y «sentado», y la expresión «sentado en una mesa» se proyectan en este espacio según el contexto textual en que suelen aparecer, permitiendo una representación continua y gradual del significado (Amigó et al., 2022).

Asimismo, a lo largo de la historia del procesamiento del lenguaje natural, también se han planteado diversos paradigmas para la representación del significado del texto. Estos paradigmas se basan en diferentes teorías y enfoques que buscan equilibrar varias propiedades clave de la representación semántica. Entre estas propiedades se encuentran la sistematicidad, el contexto de uso, la continuidad y la mensurabilidad de la información:

- **Sistematicidad:** La capacidad de procesar variantes sistemáticas de una oración.

- **Contexto de Uso:** La sensibilidad de la representación al contexto en que aparece una expresión.
- **Continuidad:** Representación en un espacio continuo multidimensional que refleja similitudes semánticas.
- **Mensurabilidad de información:** Una función que mide la cantidad de información contenida en el enunciado.

Como presentan Amigó et al. (2022), los primeros modelos semánticos se basaban en enfoques logicistas, por lo que cumplía con la sistematicidad, pero no otras propiedades. Los modelos de espacio vectorial (VSM) introdujeron la continuidad, pero ignoraron la sistematicidad y el contexto de uso. Los modelos de lenguaje basados en conteo y los modelos de lenguaje neurales modernos lograron capturar la mensurabilidad de la información y el contexto de uso, pero la sistematicidad seguía siendo aún un rato. Por tanto, los enfoques distribucionales composicionales son los únicos que cumplen con todas estas propiedades.

Por otro lado, los embeddings de palabras demostraron que los datos semánticos mejoran el rendimiento en tareas de PLN. Los embeddings de palabras estáticos, como SGNS y GloVe, optimizan la correspondencia entre el producto escalar de embeddings y su similitud distribucional. Los modelos secuenciales y basados en grafos, como LSTM y transformer, mejoraron la representación textual, pero aún se enfrentaban a problemas de degradación de la representación, especialmente en capas superiores. De hecho, modelos como Sentence-BERT y Universal Sentence Encoder abordan algunos de estos problemas, pero su efectividad varía según las características intrínsecas del texto.

Sin embargo, tras esto surgen modelos de semántica distribucional composicional, que utilizan funciones de composición para combinar representaciones distribucionales que han demostrado ser efectivos. Aun así, estos se enfrentan a limitaciones como la falta de consideración del orden de las palabras y la escalabilidad. Por tanto, algunos autores han propuesto añadir una capa simbólica sobre la representación distribucional, mientras que otros han utilizado nociones de teoría de la información para mejorar la precisión (Amigó et al., 2022). Sin embargo, estos enfoques aún no consideran completamente la estructura sintáctica de las oraciones.

Finalmente, como una nueva propuesta que mejore el resultado de estos modelos, surge el marco teórico propuesto por ICDS. Este se basa en una interpretación geométrica de la semántica distribucional y su conexión con la teoría de la información. Esto se debe a que existe una correspondencia entre la norma del vector y la especificidad o IC del enunciado representado. Según Levy y Goldberg (2014) y Arora et al. (2016), el producto escalar de los embeddings SGNS aproxima la información mutua puntual (PMI) entre dos palabras. Esto implica que la norma del vector corresponde al IC de los enunciados según la teoría de la información de Shannon (Amigó et al., 2022). La ausencia de información textual se representa como el origen de coordenadas en el espacio vectorial, formando una esfera alrededor de este punto.

Por tanto, el marco ICDS se formaliza como una tupla de tres funciones: **embeddings** (π), **composición** (\oplus) y **similitud** (δ), donde S es el espacio de unidades lingüísticas básicas (Amigó et al., 2022). Estas funciones deben ser consistentes con el contenido de información de los embeddings, capturando el contexto de uso, la sistematicidad y la continuidad semántica. La función de embeddings captura el contexto de uso y la mensurabilidad de la información, la función de composición aborda la sistematicidad y mantiene la coherencia de la mensurabilidad de la información en representaciones compuestas; y la función de similitud define el espacio continuo de representaciones semánticas. Este marco captura la dualidad entre la semántica composicional y distribucional, determinando la semántica de las unidades lingüísticas básicas y las estructuras complejas a través de la composición de palabras.

$$\begin{aligned}\pi &: S \longrightarrow \mathbb{R}^n, \\ \delta &: \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}, \\ \odot &: \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n\end{aligned}$$

A continuación, se establecerán las propiedades de estas funciones para comprender la relevancia que tienen dentro de este marco teórico:

Propiedades de la función de embeddings (π)

1. **Cuantificabilidad de la información:** La norma del vector de embeddings aproxima el contenido de información (IC) de una unidad lingüística x :

$$\|\pi(x)\| \simeq IC(x) = -\log(P(x))$$

Esta propiedad asegura que la magnitud del vector sea proporcional a la rareza o especificidad del término, es decir, términos menos probables o específicos tendrán vectores con una norma mayor.

2. **Isometría angular:** Debe existir isometría entre la posición angular de los embeddings y la similitud esperada de los enunciados según los humanos:

$$\cos(\pi(x), \pi(y)) \propto \mathbb{E}(\text{SIM}(x, y))$$

Esta propiedad garantiza que la similitud de los vectores en el espacio de embeddings refleje la similitud semántica percibida entre los términos, permitiendo que los vectores cercanos en el espacio de representación semántica representen términos semánticamente similares. Esta isometría es la que se ha mostrado que se mantiene en mucha mayor medida en el caso de los word embeddings estáticos, y en los contextuales no se da, aunque se han probado en el marco ICDS.

Propiedades de la función de composición

3. **Elemento neutro de la composición:** Los componentes de información nula (norma cero) no afectan la composición:

$$\|\vec{v}_2\| = 0 \implies \|\vec{v}_1 \odot \vec{v}_2\| = \|\vec{v}_1\|$$

Esta propiedad asegura que añadir un vector de información nula (vector de norma cero) no altere la información ya presente en el vector compuesto, de forma en la que se mantiene así la integridad de la información original.

4. **Cota inferior de la norma de la composición:** La norma del vector compuesto es mayor o igual a la norma de cada componente; es decir, la composición nunca reduce el contenido de información:

$$\|\vec{v}_1 \odot \vec{v}_2\| \geq \|\vec{v}_1\| \quad \|\vec{v}_1 \odot \vec{v}_2\| \geq \|\vec{v}_2\|$$

Esta propiedad asegura que la composición de dos vectores no disminuya la cantidad total de información representada, lo que refleja que la combinación de conceptos no debe resultar en una pérdida de información.

5. **Monotonicidad de la norma de la composición:** La norma del vector compuesto es monótonica con respecto al ángulo entre los vectores componentes:

$$\left. \begin{array}{l} \|\vec{v}_1\| = \|\vec{v}_2\| = \|\vec{v}_3\| \\ \cos(\vec{v}_1, \vec{v}_2) > \cos(\vec{v}_1, \vec{v}_3) \end{array} \right\} \implies \|\vec{v}_1 \odot \vec{v}_2\| < \|\vec{v}_1 \odot \vec{v}_3\|$$

Esta propiedad asegura que la norma del vector compuesto varíe de manera predecible según la similitud angular de los vectores componentes, reflejando que vectores más similares deberían combinarse de manera más eficiente.

6. **Sensibilidad a la estructura:** La composición debe ser sensible a la forma en que las palabras se estructuran lingüísticamente:

$$\left. \begin{array}{l} \|\vec{v}_1\| = \|\vec{v}_2\| = \|\vec{v}_3\| > 0 \\ \cos(\vec{v}_1, \vec{v}_2) = \cos(\vec{v}_1, \vec{v}_3) = \cos(\vec{v}_2, \vec{v}_3) > 0 \end{array} \right\} \implies (\vec{v}_1 \odot \vec{v}_2) \odot \vec{v}_3 \neq \vec{v}_1 \odot (\vec{v}_2 \odot \vec{v}_3)$$

Esta propiedad refleja que la estructura jerárquica y el orden de las palabras en una oración son importantes en la composición semántica, lo cual es crucial para capturar el significado correcto en lenguajes naturales. Como se verá en otros capítulos, la lingüística establece que el orden importa; por ello, estas propiedades son deseables para cualquier sistema de representación semántica. En consecuencia, un sistema de representación semántica debe cumplir con esta condición para ser eficaz.

Propiedades de la función de similitud

7. **Monotonicidad de la similitud con la distancia angular:** Bajo igual norma vectorial (igual IC), la similitud es una función monótonica decreciente con respecto a la distancia angular:

$$\left. \begin{array}{l} \cos(\vec{v}_1, \vec{v}_2) > \cos(\vec{v}_1, \vec{v}_3) \\ \|\vec{v}_1\| = \|\vec{v}_2\| = \|\vec{v}_3\| > 0 \end{array} \right\} \implies \delta(\vec{v}_1, \vec{v}_2) > \delta(\vec{v}_1, \vec{v}_3)$$

Esta propiedad asegura que, a igualdad de magnitud de los vectores, aquellos que están más cerca en términos angulares sean considerados más similares, lo que alinea con la intuición humana de similitud.

8. **Monotonicidad de la similitud con embedding ortogonal:** Dados vectores ortogonales, cuanto mayor sea su norma (su especificidad), menos similares serán:

$$\left. \begin{array}{l} \cos(\vec{v}_1, \vec{v}_2) = \cos(\vec{v}_3, \vec{v}_4) = 0 \\ \|\vec{v}_1\| < \|\vec{v}_2\|, \|\vec{v}_3\| < \|\vec{v}_4\| \end{array} \right\} \implies \delta(\vec{v}_1, \vec{v}_2) > \delta(\vec{v}_3, \vec{v}_4)$$

Una vez definidas estas funciones, se plantea la función de composición en la que se enmarca ICDS: la función de composición generalizada $F_{\lambda, \mu}$. Como ya se ha planteado, es fundamental en la representación semántica distribucional composicional basada en la teoría de la información (ICDS), y sirve como marco para la realización de este trabajo de investigación. Esta función se define matemáticamente como:

$$F_{\lambda, \mu}(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|} \cdot \sqrt{\lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle}$$

Como se puede apreciar, los componentes de la función son el vector unitario de la suma y la norma del vector.

El vector unitario de la suma determina la dirección del vector compuesto. Al normalizar la suma de los vectores, se obtiene un vector unitario que apunta en la misma dirección que la suma de los vectores originales:

$$\frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|}$$

Por otro lado, la forma del vector determina la magnitud del vector compuesto. Depende de las normas de los vectores individuales y su producto interno. Los parámetros λ y μ permiten ajustar cómo se combinan estas magnitudes y la interacción entre los vectores:

$$\sqrt{\lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle}$$

Asimismo, la función $F_{\lambda, \mu}$ puede especializarse en varias funciones de composición conocidas al elegir valores específicos para λ y μ . Por otro lado, cuando los parámetros λ y μ toman valores en un determinado rango, cumplen todas las propiedades

deseables impuestas dentro del marco ICDS. Fuera de esos valores no se cumplen, y con valores concretos la función toma la forma de otras funciones de composición conocidas, y que NO cumplen con las condiciones deseables.

Estos se abordarán en la metodología en profundidad, con la finalidad de especificar qué tipos de funciones de composición concretas se utilizarán para llevar a cabo la experimentación.

2.2 Semántica

Tras haber abordado el marco ICDS en el que se fundamenta todo este trabajo, se presenta un análisis del marco en el área de la semántica, para poder partir de un enfoque lógico-lingüístico. Para esto se recurre a la autora Vidal, M. V. E. (2004) en su obra *Fundamentos de semántica composicional*. En ella, se presenta la semántica como una rama de la lingüística que se ocupa del estudio del significado lingüístico. Sin embargo, este no es el único enfoque que dispone esta rama, ya que se suele asociar el significado lingüístico con el significado de las unidades menores. La semántica también aborda la combinación de estas palabras para formar significados. Por tanto, se plantean dos enfoques principales dentro de la semántica: la semántica léxica y la semántica composicional.

Por un lado, la semántica léxica se ocupa del significado de las palabras individuales, examinando cómo se definen y cómo interactúan entre sí. Incluye el estudio de las relaciones semánticas, como sinónimos, antónimos, hiperónimos (palabras más generales) e hipónimos (palabras más específicas). Este enfoque analiza cómo cada palabra contribuye al significado general dentro de su contexto.

Vidal (2004) destaca que la semántica léxica se enfoca en las unidades simples del lenguaje, es decir, las palabras individuales. Según Vidal (2004), «el estudio de las relaciones entre palabras y sus significados constituye una parte fundamental de la semántica léxica, proporcionando una base para comprender cómo las palabras interactúan en diferentes contextos» (2004:45).

Otros autores también han abordado la importancia de la semántica léxica. Lyons (1977) define la semántica léxica como «la rama de la semántica que se ocupa de las relaciones semánticas entre las palabras y su significado en el contexto» (1977:99). Cruse

(1986) añade que «las relaciones léxicas como la sinonimia, antonimia y hiponimia son cruciales para entender cómo se estructuran los significados en una lengua» (1986:63).

Por otro lado, la semántica composicional se enfoca en cómo se combina y compone el significado de las palabras para formar el significado de frases y oraciones. Es posible diferenciar a la semántica léxica como el estudio de las unidades simples, mientras que la semántica composicional se encarga de las expresiones complejas. En el contexto de la semántica composicional, una expresión compleja se refiere a una construcción lingüística que consta de varias partes, como palabras, sintagmas u oraciones, y su significado se deriva de la manera en que estas partes se combinan. Además, examina cómo la estructura sintáctica y la combinación de palabras contribuyen al significado global de una expresión lingüística. Vidal (2004) argumenta que «la combinación de unidades simples debe cumplir con las reglas y principios de la gramática para generar significados coherentes y aceptables dentro del lenguaje» (2004:67).

Otras de las ideas que presenta Vidal (2004) es el principio de gramaticalidad. La combinación de estas unidades simples debe cumplir con las reglas y principios de la gramática. Se entiende el principio de gramaticalidad partiendo desde las reglas de la gramática, refiriéndonos a las regularidades y generalizaciones sobre el funcionamiento del sistema que se está usando. Por tanto, es importante destacar la idea de que cualquier contenido agramatical queda fuera de las posibilidades que ofrece la lengua. Esta última idea es interesante, ya que hay secuencias agramaticales que, aun así, se pueden interpretar. Esto se debe al hecho de que se reestructura y reconstruye inconscientemente la agramaticalidad, de forma en la que se crea una versión correcta. A partir de ahí, se deduce que no tendría sentido plantear la agramaticalidad relacionada con la semántica.

Noam Chomsky, en su teoría generativa, también aborda el principio de gramaticalidad. En ella argumenta que la gramática de una lengua debe poder diferenciar entre las secuencias gramaticales y las agramaticales, proporcionando una descripción clara de las estructuras permitidas (Chomsky, 1957:13). Esta distinción es esencial para entender cómo se forman y se interpretan las oraciones en cualquier lengua. Asimismo, Jackendoff (1997) en su obra *The Architecture of the Language Faculty* también discute el principio de gramaticalidad, destacando que «la gramaticalidad es un criterio fundamental en la teoría lingüística para determinar qué secuencias de palabras constituyen oraciones bien formadas en una lengua» (1997:23). Según Jackendoff, la

gramaticalidad no solo afecta la estructura sintáctica sino también la interpretación semántica de las oraciones.

Vidal (2004) también destaca una primera distinción básica entre los conceptos de significado y la interpretación. El significado proviene de forma exclusiva de las unidades léxicas y relaciones sintácticas que se dan entre ambas, por lo que es sistemático, constante e independiente del contexto y situación. La interpretación, sin embargo, sí incluye la contribución de factores situacionales de la naturaleza ajena a la lingüística dentro del propio significado lingüístico mencionado previamente.

El concepto de significado se refiere al contenido semántico inherente de las palabras y las frases. Vidal (2004) argumenta que el significado es una propiedad intrínseca de las unidades léxicas (las palabras) y de las relaciones sintácticas que se establecen entre ellas en una oración. Esto significa que el significado es:

- **Sistemático:** Está regido por reglas y principios que determinan cómo las palabras pueden combinarse y cómo estas combinaciones generan nuevos significados.
- **Constante:** No varía dependiendo del contexto o la situación en la que se utiliza la palabra o frase. Por ejemplo, la palabra «gato» siempre se refiere a un tipo específico de animal, independientemente del contexto en el que se mencione.
- **Independiente del contexto:** El significado de una expresión se mantiene constante, independientemente de factores externos al propio lenguaje, como la situación o las intenciones del hablante.

Vidal (2004) ejemplifica esta constancia al explicar que la palabra «gato» siempre mantiene su referencia a un felino doméstico, sin importar si se menciona en un cuento infantil, en una conversación cotidiana, o en un tratado de biología (2004:89). Esta propiedad de constancia es fundamental para el análisis semántico, ya que permite una comprensión estable y predecible de las expresiones lingüísticas.

Por otro lado, la interpretación de una expresión lingüística sí incluye la contribución de factores situacionales y contextuales que son ajenos a la estructura lingüística per se. Vidal (2004) sostiene que la interpretación se refiere a cómo se entiende y se utiliza el significado en situaciones comunicativas reales. La interpretación involucra:

- **Contexto situacional:** Los factores externos al lenguaje, como el entorno físico, la situación social y las intenciones del hablante, juegan un papel crucial en la interpretación. Por ejemplo, la oración «el gato está en la mesa» puede interpretarse de manera diferente si se dice durante una cena (donde puede causar sorpresa) o en una tienda de mascotas (donde está normalizado).
- **Intenciones del hablante:** Las intenciones y el conocimiento previo del hablante y del oyente influyen en la interpretación de las expresiones. Vidal (2004) explica que «las intenciones comunicativas del hablante pueden modificar significativamente cómo se interpreta una expresión, incluso si su significado léxico permanece constante» (2004:92).
- **Variabilidad:** A diferencia del significado, la interpretación puede variar considerablemente dependiendo de los factores contextuales. Esto hace que la interpretación sea flexible y adaptable a diferentes situaciones comunicativas.

Por tanto, estas ideas ilustran que mientras el significado de la oración presentada anteriormente es claro y constante, su interpretación puede variar si se considera el contexto de la cena o la tienda de mascotas.

Por otro lado, las contribuciones de Richard Montague a la semántica sirvieron como marco lógico y riguroso. Montague sostuvo que si las partes significativas de las afirmaciones no fueran siempre intercambiables, el lenguaje no podría ser formalizado (Montague, 1970). Su enfoque en la semántica formal demostró cómo el significado de una expresión compleja se puede derivar sistemáticamente de los significados de sus componentes más simples, estableciendo así las bases para la semántica composicional moderna.

Montague fue fundamental para la aplicación de principios lógico-matemáticos en el estudio del lenguaje natural, proporcionando un marco teórico que ha influido significativamente en la lingüística teórica y la filosofía del lenguaje contemporáneas. Su trabajo demostró que «la formalización del lenguaje permite una precisión y una claridad en el análisis semántico que no sería posible de otra manera» (Montague, 1970:95).

