3. Lingüística Computacional

3.1. Introducción

La Lingüística Computacional se inició desde la introducción de la Computación y la Inteligencia Artificial (IA) a principios de la década de los cincuenta del siglo pasado. La primera propuesta en el entorno científico y académico para construir y programar una máquina capaz de conversar en lenguaje natural con seres humanos, como el español o el inglés, aparece en el artículo *Computing Machinery and Intelligence*,[1] en el que Alan Turing propuso el programa de investigación para la IA y, en particular, planteó la meta de construir y programar una máquina capaz de entender el lenguaje natural. Es también en este artículo donde se presenta el "Juego de imitación", mejor conocido como la "Prueba de Turing", que propone que una máquina capaz de conversar con un ser humano en lenguaje natural con un alto nivel de desempeño, se tiene que considerar "inteligente". La subdisciplina de la IA que engloba los esfuerzos para lograr dicha meta se conoce como Lingüística Computacional.[2]

Al igual que la lingüística descriptiva tradicional, la Lingüística Computacional aborda una gama muy amplia de fenómenos del lenguaje, tanto hablado como escrito, pero con la restricción de que los modelos se deben ca-

[1] Turing, A. (1950). **Computing Machinery and Intelligence**. *Mind,* 59: 433-460.
[2] https://en.wikipedia.org/wiki/Computational_linguistics

racterizar mediante sistemas de reglas formales y/o procedimientos computacionales, lo que impone una restricción muy severa a los tipos de modelos admisibles. En este enfoque, el objeto de investigación es la representación computacional del conocimiento lingüístico así como los procesos que lo utilizan. Por lo mismo, esta disciplina caracteriza los diferentes niveles de representación lingüística así como sus relaciones. En general se reconocen seis niveles, como sigue: i) el fonético y fonológico, ii) prosódico y entonativo, iii) léxico y morfológico, iv) sintáctico, v) semántico y vi) pragmático.

Por la magnitud de la empresa y con el fin de abordar sus objetivos específicos, la Lingüística Computacional se ha desarrollado históricamente en varias especialidades, con metáforas, teorías y metodologías diferentes, y cada una de éstas ha representado un esfuerzo de investigación y desarrollo tecnológico de dimensiones colosales. Entre las más prominentes podemos mencionar: i) Reconocimiento del habla y síntesis de voz (1952),[3] ii) Traducción automática entre lenguajes naturales (1954),[4] iii) Procesamiento del lenguaje natural (1964)[5] y iv) Lingüística de corpus (1967).[6]

Los programas y dispositivos que se presentaron inicialmente –como prueba de concepto– generaron grandes expectativas y se pensó que la construcción de la máquina del lenguaje iba a ser sólo cuestión de tiempo. Sin embargo, las propuestas no se materializaron y estas disciplinas pasaron por varios ciclos con un fuerte impulso inicial hasta su agotamiento a lo largo de la segunda mitad del siglo XX. A pesar de ello, durante este período se creó un acervo de conocimiento que se reflejó en la aparición de varias revistas emblemáticas donde se detalla la historia de la disciplina, principalmente *Computational Linguistics, Artificial Intelligence* y otras más,[7] en libros de texto de

[3] https://en.wikipedia.org/wiki/Speech_recognition
[4] https://en.wikipedia.org/wiki/History_of_machine_translation
[5] https://en.wikipedia.org/wiki/Natural_language_processing
[6] https://en.wikipedia.org/wiki/Corpus_linguistics
[7] Por ejemplo, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *Computer Speech and Language*, y *Speech Communication*s.

procesamiento de lenguaje natural[8] y reconocimiento del habla[9] y en varias conferencias internacionales de gran prestigio.[10] Asimismo, a finales de la década de los noventa y principios de este siglo aparecieron dos textos que son ahora referencia central para los estudios de posgrado en la especialidad.[11]

Con el fin de contextualizar las contribuciones de la comunidad mexicana a esta disciplina, en la Sección 3.2 se describen los seis niveles de representación lingüística principales, aunque sin comprometerse, en la medida de lo posible, con ninguna corriente de pensamiento o teoría del lenguaje en particular. Con este antecedente en la Sección 3.3 se describen los principales estudios teóricos y aplicaciones realizados por la comunidad mexicana, y en cada caso se hace referencia a los niveles de representación relevantes.

3.2. Modelos computacionales de la estructura del lenguaje

Los estudios computacionales de la estructura del lenguaje se iniciaron con la publicación de *Syntactic Structures*,[12] por Noam Chomsky, quien propuso por primera vez que la estructura sintáctica del lenguaje se puede modelar a través de un sistema de reglas completamente mecanizadas o formales. Esta teoría abrió la puerta para modelar dentro del paradigma computacional no sólo la sintaxis sino los diversos niveles de representación lingüística y tuvo

[8] Winograd, T. **Understanding Natural Language**, Academic Press, Nueva York, 1972; Allen, J. **Natural Language Understanding**, Benjamin-Cumming, 1987 y 1994; Gazdar, G. Mellish, C. **Natural language processing in Prolog**, Addison Wesley, 1989.