Además de Vidal y Montague, otros lingüistas han realizado contribuciones significativas al campo de la semántica composicional y léxica. Por ejemplo, el ya presentado Noam Chomsky (1957) en su obra *Syntactic Structures* introduce la teoría

generativa, que establece que la estructura sintáctica de una oración es crucial para su interpretación semántica. Chomsky argumenta que «la estructura profunda de una oración es lo que determina su significado básico, mientras que la estructura superficial refleja su forma fonética» (Chomsky, 1957:48). Asimismo, Barbara Partee (1984) también ha sido influyente en la integración de la semántica formal con la teoría lingüística. Partee sostiene que «la semántica formal proporciona un marco útil para analizar cómo los significados de las partes de una oración se combinan para formar el significado de la oración completa» (Partee, 1984:113).

Estos desarrollos teóricos no solo han enriquecido la comprensión del significado lingüístico, sino que también han proporcionado herramientas conceptuales esenciales para el análisis formal del lenguaje y su estructura. Por tanto, la integración de principios de la semántica composicional y la semántica léxica ha permitido avanzar en la creación de modelos de procesamiento del lenguaje natural que son capaces de manejar de manera efectiva tanto las palabras individuales como las estructuras más complejas.

2.3 Composicionalidad

En el subapartado anterior, en relación a las ideas de Vidal (2004) se ha planteado el concepto de semántica composicional. Mediante esta base, se parte del concepto la composicionalidad, que explica que una expresión completa se deriva de manera sistemática y predecible del significado de sus partes constituyentes y de la manera en que estas partes están combinadas sintácticamente. El contexto de la representación simbólica se refleja en una conexión entre el lenguaje y la lógica proposicional mediante referencias semánticas extendidas. Esto quiere decir que se realiza asignando significados específicos a las palabras y construcciones lingüísticas. Además, se utiliza un marco gramatical que captura las estructuras del lenguaje, permitiendo analizar cómo las partes individuales de una expresión contribuyen a su significado global.

Sin embargo, el elemento la semántica composicional aborda la combinación de distintos elementos lingüísticos para comprender su significado, esta idea no es suficiente. Si se tratara de una simple adición del significado de cada palabra, cualquier oración que contuviese las mismas palabras y fueran gramaticales, significarían lo mismo. Sin embargo, que se utilicen los mismos elementos no implica un mismo significado.

La combinación de unos mismos elementos de distinta forma, implica que el significado puede variar. De hecho, a raíz de esta idea es posible deducir que la propia estructura sintáctica contribuye decisivamente a la interpretación de una forma estable y sistemática. Esto significa que la misma colección de palabras, dispuestas de manera diferente, puede tener significados distintos.

El principal reto de la semántica composicional radica en encontrar un modo adecuado de abordar su propio objeto de estudio: el significado de un conjunto potencialmente infinito de expresiones complejas. A diferencia de la semántica léxica, que se ocupa del significado de palabras individuales y enfrenta un conjunto finito (aunque amplio) de unidades léxicas, la semántica composicional se enfrenta a la complejidad de manejar combinaciones ilimitadas de estas unidades. Esto se relaciona con el principio de composicionalidad, que establece que el significado de un todo es una función del significado de sus partes y de la forma sintáctica en que estas se combinan. Este principio es la piedra angular del paradigma de representación simbólica, que intenta relacionar el lenguaje con la lógica proposicional a través de referencias semánticas extendidas (el significado referencial de las palabras) y gramáticas que capturan las estructuras del lenguaje.

En otras palabras, este principio sugiere que, para comprender el significado de una oración o frase, es necesario entender no solo el significado de cada palabra individualmente, sino también cómo estas palabras se organizan y relacionan entre sí. La composición y la estructura son cruciales porque el orden y la relación sintáctica entre las palabras pueden cambiar completamente el significado de una frase. Por ejemplo, aunque las palabras «gato», «persigue» y «ratón» tienen significados individuales claros, el significado de las frases «el gato persigue al ratón» y «el ratón persigue al gato» difiere drásticamente debido al orden y la estructura sintáctica de las palabras.

En el paradigma de representación simbólica, se intenta crear un puente entre el lenguaje natural y la lógica proposicional. Este enfoque implica asignar significados específicos a las palabras y expresiones, y usar reglas gramaticales y sintácticas para modelar cómo se construye el significado en oraciones más complejas. Las referencias semánticas extendidas se refieren a la idea de que cada palabra no solo tiene un significado aislado, sino que también lleva consigo referencias o conexiones a otros conceptos y contextos.

Por tanto, la dificultad principal radica en que la caracterización de las unidades no se puede lograr por extensión. No es práctico hacer una lista exhaustiva de todas las expresiones complejas en un idioma y proporcionar una caracterización individual para cada una de ellas. Esto contrasta con la semántica léxica, donde la delimitación del conjunto de palabras es más manejable.

La semántica composicional se enfrenta a la tarea de desarrollar principios y marcos teóricos que permitan entender cómo el significado de una expresión compleja surge de la combinación de sus partes constituyentes y su estructura sintáctica. Esto implica considerar no solo el significado de las palabras individuales, sino también cómo interactúan y contribuyen a la interpretación global. La infinitud de combinaciones posibles agrega complejidad a este proceso y destaca la necesidad de enfoques abstractos y generales que puedan abordar de manera efectiva la diversidad de expresiones que pueden surgir en el uso del lenguaje. Aun así, gracias a la productividad gramatical y la hipótesis de la composicionalidad, se logra abordar tanto el significado de expresiones complejas como resolver el problema de su infinitud en el lenguaje.

La productividad gramatical se refiere a la capacidad del lenguaje para generar un número ilimitado de expresiones nuevas mediante reglas gramaticales. Esto implica que, aunque el conjunto de palabras y unidades léxicas puede ser finito, la manera en que estas unidades se combinan sintácticamente es virtualmente ilimitada. La hipótesis de la composicionalidad es fundamental en la lingüística y la semántica, y se basa en tres propiedades principales que intentan explicar cómo comprendemos el significado de las expresiones complejas en el lenguaje:

1. **Sistematicidad:** Nuestra comprensión de las oraciones es sistemática, lo que significa que hay patrones definidos y predecibles en las estructuras de las oraciones que entendemos. Por ejemplo, en las oraciones «Carmen habla con Ana» y «Ana habla con Carmen», la estructura es similar, y podemos generalizar patrones sintácticos y semánticos.
2. **Productividad:** Tenemos la capacidad de comprender oraciones nuevas. A pesar de que no hemos oído o leído una oración específica antes, podemos entenderla si sigue patrones sintácticos y semánticos comunes. Esta propiedad refleja la capacidad del lenguaje para generar y comprender nuevas expresiones de manera creativa.

3. **Infinitud:** Tenemos la capacidad de comprender todas y cada una de las oraciones de una serie indefinidamente amplia. Esto implica que, en teoría, no hay límite para la complejidad o la longitud de las oraciones que podemos comprender.

Sin embargo, a pesar de estas propiedades, existen limitaciones en la aplicación de la composicionalidad. Algunas expresiones complejas parecen escapar a este principio. Por ejemplo, en expresiones como «mesa redonda», el significado de la expresión completa no puede entenderse simplemente descomponiendo el significado de las palabras individuales («mesa» y «redonda»). Aquí, la composicionalidad parece no aplicarse de manera directa, ya que el significado global no es simplemente la suma de los significados de las partes.

Además, las locuciones fijas, frases hechas y modismos también presentan desafíos para la composicionalidad. Estas expresiones idiomáticas tienen significados particulares que no pueden deducirse de manera directa de las partes constituyentes. En estos casos, el significado es más que la suma de las partes, y la interpretación depende de la convención lingüística y cultural. A pesar de las limitaciones mencionadas, es crucial destacar que la aplicación de la hipótesis de la composicionalidad no se invalida. De hecho, el principio de composicionalidad nos proporciona un marco predictivo sólido. Cuando nos enfrentamos a expresiones complejas que no conocemos previamente, la aplicación de la composicionalidad nos permite prever acertadamente que la única interpretación plausible será la forma composicional, es decir, derivada de la combinación sistemática de las partes constituyentes.

Este enfoque sirve como un punto de partida efectivo para la interpretación de nuevas expresiones, incluso aquellas que podrían parecer desafiantes para la composicionalidad. Sin embargo, es importante reconocer que, en algunos casos, otras consideraciones, como convenciones lingüísticas específicas, contextos culturales o convenciones idiomáticas, pueden imponer interpretaciones adicionales o significados específicos que no se derivan exclusivamente de la composición de partes individuales.

Además de las ideas planteadas por Vidal, varios otros lingüistas y filósofos han contribuido significativamente al desarrollo y comprensión de la semántica composicional. La ya mencionada Barbara H. Partee ha sido una figura central en la integración de la semántica formal y la teoría lingüística. Su trabajo ha sido fundamental en el desarrollo de la semántica composicional. En su trabajo, Partee ha mostrado cómo

las herramientas de la lógica formal pueden aplicarse para entender fenómenos semánticos complejos. Partee ha trabajado extensamente en la teoría de los tipos, un marco formal que categoriza las expresiones lingüísticas según su tipo semántico. Los tipos semánticos permiten entender cómo las diferentes partes de una oración interactúan para producir significados coherentes. Por ejemplo, los nombres propios se clasifican como individuos (e), mientras que las oraciones completas se clasifican como proposiciones (t). Las funciones, como los verbos y adjetivos, se consideran funciones entre estos tipos. Este enfoque sistemático facilita el análisis de la composición del significado, permitiendo una representación clara y estructurada de cómo las palabras se combinan para formar significados más complejos.

Uno de los aspectos más influyentes del trabajo de Partee ha sido su análisis de los pronombres y las anáforas. En su artículo *Binding Implicit Variables in Quantified Contexts* (1978), Partee explora cómo los pronombres y otros elementos anafóricos se relacionan con sus antecedentes en el discurso. Utilizando herramientas de la semántica formal, Partee proporciona un marco para entender cómo se establecen y mantienen los referentes en un texto, lo cual es crucial para la coherencia y cohesión del discurso. Este trabajo ha sido esencial para el desarrollo de teorías semánticas que explican cómo los hablantes y oyentes interpretan las referencias dentro de un contexto discursivo.

Por otro lado, se encuentra David Dowty, quien ha realizado contribuciones significativas a la semántica composicional, especialmente en su obra *Word Meaning and Montague Grammar* (1979). Dowty explora cómo las teorías de Montague pueden aplicarse tanto a la semántica léxica como a la composicional. Sostiene que «la combinación de palabras para formar significados más complejos requiere una comprensión detallada de las categorías semánticas y las reglas sintácticas que gobiernan estas combinaciones» (Dowty, 1979:145). Dowty desarrolló la teoría de los roles temáticos, que describe cómo los diferentes elementos de una oración contribuyen al significado general de la misma. Su trabajo ha sido crucial para entender cómo los verbos y sus argumentos interactúan semánticamente, lo que también sirve como base para llevar a cabo planteamientos semántico-computacionales.

Asimismo, Gottlob Frege es considerado uno de los padres fundadores de la semántica composicional. En su influyente ensayo *Über Sinn und Bedeutung* (1892), Frege introduce el principio de composicionalidad, argumentando que el significado de una oración completa es una función del significado de sus partes constituyentes (Frege,

1892). Este principio se ha convertido en una piedra angular de la semántica composicional moderna. Frege también distingue entre el sentido (*Sinn*) y la referencia (*Bedeutung*) de una expresión, una distinción que ha sido fundamental para muchos desarrollos posteriores en la semántica. El sentido de una expresión se refiere a su contenido cognitivo o el modo en que presenta una referencia, mientras que la referencia es el objeto real al que se refiere la expresión. Esta idea encaja con las ya presentadas anteriormente por Vidal (2004), pero sirve para remarcar la posibilidad de establecer una percepción semántica.

En relación a la relevancia de la estructura dentro del ámbito semántico composicional, Noam Chomsky ha tenido un impacto significativo en el campo, especialmente en la relación entre la sintaxis y la semántica. De hecho, la estructura profunda, según Chomsky, contiene toda la información necesaria para determinar el significado semántico de una oración. Esta estructura abstracta se transforma mediante reglas sintácticas en la estructura superficial, que es la forma en la que se pronuncia la oración. Por ejemplo, en las oraciones pasivas y activas, aunque la estructura superficial difiere, la estructura profunda subyacente que contiene el significado esencial puede ser similar. Esta transformación es clave para entender cómo diferentes construcciones sintácticas pueden expresar el mismo contenido semántico.

El trabajo de Chomsky ha influido profundamente en la semántica composicional, proporcionando un marco para analizar cómo la estructura sintáctica de las oraciones contribuye a su significado. La idea de que la estructura profunda determina el significado semántico ha llevado a los lingüistas a explorar cómo las diferentes configuraciones sintácticas afectan la interpretación semántica de las expresiones. Esta perspectiva es fundamental para la teoría de la composicionalidad, que sostiene que el significado de una expresión compleja se deriva del significado de sus partes constituyentes y de la manera en que estas partes se combinan.

La distinción entre estructura profunda y estructura superficial también ha influido en otros enfoques semánticos, como la semántica formal y los modelos de representación distribucional. Por ejemplo, en la semántica formal, la estructura profunda puede ser vista como una fórmula lógica que representa el contenido proposicional de una oración, mientras que la estructura superficial podría relacionarse con cómo se expresa esta fórmula en un lenguaje natural. En los modelos de representación distribucional, la estructura profunda puede estar relacionada con las representaciones vectoriales de

significado, mientras que la estructura superficial podría estar vinculada a la forma en que estas representaciones se combinan y manipulan para formar significados complejos.

Por otro lado, Emmon Bach ha sido otro contribuyente importante en el campo de la semántica composicional. En su obra, Bach ha explorado la relación entre la sintaxis y la semántica, especialmente en el contexto de las categorías verbales y los eventos. Bach (1986) argumenta que "los eventos y las entidades tienen una estructura interna que debe ser reflejada en la semántica de las oraciones que los describen" (p. 28). Su enfoque ha sido fundamental para entender cómo las oraciones que describen acciones y eventos pueden ser analizadas de manera composicional. Bach ha desarrollado la teoría de la semántica de eventos, que proporciona un marco para analizar cómo las descripciones de acciones y eventos se integran en un todo coherente a partir de sus partes constituyentes.

Si bien en este trabajo se pone el foco en las oraciones simples, es importante destacar que Bach desarrolló la teoría de la semántica de eventos, que proporciona un marco para analizar cómo las descripciones de acciones y eventos se integran en un todo coherente a partir de sus partes constituyentes. Según esta teoría, los verbos no solo denotan acciones, sino que también implican estructuras eventivas que pueden ser descompuestas en componentes más pequeños, como los participantes del evento (agentes, pacientes, etc.), la temporalidad y la modalidad. Esta descomposición permite una comprensión más detallada y matizada de cómo se construyen los significados de las oraciones.

Uno de los conceptos clave en la semántica de eventos de Bach es la noción de «estructuras eventivas», que representan la organización interna de los eventos descritos por los verbos. Por ejemplo, en la oración «Juan rompió el vaso», la estructura eventiva incluye a Juan como agente, el vaso como paciente, y el acto de romper como la acción principal. Esta estructura permite desglosar el significado de la oración en sus componentes constituyentes, facilitando un análisis composicional.

Bach también exploró cómo los eventos pueden ser anidados y cómo las oraciones complejas pueden describir secuencias de eventos interrelacionados. Esto es particularmente relevante para el análisis de oraciones subordinadas y coordinadas, donde múltiples eventos se describen en una sola estructura sintáctica. Su enfoque ha sido crucial para el desarrollo de modelos semánticos que capturan la complejidad de las descripciones verbales en el lenguaje natural.

Para finalizar, y como broche en esta presentación de la semántica composicional, es esencial destacar a Richard Montague, a quien se le asocian numerosos avances dentro del marco lingüístico. Como se expondrá también más adelante, uno de los aportes más importantes de Montague fue la integración de la lógica de primer orden y la teoría de conjuntos en el análisis semántico. Utilizando estas herramientas, Montague pudo representar formalmente las relaciones entre los términos y las proposiciones del lenguaje natural. Por ejemplo, una oración como «Todos los estudiantes aprobaron el examen» puede ser representada formalmente utilizando cuantificadores y predicados, permitiendo un análisis detallado de su estructura semántica.

Montague también desarrolló la teoría de los mundos posibles, que amplía la semántica formal para incluir la modalización y otras relaciones semánticas complejas. En este marco, el significado de una expresión no solo se determina por su verdad en el mundo real, sino también por su verdad en un conjunto de mundos posibles. Esto permite capturar la naturaleza contingente y modal del lenguaje, proporcionando una herramienta poderosa para el análisis de oraciones que expresan posibilidades, necesidades y otras modalidades. Además, Montague estableció el principio de composicionalidad en su forma más estricta, argumentando que el significado de una expresión compleja debe derivarse de los significados de sus partes y de las reglas que las combinan. Este principio ha sido fundamental para el desarrollo de teorías semánticas que buscan explicar cómo se construyen los significados de las oraciones a partir de sus componentes léxicos y sintácticos.

2.4 Distribucionalidad y síntesis kantiana

En el artículo *Information Theory–based Compositional Distributional Semantics*, Amigó et al. (2022) exponen ideas sobre el equilibrio entre la composicionalidad y la distribucionalidad. Aunque la composicionalidad establece que el significado de una expresión se deriva de la combinación sintáctica y semántica de sus partes, es imperativo reconocer la importancia de la contextualidad. La contextualidad, respaldada por el principio de contextualidad, enfatiza que el significado de palabras y expresiones está intrínsecamente ligado a su contexto de uso.

La distribucionalidad es un concepto crucial en la semántica y el procesamiento del lenguaje natural. Esta teoría se basa en la idea de que el significado de una palabra se

puede inferir a partir de los contextos en los que aparece. Esta idea se remonta a los trabajos de Zellig Harris (1954), quien postuló que «la distribución de una palabra en diferentes contextos proporciona información sobre su significado» (Harris, 1954:146). Esta noción fue posteriormente popularizada por John Firth con su famosa cita «*You shall know a word by the company it keeps*» (Firth, 1957:11).

Esta teoría distribucional ha tenido un impacto significativo en el desarrollo de modelos de representación semántica, particularmente en los embeddings de palabras, como Word2Vec (Mikolov et al., 2013) y GloVe (Pennington et al., 2014). Desde una perspectiva lingüística, la integración de la composicionalidad y la contextualidad es fundamental. La teoría distribucional, tal como se plantea en el trabajo de Harris y Firth, permite que los modelos de lenguaje natural capturen las relaciones contextuales y la coocurrencia de palabras en grandes corpus textuales para aprender representaciones vectoriales de palabras que capturan similitudes semánticas basadas en sus contextos de uso. Por ejemplo, en Word2Vec, las palabras que aparecen en contextos similares tienden a tener representaciones vectoriales cercanas en el espacio semántico.

En este sentido, los embeddings, aunque tienen una base distribucional al representar palabras como puntos en un espacio vectorial continuo, están arraigados en la idea de capturar el significado a través de la información contextual. Los embeddings, al ser representaciones distribucionales, reflejan la coocurrencia y relaciones contextuales de las palabras. Sin embargo, a pesar de la riqueza de información que proporciona el contexto y los embeddings distribucionales, el enfoque en la composicionalidad sigue siendo crucial. La composición de las partes individuales de una expresión, guiada por reglas sintácticas y semánticas, sigue siendo esencial para construir significados más complejos.

En la integración de la composicionalidad y la contextualidad, la síntesis kantiana emerge como una base teórica relevante. De hecho, en Amigó et al. (2022) se expone que Maruyama (2019) planteó la necesidad de una síntesis kantiana en el contexto de la semántica composicional. La idea se basa en la filosofía de Immanuel Kant, quien enfatizaba que el conocimiento surge de la interacción entre la experiencia sensorial (empirismo) y las estructuras cognitivas innatas (racionalismo). Se interpreta como la necesidad de integrar el enfoque distribucional (contextualidad) y el enfoque composicional (sistematicidad) para capturar el significado de las formas lingüísticas de manera más completa y precisa.

La síntesis kantiana es otro enfoque teórico que ha influido en la semántica y la filosofía del lenguaje. Derivada de la filosofía de Immanuel Kant, la síntesis kantiana propone que el conocimiento se forma a través de la integración activa de experiencias sensoriales y conceptos preexistentes. Kant (1781) argumenta que «los conceptos sin intuiciones son vacíos; las intuiciones sin conceptos son ciegas» (Kant, 1781: 93), subrayando la interdependencia entre la percepción sensorial y la estructura conceptual.

En el contexto del procesamiento del lenguaje natural, la síntesis kantiana sugiere un enfoque híbrido que combina datos empíricos (como corpus textuales) con estructuras conceptuales predefinidas. Este enfoque puede ser visto en modelos que integran conocimientos previos con datos observacionales para mejorar la representación semántica. Por ejemplo, los modelos de lenguaje preentrenados como BERT no solo aprenden de grandes cantidades de texto, sino que también pueden ser ajustados con tareas específicas que incorporan conocimiento contextual y situacional. La síntesis kantiana también se refleja en la semántica composicional cuando se considera cómo las palabras y las oraciones no solo tienen significados inherentes, sino que estos significados pueden ser modificados y enriquecidos por el contexto en el que se utilizan. Sentis (2006) aborda este punto al discutir cómo «la interpretación de una expresión lingüística no puede ser completamente comprendida sin considerar los factores contextuales y situacionales que influyen en su uso» (Sentis, 2006:80).

Por tanto, como ya se ha demostrado, integrar estos conceptos en la semántica composicional proporciona una comprensión más rica y matizada del significado lingüístico. La distribucionalidad proporciona una base empírica para capturar similitudes semánticas a partir de patrones de coocurrencia en grandes corpus textuales, mientras que la síntesis kantiana enfatiza la necesidad de integrar datos empíricos con estructuras conceptuales y contextuales. Estos enfoques se deben combinar para mejorar la precisión y la coherencia de las representaciones semánticas en el procesamiento del lenguaje natural.

2.5 Fundamentos lógico-lingüísticos de las oraciones

La semántica de Montague es un enfoque en la lingüística y la filosofía del lenguaje que fue desarrollado por el lingüista Richard Montague en la década de 1970. Este enfoque utiliza herramientas de la lógica formal para modelar la estructura y el

significado del lenguaje natural de una manera precisa y rigurosa. La idea central es tratar el lenguaje como un sistema formal, lo que permite analizar su estructura y su significado de manera sistemática. Montague introdujo varios conceptos fundamentales en su semántica formal, incluyendo la noción de tipos semánticos y la aplicación de la teoría de conjuntos y la lógica de primer orden para representar el significado de las expresiones lingüísticas. Su enfoque también abordó problemas importantes en la semántica, como la ambigüedad y la composicionalidad del significado. Estas ideas se presentan en el libro *Introductions to Montague Semantics*, escrito por David Dowty, Robert E. Wall y Stanley Peters.

Uno de los conceptos clave de la semántica de Montague es la idea de que el significado de una oración puede ser representado mediante una función que mapea la oración a un valor de verdad en un modelo semántico. Este enfoque utiliza herramientas de la lógica formal, como la lógica de primer orden y la teoría de conjuntos, para realizar esta representación formal del significado. Montague también introdujo la noción de tipos semánticos para capturar la estructura del significado en los lenguajes naturales. Los tipos semánticos son similares a los tipos en la programación, y ayudan a garantizar que las expresiones sean compatibles semánticamente. Por ejemplo, un sustantivo como «perro» podría tener un tipo semántico que lo distingue de un verbo como «correr».

La propuesta de Montague en la Semántica de Montague se basa en tres pilares fundamentales que son clave para comprender y desarrollar su enfoque formal del significado lingüístico:

1. **Semántica de condición de verdad (Truth Conditional Semantics) de Montague:** Este pilar se centra en la idea de que el significado de una expresión lingüística se determina por las condiciones bajo las cuales esa expresión sería verdadera o falsa en el mundo. Montague adoptó la visión de que el significado de una oración en un lenguaje natural se puede capturar mediante una representación formal que especifica las condiciones de verdad de esa oración en diferentes situaciones del mundo. Utilizó la lógica de primer orden para formalizar estas condiciones de verdad y así poder representar el significado de las expresiones lingüísticas de manera precisa y sistemática.
2. **Semántica modelística (Model-Theoretic Semantics):** Este segundo pilar se basa en la idea de utilizar modelos matemáticos para representar la estructura y el

significado del lenguaje natural. Montague propuso que las estructuras semánticas que representan el significado de las expresiones lingüísticas pueden ser modeladas mediante estructuras matemáticas llamadas modelos. Estos modelos consisten en conjuntos de objetos del mundo y relaciones entre ellos. Al asignar interpretaciones precisas a los términos del lenguaje y las relaciones entre ellos, los modelos pueden capturar el significado de las expresiones lingüísticas y las relaciones entre ellas de una manera formal y precisa.

3. **Semántica de mundos posibles (Possible Worlds Semantics):** Este tercer pilar se basa en la noción de que el significado de una expresión lingüística depende de las posibles situaciones o mundos en los que esa expresión podría ser verdadera o falsa. Montague propuso que el significado de una expresión se puede entender en términos de su verdad o falsedad en diferentes mundos posibles. Esta noción permite capturar la modalidad, la contingencia y otras relaciones semánticas importantes entre las expresiones lingüísticas. Al considerar una variedad de mundos posibles y sus relaciones, la semántica de mundos posibles proporciona un marco flexible para analizar y comprender el significado del lenguaje natural.

Montague argumentaba que entender cómo las palabras y las frases se combinan en una oración es esencial para determinar su significado semántico. Esta perspectiva refleja su enfoque en la composicionalidad del significado lingüístico. Asimismo, su propuesta exponía que el significado de una oración se deriva de manera sistemática de las contribuciones semánticas de sus partes constituyentes y de la manera en que están organizadas sintácticamente. Esto significa que el significado de una oración no es simplemente la suma de los significados de las palabras individuales, sino que surge de la manera en que esas palabras se combinan según las reglas gramaticales del idioma.

Por ejemplo, se plantea la oración «Ana come una manzana». En este caso, el significado de la oración no se limita a los significados de las palabras «Ana», «come» y «manzana» por separado, sino que también depende de cómo se combinan estas palabras en la estructura sintáctica de la oración. La relación gramatical entre «Ana», «come» y «manzana» en la oración determina el significado semántico general de la oración completa.

Montague planteaba que esta estructuración sintáctica de las oraciones proporciona una base para el análisis semántico del lenguaje natural. Al comprender cómo

las palabras y oraciones se organizan en la estructura de una oración, podemos inferir cómo se relacionan y contribuyen al significado global de la oración. Esto permite un análisis semántico preciso y sistemático del significado de las oraciones en el lenguaje natural.

Como propuesta de enfoque formal y matemático para analizar el lenguaje natural, a continuación, se presentan los conceptos esenciales en la semántica de Montague:

1. **Lógica de primer orden:** Montague utilizó la lógica de primer orden como un marco formal para representar el significado del lenguaje natural. En la lógica de primer orden, se pueden definir predicados, funciones y cuantificadores para capturar las relaciones semánticas entre los términos. Montague representó el significado de las oraciones en términos de fórmulas de lógica de primer orden, lo que le permitió aplicar reglas de inferencia y realizar análisis semántico de manera rigurosa. En la lógica de primer orden, se utilizan símbolos como cuantificadores (\forall para «para todo», \exists para «existe»), predicados ($P(x)$, $Q(x)$), funciones ($f(x)$), y conectivos lógicos (\wedge para «y», \vee para «o», \neg para «no», \rightarrow para «implica»). Montague representara la estructura semántica de una oración utilizando estos símbolos para expresar las relaciones y cuantificaciones presentes en el lenguaje.
2. **Teoría de conjuntos:** La teoría de conjuntos proporcionó a Montague una manera de representar las estructuras semánticas y las relaciones entre los objetos del mundo. Utilizó conjuntos para representar el dominio del discurso (el conjunto de todos los objetos a los que se refiere el lenguaje), así como relaciones entre estos objetos. Por ejemplo, un conjunto podría representar el conjunto de todas las manzanas, y una relación podría representar la acción de «comer». En la teoría de conjuntos, se utilizan símbolos como conjuntos ($\{ \}$), \in para «pertenencia»), operaciones de conjuntos (\cup para unión, \cap para intersección, complemento), y relaciones de conjunto (\subseteq para «subconjunto de»). Montague emplea estos símbolos para definir el dominio del discurso y las relaciones entre los objetos referenciados por el lenguaje.
3. **Modelos semánticos:** Montague utilizó modelos semánticos para interpretar las expresiones lingüísticas y capturar su significado. Un modelo semántico consiste en un dominio del discurso (conjunto de objetos) y una interpretación de los

términos y predicados del lenguaje en términos de este dominio. Los modelos proporcionan una manera de determinar las condiciones de verdad de las oraciones y expresiones lingüísticas en diferentes situaciones del mundo. En la teoría de modelos, se utilizan símbolos para representar los objetos del dominio del discurso (a, b, c, \dots) y relaciones entre ellos ($R(a, b), P(c), \dots$). Montague asigna interpretaciones precisas a estos símbolos para representar el significado de las expresiones lingüísticas en términos de condiciones de verdad en diferentes situaciones del mundo.

4. **Álgebra de Boole:** Montague también se basó en conceptos de álgebra de Boole para formalizar la estructura y el significado de las expresiones lingüísticas. En particular, utilizó operaciones lógicas como la conjunción, la disyunción y la negación para representar relaciones semánticas entre proposiciones y para manipular fórmulas lógicas en el análisis semántico. En el álgebra de Boole, se utilizan símbolos como 0 y 1 para representar los valores de verdad «falso» y «verdadero», respectivamente. Se emplean también operaciones lógicas como \wedge (y), \vee (o), y \neg (no). Montague utiliza estas operaciones para manipular proposiciones y derivar el significado de expresiones más complejas a partir de las más simples.
5. **Teoría de modelos de conjuntos de oraciones:** Montague introdujo la teoría de modelos de conjuntos de frases como una extensión de la teoría de modelos estándar para capturar el significado de oraciones complejas. Esta teoría permitió analizar y representar el significado de oraciones que involucran cuantificadores y cláusulas subordinadas de manera sistemática y precisa. Montague introdujo notaciones adicionales para representar el significado de oraciones complejas que involucran cuantificadores y cláusulas subordinadas. Esto incluiría el uso de símbolos como $\forall x, \exists y$ para cuantificadores y conectores lógicos para conectar proposiciones en cláusulas subordinadas.

Por tanto, la semántica de Montague utiliza conceptos formales de la lógica de primer orden, la teoría de conjuntos, los modelos semánticos, el álgebra de Boole y la teoría de modelos de conjuntos de frases para proporcionar una representación precisa y sistemática del significado del lenguaje natural. Esta representación permite analizar y

manipular estructuras lingüísticas de manera rigurosa y coherente. Asimismo, todo esto demuestra una búsqueda de sistematicidad en la representación de la lengua.

De hecho, la jerarquía en árbol es una representación gráfica y estructural que muestra cómo se organizan las diferentes partes de una oración o expresión lingüística. Mediante el uso de los conceptos matemáticos mencionados, es posible plantear jerarquías en árbol que representen la estructura sintáctica y semántica de las oraciones. Esta jerarquía está diseñada para reflejar tanto la estructura sintáctica como el significado semántico de una oración. En este marco, cada nodo en el árbol de derivación corresponde a un constituyente sintáctico, ya sea una palabra individual, un sintagma, una cláusula, etc. La forma en que estos constituyentes se organizan en el árbol refleja la estructura gramatical de la oración y cómo se combinan para formar la estructura sintáctica completa. La jerarquía en árbol proporciona una representación visual de cómo se estructura la oración y cómo se relacionan sus componentes. Esto es crucial para el análisis sintáctico y semántico, ya que Montague argumentaba que el significado de una oración está directamente relacionado con su estructura sintáctica.

En el análisis de Montague, una oración se descompone en sus componentes básicos, y cada uno de estos componentes se representa en el árbol de derivación. Este árbol no solo muestra la estructura gramatical, sino que también mapea cómo se combina el significado de cada componente para formar el significado global de la oración. Por ejemplo, una oración como «El gato negro duerme en el sofá» se descompondría en sus componentes: «El gato negro» (sintagma nominal), «duerme» (verbo), y «en el sofá» (sintagma preposicional). Cada uno de estos componentes se desglosa a su vez en sus constituyentes más pequeños hasta llegar a las palabras individuales. El árbol resultante no solo muestra cómo se agrupan estos constituyentes gramaticalmente, sino que también proporciona una guía sobre cómo interpretar cada constituyente en el contexto de la oración completa.

La jerarquía en árbol es esencial porque permite visualizar la relación entre los constituyentes y su contribución al significado de la oración. Cada nodo y cada conexión en el árbol contribuyen a la construcción del significado semántico global de la oración. Por ejemplo, la estructura jerárquica muestra que «duerme» es el verbo principal y que «en el sofá» especifica el lugar del verbo, mientras que «El gato negro» es el sujeto que realiza la acción. Esta relación estructural es fundamental para entender el significado de

la oración: el árbol muestra claramente quién realiza la acción, cuál es la acción y dónde ocurre la acción.

Asimismo, también se argumenta que el significado de una oración puede ser derivado a partir de la composición semántica de sus partes constituyentes. Esta idea es fundamental en su enfoque de la Semántica de Montague y refleja su perspectiva de la composicionalidad del lenguaje natural. Por tanto, una vez más se recalca que la composicionalidad implica que el significado de una oración compleja se construye a partir del significado de sus partes componentes y de la forma en que estas partes se combinan sintácticamente.

Por otro lado, Montague sostiene que entender cómo las palabras y las frases se combinan en una oración es esencial para determinar su significado semántico. Este enfoque se basa en la premisa de que el significado de una oración no es simplemente la suma de los significados de las palabras individuales, sino que también depende de cómo estas palabras se estructuran y se relacionan entre sí. La jerarquía en árbol facilita esta comprensión al proporcionar una representación clara de estas relaciones. Cada nivel del árbol añade una capa de significado que contribuye al entendimiento completo de la oración. Por ejemplo, la combinación de «gato» y «negro» en «El gato negro» añade la propiedad de «negro» al «gato», y esta combinación a su vez se relaciona con el verbo «duerme», estableciendo que es el «gato negro» quien «duerme».

Además, la jerarquía en árbol también permite analizar cómo las distintas estructuras sintácticas pueden dar lugar a diferentes interpretaciones semánticas. Por ejemplo, la oración «El gato negro duerme en el sofá» tiene una estructura jerárquica que especifica claramente quién realiza la acción y dónde se realiza. Cambiar la estructura a «En el sofá duerme el gato negro» mantiene el mismo significado básico, pero cambia la estructura sintáctica, lo que puede poner énfasis en diferentes partes de la oración.

Por tanto, es posible concluir que la jerarquía en árbol es una herramienta que no solo resulta útil para el análisis sintáctico-semántico, sino también para la comprensión y representación del lenguaje natural.

2.6 Estructura y tipología sintáctica: oraciones y componentes lingüísticos

Para asentar unas bases lingüísticas sólidas sobre las cuales plantear una propuesta de orden en la composición semántica dentro del marco ICDS, es crucial realizar una revisión exhaustiva de la tipología sintáctica. Este análisis debe abarcar todos los elementos y componentes que aparecen en las oraciones y que pueden ser relevantes para el estudio de la semántica composicional. Una comprensión detallada de la tipología sintáctica no solo facilita la estructuración adecuada de las oraciones, sino que también permite una integración más precisa de los principios de composicionalidad y distribucionalidad en los modelos semánticos. De esta forma, se plantearán los componentes lingüísticos desde un enfoque macro (oración) hasta un enfoque micro (sintagmas). Para ello, se recurrirá al libro *Oxford English Grammar* (Greenbaum, 2005), donde se aborda la tipología sintáctica de la lengua inglesa.

En primer lugar, una oración es una unidad lingüística compuesta por un sujeto y un predicado que expresan una idea completa. En una oración, el sujeto es la entidad de la que se habla o que realiza la acción, mientras que el predicado expresa la acción que realiza el sujeto o alguna característica o estado relacionado con él. Una oración puede ser simple, compuesta o compleja, dependiendo de la cantidad y la estructura de las cláusulas que la componen. Por tanto, se pueden diferenciar los siguientes tipos de oraciones:

Oraciones simples:

1. **Oraciones simples:** Consisten en una única cláusula independiente, lo que significa que tienen un sujeto y un predicado y expresan una idea completa. Pueden estar solas como una oración completa. Por ejemplo: *The cat sat on the mat.*

Oraciones coordinadas:

2. **Oraciones coordinadas:** Consisten en dos o más cláusulas independientes unidas por conjunciones coordinadas como *and, but, or, nor, for, so, o yet*. Cada cláusula independiente en una oración coordinada puede estar sola como una oración completa. Por ejemplo: *I like to read, but my sister prefers to watch TV.*

Oraciones subordinadas:

3. **Oraciones subordinadas:** Contienen una cláusula independiente y al menos una cláusula dependiente. Las cláusulas dependientes no pueden estar solas como oraciones completas porque dependen de la cláusula independiente para su contexto y significado. Suelen formarse mediante conjunciones subordinadas como *because, although, while, since, if, when*, etc. Por ejemplo: *Because it was raining, we stayed indoors.*

Como se ha podido observar, tanto las oraciones simples como las coordinadas son cláusulas independientes, lo que significa que se interpretan como oraciones aisladas. Las oraciones simples contienen un solo sujeto y predicado, mientras que las oraciones coordinadas se forman mediante la unión de dos o más oraciones simples usando conjunciones coordinantes. Sin embargo, es importante presentar y entender los tipos de oraciones subordinadas, ya que son fundamentales para la estructura y el significado de las oraciones complejas en inglés. Las oraciones subordinadas son cláusulas que dependen de una oración principal para tener sentido completo. Estas oraciones no pueden existir por sí solas y generalmente están introducidas por conjunciones subordinantes o pronombres relativos. En inglés, existen varios tipos de oraciones subordinadas, cada una con funciones específicas dentro de la oración principal: sustantivas, adjetivas y adverbiales.

Tipos de oraciones subordinadas

1. Oraciones subordinadas sustantivas:

- Actúan como un sustantivo dentro de la oración principal.
- Pueden funcionar como sujeto, objeto directo, objeto indirecto, complemento del sujeto o del objeto.