[9] Jelineck, F. **Statistical Methods for Speech Recognition,** Cambridge, Mass.: MIT Press, 1997; Huang, X. D., Yasuo Ariki, Y., Mervyn, Y., Jack, A. **Hidden Markov Models for Speech Recognition,** Edinburgh University Press, 1990.

[10] *SpeechTEK*, *SpeechTEK Europe, ICASSP*, *Interspeech/Eurospeech*, y la *IEEE ASRU* en reconocimiento del habla; asimismo *Meeting of the ACL*, *COLING*, *NAACL*, *EMNLP* y *HLT* en procesamiento del lenguaje natural.

[11] Manning, C., Schütze, H. **Foundations of Statistical Language Processing**, MIT Press, 1999; Jurafsky, D., Martin, J. **Speech and Language Processing: An introduction to Natural Language Processing, Speech Recognition and Computational Linguistics**, Prentice Hall, 2009.

[12] Chomsky, N. **Syntactic Structures**, The Hague/Paris: Mouton, 1957.

repercusiones en una gama muy amplia de disciplinas científicas, tecnológicas y en las humanidades. Los niveles principales de estructura lingüística son como sigue:

3.2.1. Nivel fonético y fonológico

Las unidades básicas de la estructura del lenguaje hablado se conocen como fonemas. Los fonemas se reconocen por sus características combinatorias o sintagmáticas, y de contraste o paradigmáticas,[13] y corresponden, hasta cierto punto, con los símbolos del alfabeto. Por ejemplo, la palabra *casa* está constituida por cuatro fonemas en secuencia o en relación sintagmática, es decir *c*, *a*, *s* y *a*, y contrasta paradigmáticamente con las palabras *tasa* y *cara* en la primera y tercera posición respectivamente, por lo que dichas palabras se distinguen entre sí y tienen significados diferentes. El sistema fonológico se caracteriza por las reglas que permiten o restringen las secuencias y contrastes entre estas unidades, y que generan el conjunto de palabras actuales y potenciales de una lengua, como el español o el inglés.

Por su parte, cada fonema da lugar a una realización acústica específica, aunque hay fonemas que tienen diferentes realizaciones. Por ejemplo, en el español de México la *ch* en la palabra *Chihuahua* se pronuncia de manera diferente por hablantes del centro y del norte del país (africada sorda y fricativa respectivamente) pero dicha alteración no cambia su significado, por lo que el fonema es el mismo pero tiene dos realizaciones acústicas o alófonos diferentes. Mientras la fonología estudia la estructura de estas unidades abstractas, la fonética estudia las propiedades físicas de sus realizaciones acústicas.

Las unidades acústicas de los lenguajes humanos se codifican en el alfabeto fonético internacional (AFI),[14] el cual se concibió originalmente para representar la pronunciación de las palabras en los diccionarios bilingües y

[13] Saussure, F. **Curso de Lingüística General**, Editorial Losada. S. A. Moreno 3362 Buenos Aires. 1945.

[14] https://es.wikipedia.org/wiki/Alfabeto_Fonético_Internacional

permitir la comunicación entre hablantes de lenguas maternas diferentes. Por otra parte, cada región lingüística utiliza un conjunto de unidades acústicas específicas y el estudio de la fonética de su dialecto se centra en la definición de su alfabeto fonético. Los alófonos se pueden representar por un símbolo en un alfabeto fonético computacional, el cual se puede asociar a la representación de las características físicas de la señal de audio.[15] Por razones prácticas, se han desarrollado alfabetos fonéticos para cada dialecto para habilitar el reconocimiento de voz y la traducción computacional entre lenguajes hablados.[16]

3.2.2. Nivel de prosodia y entonación

En este nivel se representa la acentuación de las palabras y la estructura tonal de los enunciados. Las unidades sobre las que recaen los tonos son las sílabas y la entonación se caracteriza por la variación de la frecuencia fundamental o F0 en la elocución, la cual permite distinguir las oraciones declarativas, interrogativas y admirativas. Al igual que en el nivel fonético-fonológico, la entonación se puede estudiar desde la perspectiva de la estructura tonal abstracta o en términos de su realización física. Este nivel de representación es necesario para reconocer los tipos de intenciones expresadas en el habla y es fundamental para la creación de voces sintéticas de calidad, que pongan el acento de las palabras en la sílaba correcta y que trasmitan la intención de la elocución de manera clara.

[15] Ver el capítulo 7.

[16] Se debe considerar también que no todo símbolo ortográfico corresponde a un fonema o tiene una realización acústica, ya que el texto es una representación convencional del lenguaje hablado. Por ejemplo, la letra *h* en el español de México es sorda por lo que no se asocia a ninguna unidad fonética; asimismo, los símbolos de puntuación no corresponden a ningún fonema y representan más bien a la prosodia y la entonación. Por lo mismo, es necesario distinguir tres conjuntos de símbolos: los ortográficos, los fonemas y los símbolos acústicos, así como las reglas que traducen a cada unidad o palabra entre estos tres conjuntos.

3.2.3. Nivel léxico y morfológico

El siguiente nivel de representación se enfoca al análisis de la estructura abstracta de las palabras. Esta estructura se puede pensar en términos de una lista de palabras o "lexicón" con una entrada por cada palabra; cada entrada a su vez contiene una secuencia de fonemas, la cual se asocia a uno o varios alófonos así como a la representación de su significado convencional.[17] Intuitivamente, el primer paso para comprender el lenguaje hablado es el reconocimiento de la voz, que consiste en "alinear" a las secuencias de unidades acústicas en la elocución con secuencias de realizaciones de palabras en el lexicón y, a través del mismo, con sus significados convencionales.

Es necesario considerar que la gran mayoría de las palabras tienen variaciones debido a partículas como prefijos, infijos y sufijos que las transforman para especificarlas en algún sentido predecible y reconocible por los hablantes de la lengua; por ejemplo las palabras *inmaterial*, *corredor* y *materialista* se forman respectivamente a partir del prefijo de negación *in*, el infijo *do* y el sufijo *ista*. El primero niega la propiedad adscrita por el adjetivo que se modifica (*material*), la segunda crea el nombre de quien realiza una actividad a partir del verbo que la nombra, y la tercera crea el nombre de quien cree o hace algo a partir de la propiedad que adscribe el adjetivo modificado (*Juan es un materialista*). Si pensamos que el lexicón contiene una forma nuclear de cada palabra, asociada a un significado básico, la morfología computacional estudia las reglas formales que producen las transformaciones posibles del núcleo, que a su vez producen una alteración correspondiente en la representación de su significado. Esta ampliación dinámica del lexicón permite que se puedan reconocer no sólo las palabras actuales sino también las potenciales en conjunto con sus significados convencionales.

[17] Los conceptos o "contenidos" de las palabras se representan en una base de conocimiento; por ejemplo, como predicados lógicos o en una taxonomía conceptual en una representación estructurada. Ver capítulo 1.

3.2.4. Nivel sintáctico

La sintaxis caracteriza la estructura de las frases y las oraciones. En este nivel se considera que dentro de la oración hay palabras cuya función es central en la oración y que las otras tienen un papel subordinado. Estas relaciones se representan como una jerarquía cuyo nodo superior o cabeza es el núcleo gramatical y los nodos inferiores son los constituyentes. Dicha estructura puede ser simple o muy compleja dependiendo de los fenómenos involucrados y del formato de representación y se conoce como "estructura sintáctica". Su relevancia se debe a que es "la portadora" del significado convencional de la oración. Mientras que el significado de las palabras se codifica directamente en el lexicón, el significado convencional de la oración depende de los significados de las palabras y de cómo se combinan en la estructura sintáctica. Por esta razón, desde la propuesta original de Chomsky, inducir la estructura sintáctica a partir de las reglas de la gramática y el estímulo lingüístico se considera como una de las tareas centrales del procesamiento del lenguaje. Este proceso se conoce como "parseo" y su estudio ha sido también sujeto de una investigación de grandes dimensiones en la Lingüística Computacional.

La estructura sintáctica debe tomar en cuenta que tanto las palabras básicas como las derivadas morfológicamente pueden sufrir transformaciones adicionales cuando se ponen en el contexto de una frase o una oración, como las inflexiones debidas al género y al número de los sustantivos, que en el español deben concordar con las inflexiones de los verbos. Por lo mismo, aunque las inflexiones son también partículas que modifican a las palabras su carácter es sintáctico y no morfológico.

Un fenómeno de características morfo-sintáctico muy singular y frecuente en nuestra lengua es el de los llamados pronombres clíticos, como *se* y *lo* en *se lo comió* o *dáselo*. Lo interesante de estas partículas, independientemente de la determinación de sus referencias, es que pueden aparecer como morfemas o como palabras independientes, tanto adelante como atrás del verbo,

de manera muy flexible, además de que llevan implícito su caso (acusativo o dativo) que indica si su referencia recibe la acción verbal directamente o es beneficiario de la misma: en *se lo comió*, *lo* es un pronombre acusativo que denota el objeto de comer, lo comido, y *se* es un pronombre dativo, cuya referencia es el beneficiario de dicha acción, es decir, quien realizó la acción de comer.

De forma muy general el análisis sintáctico se ha abordado desde dos enfoques principalmente: constituyentes y dependencias. Ambos se han estudiado durante más de 40 años y representan alternativas teóricas y metodológicas para el estudio y las aplicaciones de la Lingüística Computacional, como se verá a continuación.