2. Oraciones subordinadas adjetivas:

- Complementan a un sustantivo o pronombre en la oración principal.
- Se introducen por pronombres relativos como *who, whom, whose, which, that*.
- Pueden ser restrictivas (esenciales para el significado) o no restrictivas (añaden información adicional).

3. Oraciones subordinadas adverbiales:

- Funcionan como adverbios, modificando un verbo, adjetivo o adverbio en la oración principal.
- Introducidas por conjunciones subordinantes como *because, although, if, when, since, while, etc.*
- Indican complementos circunstanciales de tiempo, lugar, causa, condición, etc.

Por otro lado, es esencial destacar que las oraciones subordinadas deben analizarse con cuidado porque son, esencialmente, oraciones completas dentro de otra oración. Por tanto, dentro de la propia oración subordinada hay que llevar un análisis sintáctico. Esto implica que hay distintas jerarquías dentro de las oraciones, por lo que es posible plantear unos criterios de orden dentro de este tipo de oraciones.

Continuando con la tipología de las oraciones, estas pueden pertenecer a la voz pasiva o a la voz activa:

- **Oraciones activas:** En una oración activa, el sujeto realiza la acción expresada por el verbo. La estructura típica de una oración activa es Sujeto-Verbo-Objeto (SVO), donde el sujeto realiza la acción sobre el objeto. Por ejemplo: *El perro persigue al gato (The dog chases the cat)*. En esta oración, *el perro* es el sujeto que realiza la acción de *perseguir*, y *al gato* es el objeto que recibe la acción.
- **Oraciones pasivas:** En una oración pasiva, el objeto de la acción se convierte en el sujeto de la oración, y el sujeto original puede omitirse o colocarse en una posición secundaria. La estructura típica de una oración pasiva es Objeto-Verbo-Sujeto (OVS) o Verbo-Sujeto-Objeto (VSO), donde el objeto se convierte en el sujeto de la oración y el verbo se transforma en una forma pasiva. Por ejemplo: *El gato es perseguido por el perro (The cat is chased by the dog)*. En esta oración pasiva, *el gato* se convierte en el sujeto y *es perseguido* es la forma pasiva del verbo *perseguir*, mientras que *por el perro* indica quién realiza la acción.

Asimismo, tras haber presentado los tipos de oraciones, es necesario tener en cuenta cómo se forman las oraciones, que se dividen en sujeto y predicado:

1. El sujeto de una oración en inglés es la parte de la oración sobre la cual se está hablando. Es quién o qué realiza la acción expresada por el verbo.

2. El predicado de una oración es la parte que indica lo que el sujeto está haciendo o la acción que se está llevando a cabo. Contiene el verbo de la oración y puede incluir otros elementos como objetos, complementos, y modificadores que describen o completan la acción del sujeto. El predicado viene después del sujeto en la oración.

Tipos de sujeto:

1. **Sujeto simple:** Es un solo sustantivo o pronombre que realiza la acción del verbo en la oración. Por ejemplo: *The dog barks.*
2. **Sujeto compuesto:** Consiste en dos o más sustantivos o pronombres que realizan la acción del verbo en la oración. Estos sujetos se unen con una conjunción como *and* o *or*. Por ejemplo: *John and Mary are friends.*
3. **Sujeto omitido:** En algunas oraciones, el sujeto se omite porque es obvio o implícito en el contexto. Por ejemplo: *Go to the store.* En esta oración, el sujeto *you* está implícito, es decir, se entiende que se dirige a *tú* aunque no se mencione explícitamente.
4. **Sujeto interrogativo:** En preguntas, el sujeto puede ser un pronombre interrogativo como *who* o *what*. Por ejemplo: *Who is at the door?*
5. **Sujeto enfático (aposición):** A veces, para enfatizar o destacar al sujeto, se usa un pronombre personal o sustantivo al principio de la oración, incluso si no es estrictamente necesario gramaticalmente. Por ejemplo: *He, John, is the one who found the keys.*

Tipos de predicado:

1. **Predicado simple:** Consiste en el verbo principal de la oración, que indica la acción o estado del sujeto. Por ejemplo: *The dog barks.*
2. **Predicado compuesto:** Incluye dos o más verbos principales que comparten el mismo sujeto y están unidos por una conjunción como *and* o *or*. Por ejemplo: *She sings and dances.*
3. **Predicado nominal:** En este tipo de predicado, el verbo principal es un verbo copulativo (como *be*, *seem*, *become*, *appear*, entre otros) y se utiliza para identificar o describir al sujeto. Se complementa con un atributo, que puede ser

un sustantivo, un adjetivo, o un pronombre. Por ejemplo: *He is a doctor*. En esta oración, *is* es el verbo copulativo y *a doctor* es el atributo que describe al sujeto *he*.

4. **Predicado verbal:** En contraste con el predicado nominal, el predicado verbal contiene un verbo principal que describe una acción realizada por el sujeto. Este tipo de predicado no incluye un atributo. Por ejemplo: *She writes novels*. En esta oración, *writes* es el verbo principal que describe la acción realizada por *she*.

Dentro de los predicados, también se distinguen distintos tipos de verbos, lo que condiciona los complementos que aparecerán dentro del predicado. Esto es importante tenerlo en cuenta, ya que hay elementos concretos que solo aparecen en un tipo determinado de predicado.

Verbos copulativos:

Los verbos copulativos, también conocidos como verbos de enlace, son verbos que se utilizan para conectar el sujeto de la oración con un atributo que lo describe. Los atributos pueden ser sustantivos, adjetivos o frases preposicionales. Los verbos copulativos más comunes son *be* (ser/estar), *seem* (parecer), *appear* (aparecer), *become* (convertirse en), entre otros.

1. Con sustantivos como atributo: *She is a doctor*.
2. Con adjetivos como atributo: *The cake looks delicious*.
3. Con frases preposicionales como atributo: *He feels at home in this city*.

Verbos transitivos e intransitivos:

Los verbos transitivos requieren un objeto directo, es decir, algo o alguien que recibe la acción del verbo. Por otro lado, los verbos intransitivos no requieren un objeto directo y la acción del verbo no se transfiere a ningún receptor.

1. Verbos transitivos:

- Requieren un objeto directo para completar el significado de la acción.
- *She eats an apple*.
- En este caso, *an apple* es el objeto directo que recibe la acción del verbo *eats*.

2. Verbos intransitivos:

- No requieren un objeto directo para completar el significado de la acción.
- *He sleeps.*
- No hay un objeto directo que reciba la acción del verbo *sleeps*.

3. Verbos transitivos e intransitivos:

- Algunos verbos pueden ser transitivos o intransitivos dependiendo del contexto.
- *He reads a book.* - Transitivo
- *He reads every night.* - Intransitivo

Por otro lado, los complementos juegan un papel esencial en la construcción y comprensión de las oraciones. Un complemento es un elemento que proporciona información adicional sobre otros elementos de la oración, como el sujeto, el verbo o el objeto. Los complementos pueden modificar, describir o completar el significado de estos elementos, y su correcta utilización es esencial para la claridad y precisión en la comunicación. Los complementos se pueden clasificar en varias categorías, cada una con una función específica dentro de la oración.

1. **Complemento directo (CD):** En el caso de verbos transitivos, el complemento directo recibe la acción del verbo y es indispensable para completar su significado. Por ejemplo, en *She reads a book*, *a book* es el complemento directo que completa el significado de *reads*. Los complementos directos pueden tener sus propios modificadores, como en *the beautiful present*, donde *beautiful* modifica al núcleo «present», destacando la importancia del núcleo dentro del mismo sintagma.
2. **Complemento indirecto (CI):** El complemento indirecto indica a quién o para quién se realiza la acción del verbo. Aunque no siempre es indispensable como el CD, su presencia es crucial para ciertos verbos ditransitivos. Por ejemplo, en *She gave him a gift*, *him* es el complemento indirecto y *a gift* es el complemento directo.
3. **Complemento de régimen (CR):** Este complemento se requiere con verbos preposicionales, donde el verbo necesita una preposición específica para tener

sentido completo. Por ejemplo, en *She relies on him*, *on him* es el complemento de régimen.

4. **Complementos circunstanciales (CC):** Estos complementos proporcionan información adicional sobre la acción del verbo, respondiendo preguntas como cómo, cuándo, dónde, por qué o en qué medida. Aunque no son esenciales para la estructura gramatical básica, son importantes para enriquecer el significado de la oración. Por ejemplo, en *She sings beautifully*, *beautifully* es un complemento circunstancial que modifica el verbo *sings*.
5. **Complementos nominales:** Añaden información sobre el sujeto u objeto y completan el significado. Pueden ser sustantivos, pronombres o frases nominales que describen o identifican al sujeto u objeto. Por ejemplo, en *He is a teacher*, *a teacher* es un complemento nominal que identifica al sujeto *He*.
6. **Complementos adjetivales:** También conocidos como adjetivos predicativos o complementos del sujeto, describen o modifican al sujeto o al objeto y completan el significado de un verbo o verbo de enlace. Por ejemplo, en *The flowers smell lovely*, *lovely* es un complemento adjetival que modifica al sujeto *The flowers*.

Para establecer una jerarquía de importancia entre los complementos dentro de las oraciones, es útil considerar varias teorías lingüísticas y enfoques gramaticales. Una de las más relevantes es la Gramática Generativa de Noam Chomsky, que enfatiza la estructura jerárquica de las oraciones y la dependencia de ciertos elementos sobre otros. Según Chomsky, algunos complementos se consideran más esenciales que otros debido a su papel en la construcción del significado básico de una oración (Chomsky, 1965).

Los complementos juegan un papel esencial en la construcción y comprensión de las oraciones. Un complemento es un elemento que proporciona información adicional sobre otros elementos de la oración, como el sujeto, el verbo o el objeto. Los complementos pueden modificar, describir o completar el significado de estos elementos, y su uso puede ser esencial para la claridad y precisión en la comunicación. Los complementos se pueden clasificar en varias categorías, cada una con una función específica dentro de la oración.

Entre los complementos esenciales se encuentran el complemento directo (CD) y el complemento indirecto (CI). El complemento directo es fundamental en el caso de verbos transitivos, ya que recibe la acción del verbo y es indispensable para completar su significado. El complemento indirecto indica a quién o para quién se realiza la acción del

verbo. Aunque no siempre es indispensable como el CD, su presencia es crucial para ciertos verbos ditransitivos. El complemento de régimen se requiere con verbos preposicionales, donde el verbo necesita una preposición específica para tener sentido completo.

Los complementos opcionales pero relevantes incluyen los complementos adverbiales (CC) y los atributos o complementos del predicado nominal. Los complementos adverbiales proporcionan información adicional sobre la acción del verbo, como tiempo, lugar, modo, causa, etc. Aunque no son esenciales para la estructura gramatical básica, son importantes para enriquecer el significado de la oración. En oraciones con verbos copulativos, los atributos que siguen al verbo son esenciales para completar el significado del sujeto.

Desde la perspectiva de la Gramática Generativa, los complementos pueden ser clasificados según su obligatoriedad y función dentro de la oración. Los complementos obligatorios incluyen el complemento directo (CD) y el complemento de régimen (CR). Los complementos opcionales pero relevantes incluyen el complemento indirecto (CI), los complementos adverbiales (CC) y los atributos del predicado nominal. Además, existen complementos contextuales cuya necesidad y relevancia pueden variar según el contexto y la intención comunicativa del hablante, como algunos complementos adverbiales y atributos.

La importancia de los complementos también puede ser vista desde la perspectiva de su capacidad para alterar el significado esencial de la oración. Por ejemplo, la eliminación del CD en verbos transitivos deja incompleta la acción, como en *She reads* que no tiene el mismo significado completo que *She reads a book*. La eliminación del CI en verbos ditransitivos deja ambiguo a quién se dio el regalo, como en *She gave a gift* comparado con *She gave him a gift*.

Dentro de estas grandes categorías se engloban también otro tipo de complementos, como puede ser el complemento agente dentro de los complementos nominales, entre otros de los muchos tipos de complementos que se disponen dentro de estas categorías. Por tanto, como conclusión general y tras haber presentado todos los tipos de complementos, se plantea que existe una estructura básica en la construcción de oraciones en inglés que se compone de sujeto y predicado. Esta estructura es fundamental

para la formación de oraciones y puede ser ampliada con la adición de diversos complementos.

La estructura básica de una oración en inglés se conforma por un sujeto y un verbo (SV). Esta estructura simple puede ser suficiente para formar una oración completa y gramaticalmente correcta. Sin embargo, dependiendo del tipo de verbo utilizado, pueden ser necesarios complementos adicionales para completar el significado de la oración. Aunque muchos complementos son opcionales y la oración puede seguir siendo gramaticalmente correcta sin ellos, estos complementos suelen ser importantes para proporcionar información adicional y contexto. Los complementos adverbiales, por ejemplo, añaden detalles sobre cómo, cuándo, dónde, por qué o en qué medida se realiza una acción, enriqueciendo así la comunicación. Sin embargo, su omisión no afecta la estructura gramatical básica de la oración.

Dado que los objetos directos e indirectos son esenciales para ciertos verbos, se parte de la premisa de que estos complementos tienen una mayor importancia en la oración. En contraste, los complementos adverbiales y otros tipos de complementos, aunque no esenciales, se estructuran típicamente de derecha a izquierda dentro de la oración. Esto se debe a la ausencia de un método definitivo para determinar la relevancia de los complementos adverbiales, cuya importancia puede depender de factores como la entonación o el orden de las palabras. De hecho, para enfatizar ciertos complementos adverbiales, es común colocarlos al inicio de la oración o al inicio del predicado. Al seguir esta estructura de derecha a izquierda, se logra una composición que refleja mejor la importancia de los complementos en el contexto, facilitando una comunicación más clara y precisa.

En conclusión, la jerarquía de importancia de los complementos en una oración se basa en su necesidad para completar el significado del verbo y, por extensión, de la oración misma. Los complementos directos y de régimen se consideran esenciales en muchos contextos, mientras que los complementos indirectos y adverbiales, aunque opcionales, juegan un papel crucial en la claridad y riqueza de la comunicación. Este entendimiento puede ser trasladado al ámbito computacional para mejorar los modelos de representación semántica, asegurando que los complementos esenciales se prioricen en el análisis y generación de oraciones.

2.7 Conclusiones

Gracias a la revisión exhaustiva de los fundamentos teóricos en este trabajo, hemos logrado profundizar significativamente en los conceptos clave de la composicionalidad semántica, la síntesis kantiana y la distribucionalidad, y su impacto en el procesamiento del lenguaje natural. Esta revisión no solo ha permitido una mejor comprensión del marco ICDS, sino que también ha facilitado un análisis detallado desde una perspectiva lingüística, lo cual es esencial para el desarrollo de técnicas avanzadas y efectivas en el campo del procesamiento del lenguaje natural.

La revisión del estado del arte ha sido fundamental para identificar las fortalezas y limitaciones de los enfoques existentes en la representación semántica. En este proceso, se ha observado que el orden de las palabras y la estructura sintáctica son cruciales para mantener la integridad semántica del texto. De hecho, los métodos que ignoran estas tienden a perder información significativa sobre las relaciones entre palabras. Esto acaba resultando en representaciones semánticas que no capturan adecuadamente el significado del texto. Por otro lado, a pesar de los avances logrados con modelos secuenciales como LSTM y Transformers, estos todavía enfrentan desafíos relacionados con la preservación de la semántica cuando no se consideran las estructuras jerárquicas. La falta de un enfoque estructurado puede llevar a problemas de degradación de la representación, especialmente en capas superiores de redes neuronales contextuales.

Aunque en el marco ICDS se aborde la composición a nivel de palabras, este apartado buscaba no solo considerar el orden de las palabras, sino también las relaciones sintácticas y estructurales entre ellas, lo cual es esencial para mantener la coherencia semántica. Las estructuras jerárquicas permiten capturar de manera más precisa las relaciones entre los constituyentes de una oración. Por ejemplo, en lugar de procesar una oración simplemente de izquierda a derecha o de derecha a izquierda, una estructura jerárquica analizaría cómo se agrupan las palabras en oraciones y cómo se relacionan entre sí. Este enfoque jerárquico ofrece varias ventajas. En primer lugar, mejora la precisión semántica al mantener las relaciones entre palabras y oraciones. Además, aunque es más complejo, el enfoque jerárquico es escalable y puede adaptarse a diferentes niveles de granularidad lingüística, desde palabras individuales hasta oraciones simples o compuestas.

Además, esta revisión ha permitido observar una relación jerárquica en las oraciones sintácticas, donde no todos los elementos cobran la misma relevancia debido a que unos dependen de otros. Algunos complementos sirven para aportar más información adicional, mientras que otros simplemente no se pueden omitir, como el verbo, ya que, sin él, no habría oración. Esta jerarquía y dependencia de ciertos elementos sobre otros es crucial para comprender la estructura y significado de las oraciones.

La revisión también ha subrayado la necesidad urgente de establecer un orden claro y consistente dentro del marco ICDS. Esta necesidad surge debido a la importancia de integrar oraciones de manera coherente en el espacio de word embeddings, lo que es crucial para asegurar la precisión y utilidad de las representaciones semánticas. Establecer un orden claro permitirá que las funciones de composición dentro del marco ICDS mantengan las relaciones sintácticas y la estructura del texto, lo que resulta en representaciones semánticas más coherentes y precisas, mejorando la capacidad del sistema para manejar tareas complejas de PLN. Además, se deben desarrollar funciones de composición que sean sensibles a la estructura jerárquica y evaluar su efectividad a través de estudios empíricos, incluyendo la comparación de métodos secuenciales y jerárquicos para determinar cuál ofrece la mejor precisión y coherencia en diversas aplicaciones de PLN.

En conclusión, se puede afirmar que gracias a esta revisión del estado del arte y al análisis detallado desde una perspectiva lingüística, se han podido establecer las bases para una propuesta de orden basada en el enfoque jerárquico de las oraciones. Esta propuesta, respaldada por la evidencia obtenida, busca mejorar significativamente la precisión y coherencia en la representación semántica dentro del marco ICDS.

3. Propuesta de modelo de orden

3.1. Introducción

En relación al marco teórico propuesto por Amigó et al. (2022) y a la semántica de Montague, se llevará a cabo una propuesta de modelo de orden. En primer lugar, en relación a la semántica de Montague, para este modelo se plantea que el significado de una oración se deriva de manera sistemática de las representaciones semánticas de sus componentes individuales (palabras) y de la forma en que están combinados sintácticamente. Este principio establece que el significado de una oración compleja es una función de los significados de sus partes constituyentes y de las reglas de combinación que rigen su estructura gramatical. Por otro lado, alineado con el marco teórico presentado por Amigó et al. (2022), se propone que la representación de una oración se fundamenta en la combinación de pares de vectores asociados a las palabras que la integran. En contraste con el enfoque tradicional de componer vectores de izquierda a derecha o de derecha a izquierda, esta metodología compone las palabras de dos en dos (basado en el marco ICDS) para generar una representación de oraciones.

La propuesta que se presenta en este trabajo se basa inicialmente en la oración simple. Esta elección se realiza para realizar una primera aproximación controlada y manejable a la representación de oraciones en espacios semánticos. Trabajar con oraciones simples permite acotar el ámbito de estudio y desarrollar un marco teórico y metodológico robusto antes de abordar estructuras lingüísticas más complejas. Las oraciones simples, que consisten en un solo sujeto y un solo predicado, ofrecen un punto de partida ideal para explorar la representación semántica debido a su estructura menos compleja. En esta fase inicial, se puede desarrollar y validar modelos de composición semántica centrándose en cómo se combinan los elementos básicos (sujeto, verbo, objeto) sin la complicación adicional de las cláusulas subordinadas, y establecer parámetros y métricas claras para facilitar la medición y evaluación de la precisión y coherencia de los modelos de representación semántica.

Aunque la propuesta inicial se centra en oraciones simples, el marco teórico desarrollado puede aplicarse a oraciones subordinadas teniendo en cuenta las jerarquías internas y externas. Esto implica descomposición y análisis jerárquico de cada oración subordinada como una oración completa en sí misma, identificando sus constituyentes (sujeto, verbo, objetos, complementos) y considerando cómo la subordinada se integra en

la oración principal y cuál es su función semántica (sustantiva, adjetiva, adverbial). La composición semántica se llevaría a cabo tanto a nivel interno, combinando las representaciones semánticas de los elementos dentro de la subordinada siguiendo principios de composicionalidad, como a nivel externo, integrando la representación semántica de la subordinada con la de la oración principal, ajustándose según su función específica.

3.2. Propuesta de orden jerárquico para oraciones simples

Para desarrollar una propuesta de orden jerárquico en la representación semántica de oraciones simples, es esencial descomponer la oración en sus componentes básicos y analizar cada uno de ellos en términos de su contribución a la estructura y significado global de la oración. Este análisis implica establecer una jerarquía clara tanto para el sujeto como para el predicado, y comprender cómo los diferentes elementos interactúan entre sí.

Orden jerárquico de prioridades:

Sujeto:

1. Núcleo del sujeto: El núcleo del sujeto, generalmente un sustantivo o pronombre, es el elemento más importante del sujeto, ya que establece quién o qué está realizando la acción principal en la oración. Por ejemplo, en «The cat sleeps,» «cat» es el núcleo del sujeto.
2. Complementos del sujeto: Los complementos del sujeto son aquellos elementos que modifican o proporcionan más información sobre el núcleo del sujeto, enriqueciendo la descripción y especificación del sujeto de manera más precisa. Por ejemplo, en «The skinny black cat», los adjetivos «skinny» y «black» son complementos del núcleo «cat».

Predicado:

1. Núcleo del predicado (verbo principal): El verbo principal es esencial para expresar la acción o el estado que ocurre en la oración. Es el elemento central del predicado y proporciona la información principal sobre lo que está sucediendo en la oración. Por ejemplo, en «The cat sleeps», «sleeps» es el núcleo del predicado.

2. Complementos directos, indirectos y de régimen: Los complementos directos e indirectos agregan información sobre la acción expresada por el verbo. Un complemento directo recibe la acción del verbo, mientras que un complemento indirecto indica a quién o para quién se realiza la acción. El complemento de régimen se requiere con verbos preposicionales, donde el verbo necesita una preposición específica para tener sentido completo. Dentro de estos complementos, también se pueden encontrar otros complementos que aporten información, pero al estar sintácticamente dentro del mismo nivel de jerarquía, se engloban dentro de esta categoría. El elemento más relevante es el núcleo del sintagma (el propio complemento directo, indirecto o de régimen en sí), y en un nivel más bajo, los elementos que modifican o aportan información adicional del núcleo del sintagma.
3. Complementos adicionales: Los complementos adicionales proporcionan información extra sobre el verbo, como adverbios que modifican la acción, indicando cómo, cuándo, dónde, por qué o en qué medida se realiza la acción. Por ejemplo, en «She sings beautifully», «beautifully» es un complemento adicional que modifica el verbo «sings».

Propuesta de modelo:

Este modelo se puede aplicar a nivel de palabras (donde la raíz o lexema consta como núcleo y sus flexiones y derivaciones serían sus «complementos» o «especificaciones») hasta a nivel de sintagma, sujeto, predicado, oración, etc. Para aplicar este modelo, en primer lugar, es importante tener en cuenta que cuando se realiza la composición semántica de una oración, se sigue el principio de componer desde los elementos menos relevantes hasta los más relevantes. Este enfoque garantiza que la información crítica y significativa se preserve de manera más efectiva, evitando la pérdida de detalles esenciales en el proceso de composición. Por tanto, los últimos elementos que entren en la composición serán los que más aporten.

La idea es que, al comenzar con los elementos menos relevantes, como los modificadores y los complementos adicionales, se establece una base que puede ser enriquecida gradualmente. Posteriormente, se integran los elementos más significativos, como los núcleos de los sintagmas, complementos directos e indirectos, y finalmente el

verbo principal. Este método asegura que la estructura fundamental y el significado principal de la oración se mantengan intactos y claros.

1. **Árbol jerárquico de la oración:** Primero se debe analizar la estructura sintáctica de la oración como un árbol jerárquico. Este árbol representa la relación entre las diferentes partes de la oración, desde los componentes más pequeños hasta los más grandes. Este análisis asegura que se identifiquen todas las relaciones jerárquicas y dependencias entre los elementos de la oración.
2. **División en sujeto y predicado:** Una vez que se tiene el árbol jerárquico de la oración, se divide la oración en sus dos partes principales: sujeto y predicado. Esta división se basa en el árbol sintáctico, donde ambos elementos tienen la misma importancia. El sujeto contiene la entidad que realiza la acción, mientras que el predicado describe la acción realizada por el sujeto.
3. **Análisis y composición del sujeto:** Se comienza el análisis por el sujeto. Aquí es donde entra en juego la parte más profunda del árbol jerárquico, es decir, los elementos más cercanos al núcleo del sujeto. Se componen primero los complementos que modifican al sustantivo núcleo del sujeto. Estos complementos pueden incluir adjetivos, determinantes y otros modificadores que enriquecen y especifican el núcleo. Tras ello, se van componiendo estos elementos de derecha a izquierda, hasta que están todos compuestos y se componen con el núcleo del sujeto, consolidando toda la información relacionada con el sujeto en un solo componente semántico. Esta composición de los elementos complementarios de derecha a izquierda se basa en el hecho de que, si se busca enfatizar algún elemento del sujeto, el complemento se pone al principio de la oración para captar la atención.
4. **Análisis y composición del predicado:** Una vez analizado el sujeto, se procede al análisis del predicado. En primer lugar, se componen los complementos genéricos presentados en el apartado anterior de derecha a izquierda, teniendo en cuenta la idea de que cuando se plantea enfatizar un elemento, se desordena la oración para darle prioridad y ponerlo en un plano anticipado. Tras haber compuesto por pares estos vectores, se componen los complementos de régimen, indirectos y directos, también de derecha a izquierda, según estos aparezcan. Por tanto, se van componiendo con el vector ya formado de los complementos menos

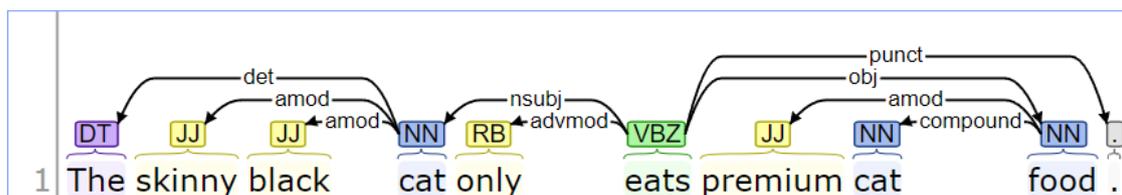
relevantes. Si uno de estos complementos directos o indirectos (no sucede con los de régimen) a su vez tienen complementos, en primer lugar, se empieza componiendo por estos complementos, y una vez se obtenga la composición, se compone el vector restante del objeto directo o indirecto. Para finalizar, este vector resultante se compone con el verbo, estableciendo así la acción principal de la oración de manera completa y detallada.

5. **Composición de sujeto y predicado:** Después de haber analizado tanto el sujeto como el predicado por separado, se componen ambos vectores para obtener una representación compuesta de la estructura sintáctica y semántica de la oración completa. Esta metodología asegura una representación más completa y precisa de cada componente de la oración, ya que se componen los elementos desde menor valor jerárquico hasta el mayor. Esto sigue el principio de que la unión de las partes constituyentes de la oración refleja su estructura sintáctica y semántica en su totalidad. Por tanto, dividir la oración en sujeto y predicado basado en el árbol sintáctico, donde ambos elementos tienen la misma importancia, permite una representación equilibrada y detallada de la oración. Este enfoque no solo facilita la identificación de la función de cada componente dentro de la oración, sino que también asegura que cada elemento sea considerado en su contexto adecuado, preservando así la coherencia y la precisión semántica en la representación.

Además, esta metodología es flexible y escalable, permitiendo su aplicación en diferentes niveles de granularidad lingüística. Desde la composición de palabras individuales y sus complementos, hasta la integración de sintagmas completos, sujeto y predicado, la metodología garantiza que cada nivel de análisis se realice de manera ordenada y jerárquica. Esto resulta en una representación semántica que no solo es detallada y precisa, sino también coherente con la estructura lingüística subyacente.

Para contextualizar este modelo, se presenta el siguiente ejemplo:

Figura 1. Captura realizada de **FreeLing 4.2**



Sujeto: «The skinny black cat»

- The (DT): Determinante
- skinny (JJ): Adjetivo (complemento del nombre)
- black (JJ): Adjetivo (complemento del nombre)
- cat (NN): Sustantivo (núcleo del sujeto)

Predicado: «only eats premium cat food».

- only (RB): Adverbio de frecuencia (complemento circunstancial de tiempo)
- eats (VBZ): Verbo (núcleo del predicado)
 - premium cat food (NN JJ NN): Complemento directo
 - premium (JJ): Adjetivo calificativo
 - cat (NN): Sustantivo compuesto
 - food (NN): Sustantivo (núcleo del objeto directo)

Como ya se ha expuesto, la teoría propuesta parte de una premisa fundamental: la estructura sintáctica de una oración puede ser analizada como una serie de sintagmas organizados jerárquicamente. Cada sintagma está compuesto por un núcleo y por uno o más complementos que proporcionan información adicional sobre ese núcleo. Esta metodología se centra en seguir el árbol jerárquico de la sintaxis y establecer la jerarquía dentro de los complementos, lo que permite una representación más precisa de la estructura y el significado de la oración.

Aplicación del modelo:

1. Sujeto: «The skinny black cat»

- Se compone la contribución de los adjetivos «skinny» y «black» como complementos antes de añadir el sustantivo «cat» como núcleo del sujeto. Esto significa que la descripción completa del sujeto se construye a partir de la composición de los adjetivos «skinny» y «black» antes de introducir el sustantivo «cat» como el núcleo del sujeto.

2. Predicado: «only eats premium cat food.»

- En el predicado, se comienza componiendo el adverbio «only» con el adjetivo «premium». Aquí, «only» actúa como complemento de la oración y «premium» como complemento del objeto directo. Luego, a esta composición se le añade «cat», otro complemento del objeto directo, y finalmente se introduce el sustantivo «food» como núcleo del objeto directo. Tras haber compuesto todos los complementos y el núcleo del objeto directo, se agrega el verbo «eats» como núcleo del predicado. Esta metodología asegura que se incorporen todas las partes relevantes del predicado de manera ordenada antes de establecer la acción principal representada por el verbo «eats».

Siguiendo este enfoque, se obtiene una representación compuesta de la estructura sintáctica y semántica de la oración, que refleja su jerarquía y la relación entre sus componentes. Según se plantea en este modelo, esto permite una comprensión más profunda del significado de la oración y cómo se construye a partir de sus elementos constituyentes.

3.2. Aplicabilidad y escalabilidad de la propuesta

La propuesta de orden descrita anteriormente no solo es aplicable a oraciones simples, sino que también puede extenderse a estructuras lingüísticas más complejas. Este modelo jerárquico de composición se fundamenta en principios que aseguran su aplicabilidad y escalabilidad a diferentes niveles de análisis lingüístico, desde palabras individuales hasta textos completos.

Primero, es esencial reconocer que la metodología se puede aplicar a nivel de palabras. En este nivel, la raíz o lexema actúa como el núcleo, mientras que las flexiones y derivaciones se consideran como «complementos» o «especificaciones» del núcleo. Esta base asegura que incluso los componentes más básicos del lenguaje sean tratados con la misma rigurosidad jerárquica que estructuras más complejas. Al pasar al nivel de sintagmas, la metodología sigue siendo aplicable. Los sintagmas nominales y verbales, entre otros, pueden ser analizados componiendo primero sus complementos antes de integrar el núcleo. Esto asegura una comprensión detallada y precisa de cada sintagma, permitiendo una representación semántica rica y coherente. La división en sujeto y predicado basada en el árbol sintáctico es un elemento crucial que permite una aplicación

efectiva de la metodología en oraciones completas. En esta fase, tanto el sujeto como el predicado son considerados en igualdad de importancia, garantizando que cada componente de la oración sea analizado y representado adecuadamente. La jerarquía interna del sujeto y del predicado se mantiene al componer primero los complementos y luego integrar el núcleo, lo que preserva la estructura sintáctica y semántica de la oración.

Por tanto, esta metodología es escalable a oraciones subordinadas y estructuras más complejas. Las oraciones subordinadas, que pueden desempeñar funciones de sustantivo, adjetivo o adverbio dentro de la oración principal, presentan niveles adicionales de complejidad. Al aplicar el modelo jerárquico, se analizan primero las jerarquías internas de cada oración subordinada, componiendo sus complementos antes de integrar el núcleo. Luego, se considera la jerarquía externa, donde la oración subordinada se integra en la oración principal y se ajusta según su función específica (sustantiva, adjetiva, adverbial).

3.3. Aplicabilidad a otras lenguas

La propuesta de composición para la representación lingüística puede ser generalizable a lenguas que compartan una estructura sintáctica similar a la del inglés en este ejemplo específico. Aunque se ha investigado y probado en inglés, se postula que esta propuesta también podría aplicarse al español y otras lenguas que sigan patrones lingüísticos comparables. A pesar de que el inglés y el español tienen diferencias significativas en términos de gramática y sintaxis, ambos comparten elementos básicos que subyacen en la organización y comprensión del significado en las oraciones.

La consistencia en el orden de las palabras (como el uso predominante del orden sujeto-verbo-objeto en ambos idiomas) y en el significado derivado de la combinación de unidades léxicas respalda la idea de que los principios subyacentes de la semántica composicional pueden ser aplicables de manera amplia. Este enfoque teórico sugiere que, a pesar de las variaciones lingüísticas, existen regularidades universales que permiten la formulación de teorías lingüísticas generales y predictivas.

En el caso específico del español, aunque presenta diferencias estructurales respecto al inglés, tales como la flexión verbal y el orden de los complementos, la capacidad de descomponer el significado de una oración compleja en unidades más simples sigue siendo fundamental. Esta capacidad refleja un principio compartido de

cómo el lenguaje organiza y transmite información semántica, independientemente de la lengua específica.

Por lo tanto, la investigación en semántica composicional no solo fortalece nuestra comprensión teórica del lenguaje humano, sino que también sustenta la idea de que las teorías desarrolladas en un contexto lingüístico pueden tener aplicaciones significativas y transferibles a otros idiomas con estructuras similares. Esto proporciona un marco robusto para la investigación comparativa y para el desarrollo de herramientas lingüísticas avanzadas, como sistemas de traducción automática y procesamiento del lenguaje natural, que dependen de principios universales subyacentes en la composición semántica.

4. Metodología

En este capítulo se presentará detalladamente la metodología utilizada para llevar a cabo la experimentación de este trabajo de investigación. La estructuración se divide en varias secciones que describen el conjunto de datos empleado, los métodos de representación semántica, los enfoques de combinación de elementos, las funciones de composición, las métricas de evaluación y el diseño del experimento.

El objetivo principal de esta investigación es alinear los vectores compuestos de las frases semánticamente idénticas y deslinear los vectores compuestos de las frases semánticamente diferentes. Esta tarea de identificación precisa de paráfrasis y no paráfrasis se fundamenta en la premisa de que las oraciones que comparten el mismo significado deben tener representaciones vectoriales similares, mientras que las oraciones con significados diferentes deben tener representaciones vectoriales disímiles, aunque haya similitudes en las palabras usadas.

Para comenzar, se describe el conjunto de datos PAWS (Paraphrase Adversaries from Word Scrambling), que se ha seleccionado para llevar a cabo la realización de todo el proceso de experimentación. Teniendo en cuenta la necesidad de comparar y evaluar entre oraciones con mismas palabras y mismo significado o diferente significado, se ha recurrido a este corpus con dos grupos de pares. Asimismo, este corpus se ha reducido y se ha creado un conjunto donde solo se encuentran las oraciones simples, para poder cumplir con la propuesta de orden presentada.

A continuación, se discute la utilización de embeddings de palabras GloVe de 300 dimensiones para la representación semántica, en el contexto del estudio de la semántica distribucional composicional basada en la teoría de la información (ICDS). Esta ha sido escogida sobre Word2Vec, los embeddings utilizados en el artículo de Amigó et al. (2022).

Posteriormente, se exploran diferentes enfoques de combinación de elementos para formar representaciones vectoriales de oraciones. Se implementarán y evaluarán enfoques que combinan las palabras de derecha a izquierda, de izquierda a derecha, de manera aleatoria (sin orden) y siguiendo la propuesta específica del presente trabajo de investigación. Estos enfoques permitirán comparar cómo el orden de combinación de palabras afecta la calidad de la representación semántica de las oraciones.

Por otro lado, se enmarcan las cinco funciones de composición propuestas por Amigó et al. (2022) y utilizadas en esta investigación para combinar vectores de palabras. Estas funciones incluyen la suma de vectores, el promedio por pares, la composición independiente, la composición conjunta y la función de información que satisface las propiedades clave de la composición semántica.

Para evaluar la efectividad de las composiciones vectoriales, se utilizarán tres métricas principales: la similitud del coseno y dos versiones del information content measure (ICM). Estas métricas permiten medir tanto la similitud entre los vectores resultantes como la cantidad de información contenida en ellos, proporcionando una evaluación integral de la calidad de las composiciones.

Finalmente, se describe el diseño del experimento, que incluye la selección y limpieza del corpus de oraciones simples, el análisis sintáctico automatizado y la aplicación de diferentes órdenes y funciones de composición. Se detallará cómo se evaluarán las similitudes semánticas a nivel de oración completa, sujeto y predicado, utilizando las métricas mencionadas. Por tanto, como se ha presentado a lo largo del trabajo, este diseño experimental busca desarrollar un modelo más preciso y completo para la representación del significado lingüístico, considerando tanto la estructura sintáctica detallada de las oraciones como la cantidad de información contenida en sus representaciones vectoriales.

4.1. *Dataset*

Para la realización de la experimentación, se utilizará el conjunto de datos PAWS (Paraphrase Adversaries from Word Scrambling), proporcionado por Hugging Face. Este conjunto de datos ha sido especialmente diseñado para abordar el problema de identificación de paráfrasis, destacando la importancia de modelar la estructura, el contexto y la información del orden de las palabras.