3.2.4.1. Enfoque de constituyentes

En este enfoque, presentado originalmente por Chomsky,[18] un constituyente es una palabra o grupo de palabras que cumplen una función específica en la oración. Estos grupos se conocen como "categorías gramaticales". El proceso sintáctico consiste en encontrar a las categorías de una oración segmentándola en sus partes de manera recurrente hasta que las partes sean las palabras básicas o derivadas en el lexicón. Aunque el número de palabras en el lexicón y el número de reglas sintácticas sea finito, la aplicación sistemática de las reglas sintácticas y morfológicas genera un número infinito de oraciones. En la Figura 3.1 se ilustra una estructura sintáctica por constituyentes, donde *O* representa a la oración, *GN* al grupo nominal y *GV* al grupo verbal. Las estructuras jerárquicas se pueden representar linealmente poniendo la expresión o subexpresión entre paréntesis seguida de un subíndice que indica la categoría del constituyente, como se indica en la parte superior de la Figura 3.1.

[18] Ver nota 12.

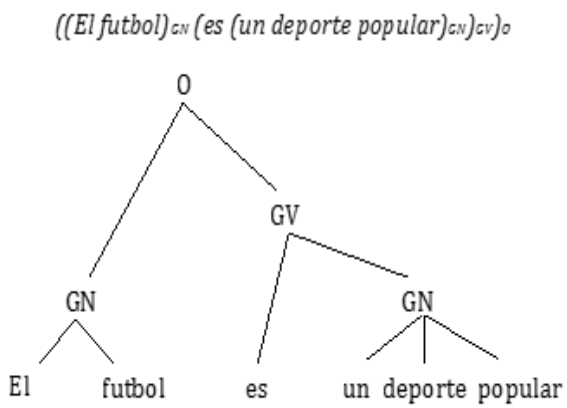

Figura 3.1. Análisis por constituyentes de *El futbol es un deporte popular*.

3.2.4.2. Enfoque de dependencias

Por su parte el enfoque de dependencias iniciado por Lucien Tesnière en 1959,[19] establece que la estructura sintáctica consiste en relaciones de dependencia entre pares de palabras donde una es la principal, rectora o cabeza, y la otra está subordinada. Si cada palabra de la oración tiene una palabra propia rectora (cabeza), la oración entera se ve como una estructura jerárquica de diferentes niveles, o como un "árbol de dependencias". La única palabra que no está subordinada es la raíz del árbol. En la Figura 3.2 se ilustra la estructura de la oración en la Figura 3.1 pero en términos de dependencias.

A diferencia del enfoque de constituyentes, en el enfoque de dependencias no se postulan categorías sintácticas abstractas por lo que las estructuras sólo contienen unidades léxicas concretas. Por lo mismo, la oposición entre el enfoque por constituyentes y el enfoque por dependencias proviene a su vez de una oposición más profunda entre la hipótesis de que la estructura sintáctica es una representación abstracta versus la hipótesis de que ésta tiene un carácter concreto.

[19] Tesnière, L., **Éléments de syntaxe structurale**, Klincksieck, Paris, 1959.

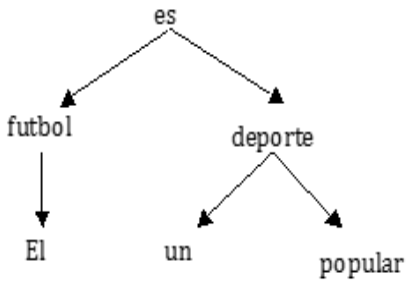

Figura 3.2. Representación en dependencias de *El futbol es un deporte popular*.

3.2.5. Nivel semántico

El propósito de la semántica es caracterizar el significado convencional o independiente del contexto de emisión o interpretación de una unidad lingüística, ya sea básica, como las palabras, o compuesta, como las frases y las oraciones. La semántica computacional se basa en el llamado "Principio de Composicionalidad",[20] que establece que el significado de una estructura compuesta es función del significado de sus partes y de su modo de combinación gramatical. En su formulación más general la semántica de una oración se caracteriza como una función proposicional, donde la palabra principal en la estructura es la función y las otras sus argumentos. Por ejemplo, la representación semántica o estructura de argumentos de *Juan se comió el pan* es la predicación *comer*(e_i, *juan*, *pan*) & *pasado*(e_i) & *beneficiario*(e_i) = *Juan*, cuya interpretación es que e_i es un evento de comer que ocurrió en el pasado, que el agente de dicho evento fue Juan, quien comió, y el paciente (el ente que es pasivo en la acción) fue el pan, lo comido, mientras que *se* es un pronombre dativo que duplica al sujeto e indica que éste se benefició con la acción de comer. Estas representaciones pueden ser de carácter concreto, como en el presente ejemplo, pero también se pueden referir a abstracciones muy profundas. Posiblemente el formalismo más general para caracterizar la semántica de los lenguajes naturales de forma computacional es la Semántica de

[20] http://en.wikipedia.org/wiki/Principle_of_compositionality

Montague[21] aunque existe una gran variedad de formatos representacionales, como redes semánticas, primitivas conceptuales, etc., que son materia de la representación del conocimiento.

3.2.6. Nivel pragmático

Toda unidad lingüística, ya sea hablada o escrita, se realiza e interpreta en relación a un "contexto". La noción de contexto se puede ilustrar con una analogía pictórica: todo cuadro tiene una imagen saliente, "la figura", que ocurre siempre en relación a una escena o "fondo". La unidad lingüística es la figura y el contexto el fondo. En general se pueden distinguir dos tipos de contextos: i) el lingüístico, que consiste en información aportada por el lenguaje mismo a lo largo de la conversación o el discurso y ii) el situacional, que consiste en información extra-lingüística y que tiene como parámetros al hablante, al oyente y posiblemente a uno o más terceros, con sus expectativas e intenciones, así como la situación espacial y temporal en la que se realiza o interpreta la elocución.

El contexto se puede apreciar, como una primera aproximación, en relación a las palabras que funcionan como índices o variables cuya referencia o denotación cambia en cada situación de uso. Los índices más directos y evidentes son los pronombres. Si éstos se interpretan en relación al contexto lingüístico, es decir a lo que se ha dicho en la conversación o el discurso, se llaman anafóricos; por su parte, si se interpretan en relación al contexto no lingüístico, como la situación espacial o temporal, se llaman deícticos o indexicales. Asimismo, las inferencias para encontrar los referentes de los pronombres se conocen como anafóricas e indexicales respectivamente. Esta distinción es esencial no sólo para interpretar el lenguaje en uso, sino también para interpretar representaciones multimodales, como los mapas o diagramas con sus anotaciones textuales.[22]

---

[21] Dowty, D. R., Wall, R. E., Peters, S. **Introduction to Montague Semantics**, Kluwer Academic Publishers, 1981.

[22] Pineda, L. A., Garza, G. (2000). **A Model for Multimodal Reference Resolution**. *Computational Linguistics,* 26 (2): 136-192.

Para ilustrar la complejidad de la inferencia anafórica considere el discurso compuesto por la oraciones *Pepe vio a Juan comprar el pan, se lo comió*. El problema es determinar quiénes son los referentes de los pronombres *se* y *lo*, para lo cual es necesario identificarlos y correferenciarlos con sus antecedentes. Una posibilidad es que *se* tenga como antecedente a *Pepe,* quien comió, y *lo* al pan, lo comido. Sin embargo, también puede ser el caso que quien se comió el pan haya sido Juan e incluso que quien comió haya sido Pepe y lo comido Juan. La inferencia anafórica es sumamente compleja y ha sido también objeto de un esfuerzo de investigación de grandes dimensiones en lingüística computacional.

Por su parte, los pronombres indexicales tienen que tomar su referente directamente del mundo en relación al hablante y al oyente, como en la interpretación de *yo* en *yo vi a Juan comerse el pan*, cuyo referente será quien profiera la elocución en el contexto particular. Hay también otras palabras como los adverbios *aquí*, *allá*, *ahora* o *ahorita* que se interpretan en relación a una locación espacial o un momento o un intervalo en el tiempo y tienen una connotación indexical pura.

Sin embargo, los pronombres que se usan normalmente como anafóricos también se pueden utilizar como indexicales, en cuyo caso el proceso de interpretación se vuelve más complejo. Por ejemplo, en una situación en que el pan esta sobre la mesa y Pepe le dice a Juan ¿*me lo pasas*? al tiempo que señala al pan con su dedo índice, la interpretación de *me* es quien profirió la elocución y es beneficiario de la acción, es decir Pepe, y de *lo* es el pan, donde el contexto relevante es el mundo y no el discurso. Aunque en principio se puede distinguir al contexto lingüístico y al que se establece por la situación en el mundo, en la práctica los contextos interactúan y el problema de interpretación es en general sumamente complejo. Un caso particular de índices son los nombres propios como Pepe, Juan y Pedro, que se refieren en cada caso al individuo en el contexto, aunque haya miles de individuos que han tenido, tienen y tendrán estos nombres.

La interpretación del lenguaje en cada situación de uso, en oposición a la interpretación convencional estudiada por la semántica, se complica aún más cuando se toman en cuenta las intenciones y las creencias de los interlocutores, así como la entonación de las elocuciones. En español las oraciones declarativas que expresan una proposición tienen una entonación relativamente plana, las interrogativas tienen una curva ascendente al final, y las imperativas como las órdenes enfatizan los tonos iniciales. Por lo mismo, es posible identificar si una elocución es un enunciado, una pregunta o una orden, sin siquiera interpretar el sentido convencional de la oración. Se dice que todo enunciado en cada situación de uso es un "acto del habla",[23] y que si la entonación corresponde al tipo del acto del habla, éste es directo, pero si esta relación se cambia, el acto del habla es indirecto. Por ejemplo, enunciar *¿Me puedes pasar el pan?* en un contexto normal de interpretación no es una solicitud de información –si el interlocutor tiene la capacidad de llevar a cabo dicha acción– sino una directiva de acción, y en ciertos contextos puede ser incluso una orden. Los actos del habla indirectos ocurren muy frecuentemente en la conversación cotidiana y su estudio es también objeto de la pragmática.