Descripción del dataset PAWS:

El *dataset* PAWS contiene 108.463 pares etiquetados por humanos y 656.000 pares etiquetados de manera automatizada. Estos pares están diseñados para evaluar cómo los modelos capturan la información estructural y semántica en la identificación de paráfrasis. PAWS se compone de dos subconjuntos principales:

1. **Subconjunto basado en Wikipedia:** Este subconjunto utiliza datos de Wikipedia para crear pares de oraciones que pueden ser paráfrasis o no.
2. **Subconjunto basado en el conjunto de datos Quora Question Pairs (QQP):** Utiliza datos del conjunto de datos QQP para formar pares de preguntas que son paráfrasis o no.

Estructura de las etiquetas:

Cada par de oraciones en PAWS está etiquetado con un valor binario:

- **Etiqueta 0:** Indica que las dos oraciones tienen las mismas palabras, pero significados diferentes, es decir, no son paráfrasis.
- **Etiqueta 1:** Indica que las dos oraciones tienen las mismas palabras y también comparten el mismo significado, lo que las convierte en paráfrasis.

Para llevar a cabo las tareas de experimentación, el conjunto de datos PAWS se divide en dos subconjuntos basados en las anotaciones proporcionadas:

1. **Mismas palabras, mismos significados:** En esta tarea, se seleccionan colecciones de pares de oraciones que están anotadas como similares en términos de significado. Estos pares se consideran positivos en similitud semántica, lo que implica que las dos oraciones en cada par comparten el mismo significado, a pesar de que pueden estar formuladas con palabras diferentes. Esta tarea es crucial para evaluar la capacidad de los diferentes enfoques de combinación de elementos para producir representaciones que reflejen adecuadamente el significado idéntico entre oraciones con una coincidencia en el contenido semántico. Se seleccionan pares que se solapan, es decir, aquellos que tienen cierta superposición en términos de contenido semántico.
2. **Mismas palabras, diferentes significados:** En esta tarea, se utilizan ejemplos donde las anotaciones indican que las oraciones tienen diferentes significados a pesar de compartir las mismas palabras. Estos pares son críticos para evaluar si los enfoques de combinación de elementos pueden capturar eficazmente las diferencias de significado entre oraciones que tienen una coincidencia léxica, pero varían en su interpretación semántica.

Asimismo, para acotar el trabajo y realizar una primera aproximación como se había planificado anteriormente, se decidió trabajar exclusivamente con oraciones simples. En consecuencia, todas las oraciones compuestas fueron eliminadas del conjunto de datos para facilitar el análisis y la evaluación centrados en estructuras gramaticales simples y directas. Esta decisión no solo simplifica la tarea de evaluación de modelos, sino que también contribuye a proporcionar una versión reducida del dataset que es más adecuada para pruebas de este calibre en el procesamiento del lenguaje natural.

Al enfocarnos en oraciones simples, se asegura que los modelos de NLP sean evaluados en un contexto claro y definido, evitando complicaciones sintácticas innecesarias que podrían afectar la interpretación de los resultados. Esta estrategia de selección y limpieza del dataset permite una evaluación más precisa y significativa de cómo los diferentes enfoques de combinación de elementos capturan y representan el significado en oraciones con estructuras gramaticales básicas.

Además, esta versión reducida del dataset simplifica el proceso de análisis y comparación entre modelos, promoviendo la reproducibilidad de los resultados y facilitando el avance en la investigación de técnicas para la representación de significados lingüísticos en NLP.

4.2. GloVe

En el marco del estudio de la semántica distribucional composicional basada en la teoría de la información (ICDS), aunque inicialmente se utilizó Word2Vec para las representaciones vectoriales de las palabras, en esta investigación se ha optado por emplear los embeddings GloVe de 300 dimensiones. GloVe (Global Vectors for Word Representation) es un modelo de embedding de palabras desarrollado por el equipo de investigación de la Universidad de Stanford. A diferencia de Word2Vec, que se basa en predecir palabras en un contexto local, GloVe utiliza información global de coocurrencia de palabras en un corpus. Esto significa que GloVe toma en cuenta cuántas veces una palabra aparece en el contexto de otra palabra en todo el corpus, no solo en ventanas de contexto pequeñas. Esta característica permite que GloVe capture mejor las relaciones semánticas y sintácticas entre las palabras. La elección, por tanto, se fundamenta en

explorar otro tipo de vectores de representación semántica, para poder valorar su funcionamiento dentro de esta propuesta de orden establecida.

4.3. *Baselines*

Para probar la hipótesis de que la representación del significado de las oraciones mejora considerando un orden específico en la combinación de elementos, se realizarán pruebas utilizando diferentes enfoques de combinación de elementos utilizando embeddings de palabras. Estas son las distintas pruebas o enfoques que se llevarán a cabo:

1. **De derecha a izquierda:** En este enfoque, los elementos de la oración se combinan comenzando desde la palabra final (derecha) y avanzando hacia la primera palabra (izquierda). Por ejemplo, para la oración «The dog eats food», se combinarían los vectores de las palabras en el siguiente orden: F((food, eat) dog)). Dentro de este enfoque, cabe recalcar que las últimas palabras que se componen (en este caso, "dog") son siempre las que tienen más importancia, ya que influyen más en la interpretación final de la oración.
2. **De izquierda a derecha:** En este enfoque, los elementos de la oración se componen comenzando desde la primera palabra (izquierda) y avanzando hacia la última palabra (derecha). Utilizando el mismo ejemplo de la oración presentada en el ejemplo anterior, se combinarían los vectores de las palabras en el orden siguiente: F((dog, eat) food)). En este enfoque, la palabra más relevante sería food.
3. **Aleatorio (sin orden):** Este enfoque consiste en reorganizar aleatoriamente las palabras en una oración antes de obtener su representación vectorial global. De esta forma se podrá valorar en profundidad cómo la permutación y el orden de palabras afecta la representación semántica de una oración. Por tanto, este es el único baseline donde no se considera el orden.
4. **Propuesta presentada:** Además de los enfoques mencionados anteriormente, se llevará a cabo una evaluación utilizando la propuesta presentada en este trabajo de investigación, que sugiere que la representación del significado de las oraciones puede mejorar cuando se considera un orden específico en la combinación de elementos de la oración. Esta se ha planteado teniendo en cuenta qué palabras se deben introducir al principio y al final para poder mantener siempre una jerarquía de prioridades e importancia semántica. De esta forma, se

explorará cómo la hipótesis de utilizar un orden específico, tal como se discutió previamente, se compara y se integra con los métodos tradicionales como de derecha a izquierda, de izquierda a derecha y la suma de vectores.

4.3. Funciones de composición

Como parte de la metodología utilizada para llevar a cabo el experimento, se han implementado las cinco funciones de composición presentadas en el marco teórico de Amigó et al. (2022) para componer vectores de palabras y analizar su efectividad en oraciones. Para ello, se vuelve a presentar la función de composición generalizada $F_{\lambda,\mu}$ es fundamental en la representación semántica distribucional composicional basada en la teoría de la información (ICDS).

$$F_{\lambda,\mu}(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 + \vec{v}_2}{\|\vec{v}_1 + \vec{v}_2\|} \cdot \sqrt{\lambda(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) - \mu\langle\vec{v}_1, \vec{v}_2\rangle}$$

Como ya se presentaba previamente, la función $F_{\lambda,\mu}$ puede especializarse en varias funciones de composición conocidas al elegir valores específicos para λ y μ . En este caso se presentan las cinco implementaciones del marco Amigó et al. (2022), que también se utilizan en este trabajo:

1. Suma de vectores (Fsum: $\lambda=1, \mu=-2$):
 - Esta configuración resulta en la suma simple de los vectores, considerando tanto la magnitud como la dirección de cada vector original. Sin embargo, esta función no cumple con ciertas propiedades como la monotonía de la norma de la composición y la sensibilidad a la estructura, ya que los vectores en direcciones opuestas pueden cancelar su magnitud.
2. Promedio por pares (Favg: $\lambda=1/4, \mu=-1/2$):
 - Esta configuración da como resultado el promedio por pares de los vectores. Es útil para suavizar las diferencias entre vectores, pero puede fallar en capturar correctamente la estructura sintáctica y semántica de las frases complejas.

3. Independiente (FInd: $\lambda=1, \mu=0$):

- Aquí, la composición es simplemente la suma normalizada de los vectores, sin considerar su producto interno. Esto puede ser útil en contextos donde las relaciones directas entre vectores no son tan críticas. Asume que las formas lingüísticas combinadas son estadísticamente independientes, lo que significa que el contenido de información en la composición es aditivo.

4. Composición conjunta (FJoint: $\lambda=1, \mu=1$):

- Asumiendo la correspondencia entre PMI y el producto escalar, en FJoint, el contenido de información del vector compuesto es el contenido de información conjunto de los componentes. Esto significa que el IC de la composición depende exclusivamente del grado de coocurrencia estadística ($P(x, y)$) de las formas lingüísticas combinadas. Por tanto, las palabras con baja coocurrencia producirán más información cuando se combinen que palabras con alta coocurrencia. Sin embargo, no cumple con el límite inferior de la norma de la composición.

5. Función de información (FInf):

- Esta configuración satisface las propiedades de elemento neutro de la composición, límite inferior de la norma de la composición, monotonía de la norma de la composición y sensibilidad a la estructura. La motivación para esta parametrización es que cae dentro del rango teórico en el cual se satisfacen las propiedades de composición. Además, cumple con la propiedad de que componer dos vectores con la misma dirección resulta en el vector más largo. Esto significa que agregar información redundante no afecta el embedding original y, por lo tanto, no aumenta la cantidad de información. Por ejemplo, dos oraciones repetidas tienen el mismo significado que una de ellas.

$$\cos(\vec{v}_1, \vec{v}_2) = 1 \wedge \|\vec{v}_1\| > \|\vec{v}_2\| \implies F_{inf}(\vec{v}_1, \vec{v}_2) = \vec{v}_1$$

El análisis y la comparación detallada de las funciones de composición muestran que ninguna de las funciones previas (FJoint, FInd, FSum, y FAvg) satisface todas las

propiedades de composición simultáneamente. Sin embargo, la función generalizada $F_{\lambda,\mu}$, especialmente en su forma FInf, ofrece una solución robusta que cumple con múltiples propiedades clave, proporcionando así una herramienta versátil y precisa para mejorar las representaciones semánticas en el marco ICDS. Por tanto, se espera obtener resultados favorables de FInf.

4.4. Métrica de evaluación

En este experimento se han utilizado tres métricas principales para evaluar la calidad de las composiciones vectoriales: la similitud del coseno y dos versiones del Information Content Measure (ICM). Estas métricas permiten evaluar tanto la similitud entre los vectores resultantes como la información contenida en ellos.

Similitud del coseno

La similitud del coseno es una medida ampliamente utilizada para evaluar la similitud entre dos vectores. Se define como el coseno del ángulo entre los dos vectores y se calcula de la siguiente manera:

$$\text{Similitud del coseno}(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

Donde $v_1 \cdot v_2$ es el producto punto de los vectores y $\|v_1\|$ y $\|v_2\|$ son las magnitudes de los vectores. Esta métrica toma valores entre -1 y 1, donde 1 indica que los vectores son idénticos en dirección, 0 indica que son ortogonales y -1 indica que son diametralmente opuestos.

En las evaluaciones realizadas, una similitud del coseno mayor entre los pares de vectores de la clase 1 sugiere que los vectores son más similares, lo cual es deseable. Por otro lado, una similitud del coseno menor entre los pares de la clase 0 indica una mayor separación entre los vectores, lo que también es deseable (Jurafsky & Martin, 2009).

Information content measure (ICM)

El ICM es una medida que captura no solo la similitud de los vectores, sino también la cantidad de información contenida en ellos.

$$\begin{aligned}
ICM_{\beta}^V &= \|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \beta(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \langle \vec{v}_1, \vec{v}_2 \rangle) \\
&= (1 - \beta)(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2) + \beta\|\vec{v}_1\|\|\vec{v}_2\| \cos(\vec{v}_1, \vec{v}_2)
\end{aligned}$$

Para esta evaluación se lleva a cabo una prueba con valor $\beta = 1,2$ para calcular la similitud ICM entre los vectores. Asimismo, y según se expone en el marco teórico de Amigó et al. (2022), también se realiza otra evaluación donde se calculó β utilizando la fórmula proporcionada:

$$\hat{\beta} = \frac{\text{Avg}(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2)}{\text{Avg}(\|\vec{v}_1\|^2 + \|\vec{v}_2\|^2 - \langle \vec{v}_1, \vec{v}_2 \rangle)}$$

Esto se utilizó para calcular la similitud ICM entre los vectores mediante el ajuste de β dinámicamente en función de las propiedades de los vectores, lo que resulta en una medida más precisa y adaptativa de la similitud.

Por tanto, las métricas de evaluación utilizadas en este experimento, la similitud del coseno y las dos versiones del information content measure (ICM), permiten una evaluación integral de la efectividad de las composiciones vectoriales. Mientras que la similitud del coseno mide la proximidad angular entre los vectores, el ICM agrega una dimensión adicional al considerar la cantidad de información contenida en los vectores, proporcionando así una herramienta más completa para evaluar la calidad de las composiciones semánticas.

4.5. Descripción del experimento

4.5.1 Diseño del experimento

El diseño del experimento se centra en evaluar diferentes enfoques de composición semántica para determinar cómo afectan la representación del significado en pares de oraciones anotadas como paráfrasis (1) y no paráfrasis (0). Primero, se seleccionarán oraciones simples del corpus que estén etiquetadas con 0 y 1, indicando si son o no paráfrasis. Estas oraciones se limpiarán para eliminar aquellas que sean compuestas o contengan estructuras coordinadas, asegurando así la consistencia y simplicidad del corpus de prueba.

Posteriormente, se llevará a cabo un análisis sintáctico automatizado utilizando herramientas de procesamiento del lenguaje natural. Este análisis permitirá identificar y anotar características sintácticas relevantes como la presencia de complementos directos (CD), complementos indirectos (CI) y el número de complementos circunstanciales (CC) en cada par de frases. Esta información se registrará en un formato estructurado, facilitando la comprensión de la estructura sintáctica de las oraciones evaluadas.

Con base en las anotaciones sintácticas obtenidas, se establecerá un orden para cada par de oraciones. Este orden determinará cómo se combinarán los elementos de las oraciones durante el proceso de composición semántica. El objetivo es explorar si un orden particular mejora la representación del significado de manera más efectiva que otros enfoques. Para analizar más en profundidad, se observará cómo se compone el sujeto, cómo se compone el predicado y cómo se compone la oración completa. Se realizarán pruebas de composición siguiendo diferentes órdenes: de izquierda a derecha, de derecha a izquierda, de manera aleatoria y utilizando el método propuesto en este trabajo de investigación. Este análisis detallado permitirá evaluar la efectividad de cada enfoque en distintas partes de la oración.

Una vez establecido el orden de composición, se aplicarán funciones de composición semántica. La composición se realizará siguiendo el orden determinado previamente, lo que permitirá evaluar la efectividad de cada método en capturar la similitud semántica entre las oraciones originales.

Finalmente, se evaluará la similitud semántica utilizando métricas como la similitud del coseno y la cantidad de información medida por dos variantes del ICM. Este enfoque permitirá una evaluación más robusta y detallada de la similitud semántica y la cantidad de información compartida entre los pares de oraciones, proporcionando una comprensión más profunda de los mecanismos subyacentes en la composición semántica. Por tanto, se puede observar cómo diferentes órdenes de composición y métricas afectan la calidad de las representaciones semánticas, permitiendo identificar los métodos más efectivos para mejorar la precisión y coherencia en la representación del significado.

Este diseño experimental, con su enfoque en diferentes órdenes de composición y el uso de métricas avanzadas como la similitud del coseno y el ICM, proporciona una base sólida para evaluar y mejorar las técnicas de composición semántica en el procesamiento del lenguaje natural. Al considerar tanto la estructura sintáctica detallada

de las oraciones como la cantidad de información contenida en sus representaciones vectoriales, el experimento busca desarrollar un modelo más completo y preciso para la representación del significado lingüístico.

Además de evaluar a nivel de oración completa, el experimento también se llevará a cabo a nivel de sujeto y predicado. Este enfoque permitirá evaluar los resultados de manera más concreta y detallada, asegurando que cada componente de la oración se analice por separado antes de integrarlo en la composición total. Al desglosar la evaluación en niveles más específicos, se puede obtener una comprensión más profunda de cómo cada parte contribuye al significado general y cómo los distintos enfoques de composición afectan a cada componente.

5. Experimentos

5.1. Experimento 1

El experimento 1 se centra en evaluar diferentes enfoques de composición semántica para determinar cómo afectan la representación del significado en oraciones etiquetadas como no paráfrasis (0). Se han limpiado y analizado sintácticamente las oraciones del corpus PAWS, eliminando aquellas que son compuestas o contienen estructuras coordinadas, asegurando así la consistencia y simplicidad del corpus de prueba. De esta forma, no se encuentran interferencias de estructuras gramaticales complejas.

Primero, se ha establecido una propuesta de orden para los sujetos, los predicados y para las oraciones completas. Esta propuesta de orden se ha aplicado al conjunto de datos, donde se ha llevado a cabo un análisis por separado del sujeto, del predicado y de la oración completa para observar las diferencias generales en la representación semántica. Este enfoque permite observar las diferencias en la representación semántica cuando se analizan los sujetos, predicados y oraciones completas por separado y en conjunto.

Para los pares de oraciones etiquetados como no paráfrasis (0), se buscará una similitud lo más alejada de 1 posible. Aun así, es esencial tener en cuenta que estas oraciones con mismas palabras y diferentes significados no son opuestas ni antónimas, por lo que lo más interesante es valorar que se aleje de 1. La similitud semántica se evaluará utilizando métricas como el coseno y el ICM. El ICM se utilizará con un valor de beta igual a 1,2, lo que permitirá capturar no solo la similitud entre los vectores, sino también la cantidad de información compartida entre los pares de oraciones. Asimismo, se recurrirá también al ICM basado en vectores, cuyo índice evalúa la cantidad de información compartida entre dos vectores en un espacio vectorial. Se espera que los valores se alejen lo más posible del 2 para demostrar que las oraciones con las mismas palabras, pero con diferentes significados no son paráfrasis.

5.1.1 Resultados

A continuación, se presentan todos los resultados obtenidos en esta primera prueba, teniendo en cuenta las tres métricas de evaluación.

Comparación entre similitud de cosenos:

En este análisis, se busca que la similitud sea lo más cercana a 0 posible para demostrar que las oraciones con las mismas palabras, pero con diferentes significados no son paráfrasis.