Los actos del habla indirectos pueden también alterar tanto el material léxico como la forma sintáctica del enunciado; por ejemplo, si Pepe y Juan están en una habitación donde hace mucho frio y el primero le dice al segundo, quien está junto a la ventana, que está abierta, ¿Qué tal tu veranito*?*, la intención es realmente la directiva de acción o incluso la orden de que Juan cierre la ventana. Este ejemplo ilustra cómo el contexto o situación de interpretación espacial y temporal, que involucra también a las expectativas e intenciones de los interlocutores, es indispensable para interpretar al lenguaje en uso.

En resumen, el propósito del análisis pragmático computacional es inferir la intensión o significado de una elocución de manera automática a partir de su significado convencional y del contexto. Asimismo, la salida del módulo

[23] Por ejemplo, ver Austin, J. L. **How to Do Things With Words**, Oxford University Press, 1962. Para una discusión más didáctica ver Levinson, S. C. **Pragmatics**, Cambridge University Press, 1983.

pragmático se puede conceptualizar como la especificación de las acciones que el intérprete tiene que realizar, entre un conjunto posiblemente muy amplio de tipos de acciones, como respuesta a la intención expresada por su interlocutor. Estas acciones pueden ser mentales, motoras e incluso emotivas. Mental, como consultar y reportar información, o hacer una inferencia conceptual, como determinar si dos expresiones son sinónimas, o deliberativa, como hacer un diagnóstico, tomar una decisión o inducir un plan; motora, como tomar un objeto y entregarlo a un destinatario; y emotiva, como ponerse muy contento al recibir una buena noticia.

3.2.7. Ambigüedad

Un problema que aqueja a los modelos computacionales de la interpretación del lenguaje natural es la ambigüedad. Este fenómeno se puede apreciar directamente en el lexicón, ya que hay muchas palabras que tienen más de un significado convencional, como *banco* o *gato*, y su resolución consiste en inferir qué es lo que quiso decir quien las expresó en la situación de uso, tanto en el lenguaje hablado como en el escrito. La ambigüedad aparece también en el nivel sintáctico; por ejemplo, en *Pepe vio a Juan en el banco comiéndose el pan*, es ambiguo quién estaba en el banco, quién estaba comiéndose el pan, y qué quiere decir *banco*. Se invita al lector a visualizar las diversas escenas descritas por esta oración, incluyendo la escena en que los dos están comiendo pan sentados en un banco en el banco. El problema es más agudo si se toma en cuenta que la ambigüedad puede también surgir en el nivel fonético y la entonación, y frecuentemente se sostiene que puede ocurrir en todos los niveles de representación lingüística.

Los modelos generativos o basados en reglas seleccionan una entrada léxica para cada palabra y asignan una estructura sintáctica particular a cada interpretación posible, que se refleja como una predicación particular en la semántica, por lo que la aplicación sistemática de estos procesos produce un número significativo de estructuras sintácticas y, consecuentemente, de

interpretaciones para cada elocución, aunque normalmente sólo una es la apropiada en el contexto. La resolución de la ambigüedad se ha abordado tradicionalmente tanto con la inclusión de restricciones que prevengan la generación de interpretaciones incorrectas como con la generación de todas las interpretaciones para después eliminar las incorrectas, o mediante combinaciones de estas dos estrategias. Sin embargo, si sólo se toma en cuenta el significado convencional de las palabras o las oraciones, el problema es realmente muy complejo, y frecuentemente se sostiene que éste es el mayor problema de la lingüística computacional y que todo lenguaje bien regimentado debería excluir completamente a la ambigüedad, como sucede en los lenguajes formales.

Sin embargo, la ambigüedad es también un recurso expresivo que se puede capitalizar para expresar varias figuras lingüísticas, como la generalización y la ironía, entre muchas otras, y normalmente los seres humanos no generamos interpretaciones incorrectas, cuando menos conscientemente; más aún, somos capaces de apreciar los diferentes sentidos posibles y entender el chiste. Esto se debe a que a diferencia de los modelos formales que utilizan tan sólo significados convencionales, los seres humanos interpretamos el lenguaje en relación al contexto. Considere que si las palabras u oraciones se expresan en un contexto espacial o temporal especifico, digamos por Pedro, quien profiere *Pepe vio a Juan en el banco comiéndose el pan* al tiempo que ve la escena cuando ocurre, o viendo una fotografía del evento, la información extralingüística previene parcial o totalmente la generación de hipótesis incorrectas.

3.2.8. Arquitectura de la máquina del lenguaje

A primera vista, la funcionalidad y relaciones entre los niveles de representación lingüística sugieren que el lenguaje se procesa, es decir, se interpreta y se genera, a través de una estructura "de tubería" (*pipeline*) o "de abajo hacia arriba" (*bottom-up*), en el que los módulos de procesamiento correspondientes a cada nivel de estructura se alinean en serie. Este modelo se adopta explícita

o implícitamente en la gran mayoría de los modelos y aplicaciones de la Lingüística Computacional.

En esta arquitectura el problema de la ambigüedad se resuelve ya sea imponiendo restricciones mutuas entre las representaciones que se tienen que combinar en cada nivel, o restricciones entre los objetos de diferentes niveles, filtrando las posibles interpretaciones a lo largo del proceso hasta el final de la tubería, donde se produce sólo la interpretación correcta. De manera análoga, la generación procede a partir de una especificación abstracta de la intención la cual se especifica de manera semántica, sintáctica y léxica, hasta su realización fonética o textual.

Sin embargo, es también posible pensar que el contexto no sólo restringe de manera activa las posibles interpretaciones durante el proceso de interpretación, sino que participa activamente en la síntesis de la interpretación correcta de manera simultánea con los procesos que actúan en cada uno de los niveles de representación lingüística. Desde este punto de vista, el modelo de tubería se puede sostener, pero en lugar de pensar en el nivel pragmático al final de la tubería, éste se tendría que pensar como un módulo paralelo, que impacta a todos los módulos de manera directa, tanto en el proceso de interpretación como en el de generación.

Sin embargo, en muchas de las aplicaciones tradicionales de la Lingüística Computacional no se tiene información del contexto o no es posible representarlo, por lo que se tiene que adoptar el modelo de tubería y enfrentar el problema de la ambigüedad. Pero, como se sugiere a partir de esta discusión, el problema de fondo no es el de la ambigüedad en sí, sino el de la representación del contexto junto con los problemas asociados a los procesos del módulo pragmático, y cómo estos procesos inciden en el resto de los niveles de representación lingüística.

3.3. Especialidades cultivadas en México

En esta sección se presentan las áreas de investigación desarrolladas por la comunidad mexicana. Iniciamos con el trabajo en lingüística de corpus en la Sección 3.3.1 ya que estos recursos se utilizan en todos los niveles de representación. En la sección 3.3.2 se abordan las contribuciones en tecnologías del habla, donde se hace referencia tanto a los niveles de estructura fonético-fonológica y léxica, como a la lingüística de corpus. La sección 3.3.3 se dedica a los sistemas conversacionales en lenguaje natural que involucran también a todos los niveles. En la sección 3.3.4 se abordan las especialidades relacionadas con el procesamiento de textos que involucran principalmente a los niveles léxico, sintáctico y semántico.

3.3.1. Lingüística de corpus

La lingüística de corpus se enfoca a la creación de recursos lingüísticos, tanto en la modalidad hablada como en la textual. Los corpus son recursos empíricos codificados en formatos computacionales para el estudio de todos los niveles de representación lingüística, así como para el desarrollo de aplicaciones diversas. El desarrollo de corpus exige procedimientos y criterios rigurosos por su magnitud y para su diseño, recolección y organización, de manera que sean confiables y apropiados para las tareas de interés.

La práctica de recabar corpus se inició en el entorno computacional de manera muy temprana con la recopilación del conjunto de textos de diferentes géneros del *Brown Corpus*[24] y más tarde con la construcción de textos anotados sintácticamente, llamados *Treebanks*,[25] particularmente con el *Penn Treebank Project*.[26] Un corpus muy conocido para el lenguaje hablado es el

[24] https://en.wikipedia.org/wiki/Brown_Corpus

[25] https://en.wikipedia.org/wiki/Treebank

[26] Taylor, A., Marcus, M., Santorini, B. (2003). **The Penn treebank: an overview**. En A. Abeille (Ed.) Treebanks: The state of the art in syntactically annotated corpora. Kluwer, pp. 41-70.

Switchboard.[27] Para el español, un corpus textual anotado o transcrito en varios niveles es el Cast3LB Corpus.[28]

Los corpus textuales de primera generación contenían alrededor de un millón de palabras, mientras que los mega corpus actuales contienen más de cien millones de palabras; esto es en parte posible gracias a la Web. Los corpus deben prepararse para su tratamiento computacional, por ejemplo, para la búsqueda de ocurrencias de determinado tipo y número de palabras, la extensión oracional, etc. A continuación se muestran diversas herramientas de utilidad para este tratamiento y en especial para corpora de gran magnitud.

3.3.1.1. Modelos del lenguaje

Una tarea básica en el análisis del lenguaje es la predicción de las unidades que se siguen en una elocución o un texto (por ejemplo, alófonos o palabras) dadas las unidades que se han observado. Esta predicción se puede hacer utilizando modelos probabilísticos, llamados modelos de *n-gramas*. Un *n*-grama es una secuencia de *n tokens* o instancias de palabras; un 2-grama (o bigrama) es una secuencia de dos palabras como *favor de*, *de guardar* o *guardar silencio*; un 3-grama (trigrama) es una secuencia de tres palabras como *favor de guardar* o *de guardar silencio*. Un modelo de lenguaje consiste en la asignación de una probabilidad para todos los *n*-gramas diferentes en relación al corpus de referencia, utilizando para este efecto un algoritmo, normalmente de carácter probabilístico. Por ejemplo, la probabilidad de un unigrama es simplemente la probabilidad de ocurrencia de cada palabra en relación al corpus, y la probabilidad de un bigrama es la probabilidad de que la palabra al final ocurra dada la probabilidad de que la palabra inicial ocurre. Este procedimiento se aplica de manera recurrente para obtener la probabilidad de *n*-gramas de orden mayor.