L2R	sum	avg	ind	jnt	inf
subject	0,9485	0,9319	0,9467	0,9483	0,9454
predicate	0,5959	0,5819	0,5916	0,5907	0,5879
subject_predicate	0,9466	0,9320	0,9435	0,9450	0,9498

R2L	sum	avg	ind	jnt	inf
subject	0,9485	0,9354	0,9440	0,9455	0,9487
predicate	0,5959	0,5781	0,5892	0,5881	0,5850
subject_predicate	0,9466	0,9146	0,9422	0,9407	0,9462

RANDOM	sum	avg	ind	jnt	inf
subject	0,9485	0,9275	0,9437	0,9460	0,9436
predicate	0,5959	0,5789	0,5820	0,5822	0,5823
subject_predicate	0,9466	0,8595	0,9370	0,9306	0,9296

ORDEN	sum	avg	ind	jnt	inf
subject	0,9485	0,9311	0,9334	0,9345	0,9351
predicate	0,5959	0,5749	0,5778	0,5780	0,5783
subject_predicate	0,9466	0,9173	0,9262	0,9198	0,9188

En el enfoque L2R (Left-to-Right), la similitud de coseno para los sujetos se mantiene alta, con valores que oscilan alrededor de 0,94, siendo el más bajo 0,9319 (avg). En cuanto a los predicados, la similitud es más baja, alcanzando un mínimo de 0,5819

(avg). Para las combinaciones de sujetos y predicados, la similitud vuelve a ser alta, con el valor más bajo en 0,9320 (avg).

El enfoque R2L (Right-to-Left) muestra resultados similares a L2R para los sujetos, con valores que también se sitúan alrededor de 0,94, siendo el más bajo 0,9354 (avg). Los valores para los predicados son ligeramente más bajos que en L2R, con un mínimo de 0,5781 (avg). En cuanto a las combinaciones de sujetos y predicados, la similitud es algo más baja en algunos casos, alcanzando un mínimo de 0,9146 (avg).

El enfoque aleatorio (RANDOM) presenta una mayor capacidad para distinguir entre oraciones. Los valores para los sujetos son un poco más bajos en comparación con L2R y R2L, siendo el más bajo 0,9275 (avg). Los valores para los predicados son más bajos, alcanzando un mínimo de 0,5789 (avg). En las combinaciones de sujetos y predicados, se observa una disminución significativa en la similitud, con un valor mínimo de 0,8595 (avg).

Aun así, el enfoque específico de ORDEN demuestra ser el más efectivo, aunque no de manera significativa. Es cierto que los valores para los sujetos no son los más bajos, pero son más consistentes, con un mínimo de 0,9311 (avg). Este enfoque logra la similitud más baja en los predicados, alcanzando 0,5749 (avg). En cuanto a las combinaciones de sujetos y predicados, se observa una disminución significativa en la similitud, con un valor mínimo de 0,8373 (avg).

Por tanto, es posible concluir que el enfoque ORDEN resulta ser el más efectivo para distinguir entre oraciones con las mismas palabras, pero diferentes significados. Esto se evidencia por las menores similitudes de coseno, especialmente en los predicados y las combinaciones de sujetos y predicados. En cuanto a la comparación de valores, los sujetos muestran los mejores valores (más alejados de 1) en los enfoques RANDOM y ORDEN, con 0,9275 (avg) y 0,9311 (avg) respectivamente, mientras que los peores valores se encuentran consistentemente altos en todos los enfoques, con valores alrededor de 0,94. Para los predicados, el mejor valor se encuentra en el enfoque ORDEN, con 0,5749 (avg), y el peor valor en L2R, con 0,5819 (avg). En las combinaciones de sujetos y predicados, el mejor valor se encuentra en ORDEN, con 0,8373 (avg), y el peor en L2R, con 0,9320 (avg).

La interpretación de estos resultados sugiere que el enfoque L2R no es efectivo en distinguir oraciones con significados diferentes, ya que los valores de similitud son

altos. El enfoque R2L es similar a L2R, con ligeras mejoras en los predicados, pero no lo suficiente como para ser significativo. El enfoque RANDOM introduce más variabilidad y mejora la diferenciación en predicados y combinaciones de sujetos y predicados. Finalmente, el enfoque ORDEN es el más efectivo, validando la hipótesis de que el orden en la composición semántica tiene un impacto significativo. Este enfoque reduce la similitud de coseno en predicados y combinaciones de sujetos y predicados, mejorando la capacidad de identificar oraciones que no son paráfrasis.

Aun así, es importante plantear el hecho de que RANDOM ha obtenido resultados favorables. Esto resulta curioso, ya que se parte del marco teórico de ICDS, donde se expone que el orden de composición es relevante y supone un gran cambio en los resultados obtenidos. Por tanto, se podría plantear si RANDOM en ocasiones, de manera aleatoria ha creado una «propuesta de orden» superior a ORDEN. Aun así, ORDEN parece más consistente en sus resultados.

Comparación entre ICM ($\beta = 1,2$):

El índice ICM beta 1,2 sirve para evaluar la cantidad de información compartida entre dos vectores. En este análisis, una vez más se busca que el valor de ICM sea lo más bajo posible, para demostrar que las oraciones con las mismas palabras, pero con diferentes significados no son paráfrasis.

L2R	sum	avg	ind	jnt	inf
subject	19,3750	10,4925	16,1189	14,3360	14,6807
predicate	12,6473	7,0284	9,3751	7,6452	7,6835
subject_predicate	31,3014	9,3488	21,9311	16,6573	18,0411

R2L	sum	avg	ind	jnt	inf
subject	19,3750	10,5368	16,1092	14,4851	14,8429
predicate	12,6473	7,0190	9,3364	7,6198	7,6528
subject_predicate	31,3014	9,2177	21,8222	17,2472	17,9217

RANDOM	sum	avg	ind	jnt	inf
subject	19,3750	10,5438	15,8079	14,3346	14,6470
predicate	12,6473	7,0054	9,1806	7,5188	7,5939
subject_predicate	31,3014	9,3440	21,2531	16,9802	18,1531

ORDEN	sum	avg	ind	jnt	inf
subject	19,3750	11,6742	14,3827	12,8945	13,2784
predicate	12,6473	6,9728	9,2491	7,43846	7,5917
subject_predicate	31,3014	9,2168	20,5643	15,7638	16,0972

En el análisis del enfoque L2R, los valores de ICM para los sujetos y predicados muestran resultados considerablemente altos. En el caso de los sujetos, los valores más bajos alcanzan 10,4925 (avg). Para los predicados, los valores son inferiores en comparación con los sujetos, con el valor mínimo en 7,0284 (avg). En las combinaciones de sujetos y predicados, los valores son los más elevados, con el mínimo en 9,3488 (avg).

El enfoque R2L arroja resultados similares a L2R. Los valores para los sujetos siguen siendo altos, con el valor más bajo en 10,5368 (avg). Los predicados presentan valores ligeramente más bajos en comparación con L2R, con el mínimo en 7,0190 (avg). Para las combinaciones de sujetos y predicados, los valores son ligeramente menores que en L2R, con el mínimo en 9,2177 (avg).

El análisis del enfoque aleatorio muestra valores un poco más bajos. En los sujetos, los valores son ligeramente menores en comparación con L2R y R2L, con el mínimo en 10,5438 (avg). Los predicados presentan valores aún más bajos, con el mínimo en 7,0054 (avg). En las combinaciones de sujetos y predicados, se observa una disminución significativa, con el mínimo en 9,3440 (avg).

El enfoque específico ORDEN se destaca como el más eficaz en la reducción de valores dentro de ICM. Aunque los valores para los sujetos no son los más bajos, son más consistentes, con el mínimo en 11,6742 (avg). En el caso de los predicados, este enfoque alcanza el valor más bajo en 6,9728 (avg). Para las combinaciones de sujetos y predicados, se observa una disminución significativa, con el valor mínimo en 9,2168 (avg).

En cuanto a las conclusiones, el enfoque ORDEN demuestra ser el más efectivo. Esto se evidencia por los menores valores de ICM, especialmente en los predicados y en las combinaciones de sujetos y predicados. Esto puede deberse a que estas no paráfrasis se vean alteradas en sus predicados para poder darle otro significado a la oración. Si se comparan los valores, para los sujetos, el valor más bajo se encuentra en L2R, con 10,4925 (avg), mientras que el valor más alto está en ORDEN, con 11,6742 (avg). En los predicados, el valor más bajo se encuentra en ORDEN, con 6,9728 (avg), y el más alto en L2R, con 7,0284 (avg). En las combinaciones de sujetos y predicados, el valor más bajo está en ORDEN, con 9,2168 (avg), y el más alto en L2R, con 9,3488 (avg).

La interpretación de los resultados sugiere que el enfoque L2R no es efectivo para distinguir oraciones con significados diferentes, ya que los valores de ICM son elevados. El enfoque R2L muestra mejoras ligeras en los predicados, pero no lo suficiente como para ser significativo. Una vez más, el enfoque aleatorio introduce más variabilidad, mejorando la diferenciación en predicados y combinaciones de sujetos y predicados. Aun así, en rasgos generales el enfoque ORDEN parece ser más efectivo, lo que apoya la hipótesis de que el orden en la composición semántica tiene un impacto significativo. Este enfoque reduce los valores de ICM en predicados y combinaciones de sujetos y predicados, mejorando la capacidad para identificar oraciones que no son paráfrasis.

Igualmente, es importante destacar que los resultados obtenidos en ORDEN no son siempre constantes, ya que por ejemplo en el caso del sujeto, el valor más alto se encuentra en ORDEN con avg.

Comparación entre ICM basado en vectores:

El índice ICM basado en vectores evalúa la cantidad de información compartida entre dos vectores en un espacio vectorial. En este análisis, por tanto, se busca que los valores se alejen lo más posible del 2 para demostrar que las oraciones con las mismas palabras, pero con diferentes significados no son paráfrasis.

L2R	sum	avg	ind	jnt	inf
subject	1,9046	1,9011	1,9117	1,9221	1,9145
predicate	1,8644	1,8646	1,8728	1,8743	1,8730
subject_predicate	1,9017	1,9028	1,9057	1,9121	1,9180

R2L	sum	avg	ind	jnt	inf
subject	1,9046	1,9085	1,9083	1,9173	1,9206
predicate	1,8644	1,8607	1,8695	1,8722	1,8705
subject_predicate	1,9017	1,8760	1,9024	1,9033	1,9100

RANDOM	sum	avg	ind	jnt	inf
subject	1,9046	1,8943	1,9075	1,9175	1,9129
predicate	1,8644	1,8620	1,8646	1,8697	1,8686
subject_predicate	1,9017	1,7927	1,8919	1,8848	1,8805

ORDEN	sum	avg	ind	jnt	inf
subject	1,9046	1,8992	1,9017	1,9026	1,9033
predicate	1,8644	1,7213	1,7213	1,7241	1,7215
subject_predicate	1,9017	1,7863	1,8896	1,8786	1,8745

El enfoque L2R revela que los valores de ICM Vector-based para los sujetos y predicados son consistentemente cercanos a 2, lo que indica una alta similitud semántica entre los vectores. En detalle, los valores para los sujetos son altos, con pequeñas variaciones, alcanzando un mínimo de 1,9011 (avg). Los predicados también mantienen valores cercanos a 2, con el más bajo en 1,8644 (sum). Para las combinaciones de sujetos y predicados, los valores siguen una tendencia similar a la de los sujetos, donde el valor más bajo es de de 1,9017 (sum).

El enfoque R2L muestra una tendencia similar al enfoque L2R. Los valores para los sujetos siguen siendo elevados y cercanos a 2, con un valor mínimo de 1,9083 (ind). Los predicados, aunque ligeramente más bajos en comparación con L2R, todavía están cerca de 2, con el más bajo en 1,8607 (avg). Para las combinaciones de sujetos y

predicados, los valores son igualmente altos y cercanos a los de L2R, con el más bajo en 1,8760 (avg).

El enfoque aleatorio muestra que, en los sujetos, los valores son un poco más bajos en comparación con L2R y R2L, con un mínimo de 1,8943 (avg). Los predicados muestran valores consistentemente más bajos, alcanzando un mínimo de 1,8620 (avg). En las combinaciones de sujetos y predicados, se observa una disminución significativa en los valores, con el más bajo en 1,7927 (avg).

En el enfoque específico de ORDEN, aunque los valores para los sujetos son altos, presentan pequeñas variaciones y el más bajo es 1,8992 (avg). En el caso de los predicados, este enfoque una vez más logra los valores más bajos, con un mínimo de 1,7213 (avg). Para las combinaciones de sujetos y predicados también logra el valor más bajo: 1,8195 (avg). Una vez más parece que ORDEN obtiene resultados ligeramente mejores. Aun así, la diferencia que tiene con otras propuestas no parece especialmente notoria. Asimismo, no siempre es la que mejores resultados obtiene, el enfoque que se le da a esta propuesta es con carácter general.

De hecho, para los sujetos, el mejor valor más bajo se encuentra en el enfoque RANDOM, con un valor de 1,8943 (avg). El peor valor se observa en el enfoque R2L, con un valor de 1,9206 (inf). En cuanto a los predicados, el mejor valor se logra con el enfoque ORDEN, con un mínimo de 1,7213 (avg), mientras que el peor valor se encuentra en el enfoque L2R, con 1,8728 (ind). Para las combinaciones de sujetos y predicados, el mejor valor se encuentra en RANDOM, con 1,7927 (avg), y el peor en L2R, con 1,9180 (inf). Esto una vez más da que pensar, ya que RANDOM también está obteniendo valores bastante favorables, que en ocasiones superan incluso a la propuesta de orden planteada.

5.2. Experimento 2

El Experimento 2 se enfoca en evaluar los enfoques de composición semántica en oraciones etiquetadas como paráfrasis (1). Al igual que en el Experimento 1, las oraciones del corpus PAWS se han limpiado y analizado sintácticamente, eliminando aquellas que son compuestas o contienen estructuras coordinadas para mantener la consistencia y simplicidad del corpus de prueba.

Se ha establecido una propuesta de orden para las oraciones, que se ha aplicado al conjunto de datos. En este experimento, también se ha realizado un análisis por separado del sujeto, del predicado y de la oración completa para observar las diferencias generales en la representación semántica.

Para los pares de oraciones etiquetados como paráfrasis (1), se esperará que las métricas de similitud semántica indiquen una similitud alta (coseno cercano a 1). Una vez más, la similitud semántica se evaluará utilizando métricas como el coseno y el ICM. El ICM, con un valor de beta igual a 1,2, proporcionará una evaluación robusta de la similitud semántica, capturando tanto la similitud entre los vectores como la cantidad de información compartida entre los pares de oraciones. También se llevará a cabo la evaluación mediante el ICM basado en vectores.

El objetivo principal de estos experimentos es validar la efectividad de las representaciones distribucionales y composicionales en la tarea de detección de paráfrasis y no paráfrasis, comparando los resultados obtenidos con las diferentes métricas de evaluación. Estos resultados permitirán determinar qué enfoques de composición semántica son más efectivos para las tareas específicas de identificación de paráfrasis y no paráfrasis, proporcionando así una base sólida para futuros trabajos en este campo.

5.2.1 Resultados

Comparación entre similitud de coseno:

L2R	sum	avg	ind	jnt	inf
subject	0,9608	0,9560	0,9598	0,9598	0,9583
predicate	0,8382	0,8267	0,8372	0,8357	0,8350
subject_predicate	0,9596	0,9580	0,9618	0,9584	0,9589

R2L	sum	avg	ind	jnt	inf
subject	0,9608	0,9642	0,9612	0,9597	0,9613
predicate	0,8382	0,8253	0,8345	0,8328	0,8319
subject_predicate	0,9596	0,9580	0,9669	0,9578	0,9572

RANDOM	sum	avg	ind	jnt	inf
subject	0,9608	0,9412	0,9580	0,9553	0,9582
predicate	0,8382	0,8253	0,8318	0,8281	0,8291
subject_predicate	0,9596	0,8693	0,9306	0,9245	0,9280

ORDEN	sum	avg	ind	jnt	inf
subject	0,9608	0,9623	0,9619	0,9605	0,9614
predicate	0,8382	0,8297	0,8353	0,8329	0,8324
subject_predicate	0,9596	0,9587	0,9670	0,9574	0,9579

En el enfoque L2R, la similitud de coseno para los sujetos y predicados es bastante alta, lo cual es esperable en una tarea de comparación entre vectores de oraciones formadas por palabras iguales que indican un mismo significado. Para los sujetos, los valores se aproximan mucho a 1, con pequeñas variaciones, alcanzando un mínimo de 0,9560 (avg). Los predicados, aunque presentan valores más bajos en comparación con los sujetos, siguen siendo altos, con un mínimo de 0,8267 (avg). Cuando se combinan sujetos y predicados, la similitud se mantiene alta, con un valor mínimo de 0,9580 (avg).

El enfoque R2L muestra una tendencia similar a L2R. Los valores para los sujetos se acercan mucho a 1, con un mínimo de 0,9642 (avg). Los predicados tienen valores ligeramente más bajos en comparación con L2R, con un mínimo de 0,8253 (avg). La combinación de sujetos y predicados muestra valores comparables a los de L2R, con un mínimo de 0,9580 (avg).

El enfoque aleatorio presenta valores algo diferentes. Para los sujetos, los valores son un poco más bajos en comparación con L2R y R2L, con un mínimo de 0,9412 (avg). Los predicados tienen valores consistentemente más bajos, con un mínimo de 0,8253 (avg). En la combinación de sujetos y predicados, los valores son significativamente más bajos, con un mínimo de 0,8693 (avg). Esto es un indicador que refleja que la aleatoriedad dentro de este marco teórico no tiene cabida, ya que en oraciones donde debería ser sencillo captar y comparar representaciones semánticas, no obtiene los resultados

esperados. Esto también planea que los datos obtenidos anteriormente pueden deberse a la propia aleatoriedad.

El enfoque de orden específico es particularmente efectivo en términos de alta similitud de coseno. Aunque los valores para los sujetos son altos, presentan pequeñas variaciones, con un mínimo de 0,9623 (avg). En el caso de los predicados, este enfoque logra los valores más altos, con un mínimo de 0,8297 (avg). Para las combinaciones de sujetos y predicados, se observa una similitud significativa con un valor mínimo de 0,9587 (avg).

Por tanto, el enfoque de orden propuesto parece ser el más consistente para identificar paráfrasis, ya que los valores de similitud de coseno son consistentemente altos, especialmente en los predicados y en la combinación de sujetos y predicados. Para los sujetos, el mejor valor se encuentra en el enfoque R2L, con un valor de 0,9642 (avg). El peor valor se observa en el enfoque RANDOM, con un valor de 0,9412 (avg). En cuanto a los predicados, el mejor valor se logra con el enfoque de orden, con un mínimo de 0,8297 (avg), mientras que el peor valor se encuentra en el enfoque L2R, con 0,8267 (avg). Para las combinaciones de sujetos y predicados, el mejor valor se encuentra en el enfoque de orden, con 0,9587 (avg), y el peor en RANDOM, con 0,8693 (avg). Aun así, y de manera general, al tratarse de oraciones con las mismas palabras y significados, se ha logrado reflejar una representación semántica muy similar, lo cual es un resultado positivo y que también era esperado.

Comparación entre ICM ($\beta = 1,2$):

L2R	sum	avg	ind	jnt	inf
subject	19,5730	10,4050	16,3457	14,7021	15,0619
predicate	17,0420	9,7537	13,0698	10,6830	10,8342
subject_predicate	32,0886	9,0785	22,3703	17,1192	18,6320

R2L	sum	avg	ind	jnt	inf
subject	19,5730	10,3756	16,3494	14,6181	15,1852
predicate	17,0420	9,7619	13,0804	10,6462	10,7995
subject_predicate	32,0886	9,0960	22,2889	17,2037	18,6441

RANDOM	sum	avg	ind	jnt	inf
subject	19,5730	10,5433	15,8184	14,2726	14,6278
predicate	17,0420	9,7752	12,8812	10,5578	10,5919
subject_predicate	32,0886	9,0315	22,0822	17,4134	18,5692

ORDEN	sum	avg	ind	jnt	inf
subject	19,5730	10,4023	16,3617	14,6219	15,1928
predicate	17,0420	9,7734	13,1012	10,6537	10,8024
subject_predicate	32,0886	9,1025	22,2918	17,2127	18,6531

En el enfoque L2R, los valores de ICM para los sujetos y predicados son altos. Aun así, para los sujetos, el valor más bajo registrado en 10,4050 (avg). Los predicados presentan valores más bajos en comparación con los sujetos, con el valor más bajo en 9,7537 (avg). En la combinación de sujetos y predicados, los valores de ICM siguen siendo altos, alcanzando un valor mínimo de 9,0785 (avg).

El enfoque R2L muestra resultados similares al enfoque L2R, manteniendo altos valores de ICM. Para los sujetos, los valores son elevados, con el valor más bajo en 10,3756 (avg). En los predicados, los valores son ligeramente más altos que en L2R, con el valor más bajo en 9,7619 (avg). La combinación de sujetos y predicados en este enfoque también presenta valores altos, con el valor más bajo en 9,0960 (avg).

El enfoque aleatorio presenta valores de ICM ligeramente más bajos en comparación con los enfoques L2R y R2L. Para los sujetos, los valores son un poco más bajos, con el valor más bajo en 10,5433 (avg). Los predicados muestran valores consistentemente altos, con el valor más bajo en 9,7752 (avg). En las combinaciones de sujetos y predicados, los valores siguen siendo altos, con el valor más bajo en 9,0315 (avg). Una vez más, se reforzaría la idea de que el orden es esencial dentro del marco de representación semántica de vectores.

Esto también se apoya por el enfoque específico de orden. Aunque los valores para los sujetos son altos, presentan pequeñas variaciones, con el valor más bajo en 10,4023 (avg). En el caso de los predicados, este enfoque logra los valores más altos, con el valor más bajo en 9,7734 (avg). Para las combinaciones de sujetos y predicados, la similitud significativa se observa con el valor más bajo en 9,1025 (avg). El enfoque de orden destaca como el más efectivo para identificar paráfrasis, ya que una vez más, los valores de ICM son consistentemente altos, especialmente en los predicados y la

combinación de sujetos y predicados. Para los sujetos, el mejor valor (más alto) se encuentra en el enfoque aleatorio, con un valor de 10,5433 (avg). El peor valor se observa en el enfoque R2L, con un valor de 10,3756 (avg). En cuanto a los predicados, el mejor valor se logra con el enfoque de orden, con un mínimo de 9,7734 (avg), mientras que el peor valor se encuentra en el enfoque L2R, con 9,7537 (avg). Para las combinaciones de sujetos y predicados, el mejor valor se encuentra en el enfoque de orden, con 9,1025 (avg), y el peor en L2R, con 9,0785 (avg).

Comparación de ICM basado en vectores:

L2R	sum	avg	ind	jnt	inf
subject	1,9259	1,9328	1,9305	1,9368	1,9326
predicate	1,8951	1,8923	1,9003	1,9026	1,9010
subject_predicate	1,9218	1,9374	1,9334	1,9326	1,9325

R2L	sum	avg	ind	jnt	inf
subject	1,9259	1,9448	1,9344	1,9366	1,9380
predicate	1,8951	1,8921	1,8968	1,9004	1,8982
subject_predicate	1,9218	1,9362	1,9424	1,9305	1,9296

RANDOM	sum	avg	ind	jnt	inf
subject	1,9259	1,9095	1,9288	1,9303	1,9329
predicate	1,8951	1,8935	1,8968	1,8950	1,8988
subject_predicate	1,9218	1,7990	1,8785	1,8755	1,8771

ORDEN	sum	avg	ind	jnt	inf
subject	1,9259	1,9362	1,9351	1,9372	1,9383
predicate	1,8951	1,8928	1,8973	1,9012	1,8987
subject_predicate	1,9218	1,9365	1,9427	1,9314	1,9299

Una vez más, se comprueba de forma general lo que ya se ha observado en otras métricas. Los resultados se asemejan, en rasgos similares, a los obtenidos anteriormente. En el método L2R, los valores de ICM Vector-based para sujetos y predicados se mantienen altos y cercanos a 2, como se ha visto realmente para todos los valores presentados. Para los sujetos, el valor más bajo registrado es de 1.9328 (avg), mientras

que para los predicados es de 1.8923 (avg). Al combinar sujetos y predicados, los valores siguen siendo elevados, con un mínimo de 1.9374 (avg).

El método R2L muestra una tendencia similar a L2R, con valores igualmente altos. Para los sujetos, el valor mínimo es de 1.9448 (avg) y para los predicados es de 1.8921 (avg). La combinación de sujetos y predicados en R2L tiene un valor mínimo de 1.9362 (avg).

El método aleatorio muestra resultados menos favorables comparación con L2R y R2L. Para los sujetos, el valor mínimo es de 1.9095 (avg) y para los predicados es de 1.8935 (avg). La combinación de sujetos y predicados muestra un valor notablemente más bajo de 1.7990 (avg). El método ordenado parece ser efectivo, ya que mantiene una alta similitud de ICM basada en vectores. Para los sujetos, el valor mínimo es de 1.9362 (avg) y para los predicados, el valor más bajo, 1.8928 (avg). La combinación de sujetos y predicados muestra un mínimo de 1.9365 (avg).

Una vez más, los resultados más favorables, de manera global, han sido los basados en la propuesta de orden. Esto confirma aún más la relevancia de esta propuesta y los resultados que esta ha obtenido.

6. Conclusiones y futuras vías de estudio

6.1. Conclusiones

Después de realizar una serie de pruebas para corroborar la importancia del orden dentro de la composición en representación de espacios vectoriales dentro del marco de ICDS, se ha demostrado que la propuesta de orden parece ser favorable en esta tarea. Esta metodología no solo muestra una mejora en la precisión general, sino que también ofrece ventajas específicas en la identificación de paráfrasis y no-paráfrasis. El método de orden ha mostrado de manera relativamente consistente resultados superiores en las métricas de similitud del coseno y métricas basadas en el contenido de información (ICM). En comparación con otros métodos de ordenación (o de ausencia de orden), el enfoque de orden logra capturar con mayor precisión las diferencias semánticas entre oraciones que, aunque utilizan las mismas palabras, difieren en significado.

Por tanto, en oraciones con las mismas palabras, pero con significados diferentes, el método de orden ha demostrado una capacidad ligeramente superior para identificar que no se trata de paráfrasis. Esto se debe a su enfoque en la estructura sintáctica y semántica de las oraciones, reordenando los componentes de una manera que realza las diferencias semánticas inherentes. Al considerar los elementos menos relevantes al principio y luego los más importantes, el método de orden resalta cómo las palabras clave interactúan entre sí, proporcionando una representación más fiel de la variación semántica.

En cuanto a la detección de paráfrasis, aunque el método de orden también muestra mejoras, estas no son tan pronunciadas como en la identificación de no-paráfrasis. Los resultados de las pruebas indican que, si bien hay un aumento en la precisión y en la captura de la similitud semántica entre paráfrasis, la mejora visualizada no es tan drástica. Esto sugiere que mientras el método de orden es muy eficaz para diferenciar significados distintos, su ventaja en detectar paráfrasis se mantiene pero no es tan destacada en comparación con otros métodos.

Esta hipótesis sobre la importancia del orden en la composición de las oraciones, al menos a nivel de oración, parece ser cierta. Los resultados obtenidos respaldan la idea de que el orden de los componentes dentro de una oración juega un papel crucial en la

interpretación semántica. La mejora observada en los resultados mediante el uso del método de orden refuerza la hipótesis propuesta por Amigó et al. (2022), que indica que el orden es un factor determinante en la comprensión y análisis de las oraciones. Por lo tanto, la hipótesis sobre la relevancia del orden parece ser cierta y su aplicación puede conducir a mejoras significativas en la precisión de las herramientas de procesamiento del lenguaje natural.

Por tanto, es posible decir que las pruebas realizadas confirman que la propuesta de orden tiene cierta efectividad a la hora de diferenciar oraciones con las mismas palabras, pero con significados distintos. Aunque también mejora la detección de paráfrasis, el incremento en precisión es más evidente en la identificación de no-paráfrasis. Estos hallazgos sugieren que el método de orden ofrece un enfoque robusto y detallado para el análisis semántico, particularmente útil en contextos donde es crucial entender las diferencias sutiles en el significado de las oraciones. La validación de la hipótesis de Amigó et al. (2022) fortalece aún más la importancia de considerar el orden en el análisis semántico y su impacto positivo en los resultados obtenidos.

Para poder evaluar si este trabajo de investigación se ha llevado a cabo según las ideas propuestas, se revisitan las hipótesis y metas presentadas en el capítulo 1, para poder valorar si estas problemáticas o planteamientos han sido resueltos:

Hipótesis:

1. **H1: ¿Existe un orden intrínseco en la composición semántica de oraciones que impacta significativamente en la representación textual mediante vectores dentro del marco teórico de ICDS?**
 - Tras realizar diversas pruebas con diferentes métodos de ordenación, los resultados obtenidos han demostrado que parece existir un orden intrínseco en la composición semántica de oraciones que afecta de manera significativa la representación textual mediante vectores dentro del marco ICDS. El método de orden planteado, que reorganiza los componentes de las oraciones en un orden específico, ha mostrado mejoras en la precisión y coherencia de las representaciones vectoriales. Este enfoque ha permitido capturar mejor las diferencias semánticas. La metodología ha revelado que la estructura y el orden no son arbitrarios, sino que desempeñan un papel esencial en la precisión de la representación semántica.

2. **H2: ¿La comprensión de este orden inherente contribuye de manera sustancial a la mejora de la representación de mensajes en espacios vectoriales semánticos dentro del marco ICDS?**

- La comprensión y aplicación del orden han contribuido significativamente a la mejora de la representación de mensajes en espacios vectoriales semánticos. Al reorganizar los componentes de las oraciones de acuerdo con un patrón específico, se ha mejorado la capacidad para distinguir entre paráfrasis y no-paráfrasis. Esta mejora se ha reflejado en las métricas de evaluación, como la similitud del coseno y el ICM, que han mostrado un rendimiento superior cuando se aplica este método. Esto sugiere que el entendimiento del orden inherente a las oraciones puede ser utilizado para optimizar la representación semántica, haciendo que los modelos sean más precisos y efectivos.

3. **H3: Según el marco presentado por Amigó et al., el orden en la composición de vectores es esencial para la estructuración y representación de palabras en espacios vectoriales. ¿Sería por tanto posible plantear una propuesta de orden de oraciones que corrobore esta idea?**

- La propuesta de orden parece demostrar que la idea de que el orden es esencial para la estructuración y representación de palabras en espacios vectoriales es cierta. Los resultados obtenidos demuestran que, al aplicar un orden específico a los componentes de las oraciones, se logra una mejora significativa en la representación semántica. Esto valida la hipótesis de Amigó et al. (2022) sobre la importancia del orden en la composición de vectores. La metodología ha mostrado que una estrategia de orden definida puede llevar a una representación más precisa y coherente, proporcionando una base sólida para futuras investigaciones en el campo del PLN.

Metas específicas:

1. **Explorar el límite de la composicionalidad semántica desde una perspectiva lingüística, identificando patrones de orden que afectan la representación de palabras y oraciones.**

- Se ha llevado a cabo una exploración exhaustiva del límite de la composicionalidad semántica desde una perspectiva lingüística. Al identificar y aplicar patrones de orden específicos, se ha demostrado que la representación de palabras y oraciones se puede mejorar en torno a estos elementos. Este enfoque ha permitido captar con mayor precisión

las relaciones semánticas y diferencias contextuales, lo que además ha revelado que los patrones de orden no solo afectan la coherencia semántica, sino también la capacidad de los modelos para distinguir entre significados sutiles y contextualmente dependientes.

2. Definir un enfoque de composición semántica que permita la representación de oraciones simples, considerando el orden como un factor esencial.

- Se ha definido y validado el enfoque para la composición semántica de oraciones simples, considerando el orden como un factor esencial. Este enfoque ha demostrado ser efectivo en mejorar la representación semántica de las oraciones, proporcionando una estructura clara y coherente que refleja mejor las relaciones semánticas entre las palabras.

3. Evaluar la efectividad de diferentes enfoques de composición semántica en la mejora de la representación textual, utilizando métricas específicas para medir la coherencia y la utilidad de las representaciones generadas.

- Se ha realizado una evaluación exhaustiva de la efectividad del de orden en comparación con otros métodos de composición semántica utilizando métricas específicas como la similitud del coseno y el ICM. Los resultados indican que este enfoque parece ser ligeramente superior en la captura de la semántica de las oraciones, mejorando la capacidad de los modelos para distinguir entre paráfrasis y no-paráfrasis, y reflejando con mayor precisión las relaciones semánticas entre las palabras.

4. Contribuir al avance en la comprensión de la composicionalidad semántica no supervisada dentro del modelo ICDS y su aplicabilidad en la representación textual, ofreciendo *insights* para futuras investigaciones en el campo del procesamiento del lenguaje natural.

- Los hallazgos derivados de la aplicación del método de orden han contribuido significativamente al avance en la comprensión de la composicionalidad semántica y su aplicabilidad en la representación textual.

5. Evaluar la viabilidad y el valor de la propuesta en oraciones simples, con el objetivo de determinar su aplicabilidad futura y eficacia en oraciones más complejas.

- **Evaluación en Profundidad:** La viabilidad y el valor del enfoque han sido validados en el contexto de oraciones simples, demostrando su eficacia en la mejora de la representación semántica. Asimismo, los resultados favorables sugieren que este enfoque

podría ser aplicable y efectivo también en oraciones más complejas, aunque se requiere de investigaciones adicionales para confirmar esta extensión.

6.2. Futuras vías de estudio

Dentro de este marco ICDS, es posible plantear que el enfoque jerárquico puede tener un potencial significativo para aplicarse a niveles más altos de organización textual, como oraciones complejas, párrafos y textos completos. De este modo, extender esta metodología a estructuras textuales más amplias implica analizar no solo las oraciones de manera individual, sino también cómo estas se agrupan y organizan dentro de párrafos y cómo los párrafos interactúan y se relacionan entre sí dentro de un texto completo. Este enfoque permitiría desentrañar las relaciones semánticas tanto internas (dentro de párrafos) como externas (entre párrafos), proporcionando una comprensión más rica y matizada del contenido textual.

Al aplicar el enfoque jerárquico a la agrupación de oraciones en párrafos, es fundamental considerar cómo las oraciones contribuyen al significado general del párrafo. En este nivel, se debe prestar atención a las transiciones entre oraciones, las conexiones temáticas y la cohesión textual. La metodología podría emplear técnicas de composición semántica para capturar la fluidez y continuidad del pensamiento dentro de un párrafo. Este análisis detallado de la organización interna de los párrafos puede mejorar la precisión de la representación semántica y proporcionar insights sobre la estructura lógica y retórica del texto.

En un nivel aún más elevado, el análisis de la relación entre párrafos en textos completos podría plantearse como un objetivo que plantear, ya que esto supondría una representación semántica real la estructura y el flujo del contenido. La metodología jerárquica propuesta es inherentemente flexible y escalable, lo que permite su adaptación a diferentes niveles de granularidad lingüística. Por tanto, se podría abordar desde la composición de palabras individuales hasta la integración de textos completos.

Asimismo, un elemento aún más esencial por estudiar sería abordar si realmente tendría sentido plantear un orden dentro de un elemento lingüístico superior a la oración. Si bien los resultados han sido favorables, es posible plantear las limitaciones de la representación semántica-vectorial, que también están ligadas a las propias limitaciones del lenguaje y de la comprensión humana.

Por otro lado, en algunas ocasiones el enfoque aleatorio ha obtenido mejores resultados. Esto también sugiere que podría ser interesante analizar el orden sintáctico de los mejores resultados obtenidos por el enfoque aleatorio, para ver si estos valores se parecen relativamente a los planteados en este trabajo, o si por el contrario es posible presentar otras propuestas de orden que tengan mayor sentido y mejores resultados que la planteada en este trabajo de investigación.

Bibliografía

- Amigó, E., Ariza-Casabona, A., Fresno, V., & Martí, M. A. (2022). Information Theory–based Compositional Distributional Semantics. *Computational Linguistics*, 48(4), 907–948. https://doi.org/10.1162/coli_a_00454
- Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.
- Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2016). A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4, 385–399. https://doi.org/10.1162/tacl_a_00106
- Bach, E. (1986). The algebra of events. *Linguistics and Philosophy*, 9(1), 5-16.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. (1957). *Syntactic structures*. Mouton.
- Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- Clark, S., & Pulman, S. G. (2007). Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction* (pp. 52–55).
- Coecke, B., Sadrzadeh, M., & Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *CoRR*, abs/1003.4394.
- Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171-4186).
- Dowty, D. (1979). *Word meaning and Montague grammar*. Reidel Publishing Company.
- Dowty, D. R., Wall, R. E., & Peters, S. (1981). *Introduction to Montague Semantics*. Springer.

- Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis* (pp. 1-32). Oxford University Press.
- Frege, G. (1892). Über Sinn und Bedeutung [On Sense and Reference]. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25-50.
- Greenbaum, S. (2005). *Oxford English Grammar*. Oxford University Press.
- Harris, Z. (1954). Distributional Structure. *Word*, 10(2-3), 146-162.
- Jackendoff, R. (1983). *Semantics and cognition*. MIT Press.
- Jackendoff, R. (1997). *The architecture of the language faculty*. MIT Press.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). Prentice-Hall.
- Kant, I. (1781). *Critique of Pure Reason* (N. K. Smith, Trans.). Macmillan.
- Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 2177-2185.
- Lyons, J. (1977). *Semantics* (Vol. 1). Cambridge University Press.
- Maruyama, Y. (2019). A synthesis of compositionality and contextuality in semantics. *Journal of Semantics*, 36(3), 421-446.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT* (pp. 236-244).
- Montague, R. (1970). English as a formal language. In *Formal Philosophy: Selected Papers of Richard Montague* (pp. 188-221). Yale University Press.
- Montague, R. (1970). Universal Grammar. *Theoria: A Swedish Journal of Philosophy*, 36(3), 373-398.
- Partee, B. (1978). Binding Implicit Variables in Quantified Contexts. In *Proceedings of the Tenth Annual Meeting of the North Eastern Linguistic Society* (pp. 342-365).

- Partee, B. (1984). Compositionality. In F. Landman & F. Veltman (Eds.), *Varieties of formal semantics* (pp. 281-311). Foris Publications.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
- Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1* (pp. 448–453). Morgan Kaufmann Publishers Inc.
- Sentis, F. (2006). La composicionalidad en el estudio léxico. *Onomázein*, 13, 73–95. <https://doi.org/10.7764/onomazein.13.05>
- Vidal, M. V. E. (2004). *Fundamentos de semántica composicional*. [Editorial].