Es muy importante que el corpus sea representativo o informativo del dominio de interés ya que su calidad impacta directamente en la calidad de

[27] https://catalog.ldc.upenn.edu/ldc97s62
[28] http://www.aclweb.org/anthology/W04-0209.pdf

los modelos del lenguaje. Por lo mismo es común utilizar medidas entrópicas para evaluarlo. Una medida común es la perplejidad, que cuantifica el promedio de las secuencias del lenguaje que pueden seguir a una secuencia dada. Mientras menos sean éstas más predecible e informativo es el corpus, y su entropía tiene un valor más bajo, por lo que el corpus será de mayor utilidad.

Por su parte, en todos los corpus hay contextos que no aparecen de manera contingente aunque puedan ocurrir en el lenguaje, por lo que sus *n*-gramas tendrán un valor de cero. Para enfrentar este problema es común reasignar la masa de probabilidad de los *n*-gramas, quitándole a los que más tienen y distribuyéndola entre los que menos tienen, de forma que la masa total se conserve. Este proceso se conoce como "suavizado" y hay diversas estrategias y algoritmos para llevarlo a cabo. Por ejemplo, si un trigrama tiene probabilidad cero, se le puede asignar la misma probabilidad del trigrama que menos probabilidad tiene en el corpus, al tiempo que se les quita esta probabilidad de manera proporcional a los que más tienen. Este tipo de estrategias son muy útiles en la práctica ya que en un corpus suavizado todos los *n*-gramas tienen un valor, a pesar de que sus partes constitutivas no lo tengan.

Los modelos de lenguaje tienen una gran variedad de aplicaciones, como la corrección de errores gramaticales, reconocimiento de voz, análisis de opiniones, generación de lenguaje natural, medida de semejanza entre palabras, identificación de autoría, entre muchas otras, como se explica más adelante en este texto.

3.3.1.2. Corrección ortográfica

Supongamos por ejemplo que la frase *en quince monitos* ocurre en un texto. El error se puede detectar si la probabilidad del 3-grama *en quince monitos* es cero, pero, al mismo tiempo, la probabilidad del 2-grama *en quince* y del 1-grama *monitos* no lo son; por lo mismo, *monitos* en dicho contexto debe ser un error. Por otra parte *monitos* se parece a *minutos*, lo cual se puede evaluar por algún tipo

de medida de similitud entre palabras; supongamos adicionalmente que el 3-grama *en quince minutos* tiene una probabilidad diferente de cero. Consecuentemente es posible corregir el error substituyendo *monitos* por *minutos*. Por supuesto, esta explicación es sumamente simplificada pero se incluye para dar una idea de la intuición subyacente a este tipo de algoritmos. Algunos trabajos[29] proponen la corrección de errores gramaticales mediante un modelo de lenguaje de trigramas sintácticos[30] continuos y no continuos[31] obtenidos de un corpus de texto de dimensiones muy significativas.

3.3.2. Reconocimiento de voz

El proceso de traducir el habla a su representación textual por medio de un proceso computacional se conoce como "reconocimiento de voz". Este proceso se conceptualiza en relación a los niveles fonético-fonológico, léxico y morfológico. El reconocimiento se inicia con el análisis de la señal del habla para identificar la secuencia de unidades acústicas de la elocución. El proceso propiamente consiste en alinear dicha secuencia con una secuencia de pronunciaciones de un conjunto de palabras en el lexicón. Como el lexicón contiene a las palabras representadas como secuencias de fonemas, posiblemente agrupados en morfemas, se requiere adicionalmente traducir dicha forma abstracta a su realización ortográfica, lo cual se puede hacer conceptualmente mediante reglas que traduzcan fonemas y morfemas a grafemas (o símbolos ortográficos). Los algoritmos de reconocimiento de voz son normalmente probabilísticos, por lo que cada elocución puede dar lugar a un número muy significativo de hipótesis, cada una asociada a un valor de preferencia o *score*.

[29] Hernandez, S. D., Calvo, H. (2014). **Shared Task: Grammatical Error Correction with a Syntactic N-gram Language Model from a Big Corpora**. En *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task* (*CoNLL 2014)*, pp. 53–59.

[30] Sidorov, G., Gupta, A., Tozer, M., Catala, D., Catena, A., Fuentes, S. (2013). **Rule-based System for Automatic Grammar Correction Using Syntactic N-grams for English Language Learning (L2)**. En *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task* (*CoNLL 2013),* pp. 96–101.

[31] Sidorov, G. (2013). **N-gramas sintácticos no-continuos**. *Polibits*, 1(48):69–78.

Por otra parte, los modelos de lenguaje asignan una probabilidad a la elocución, la cual se puede tomar como su probabilidad a priori. Mediante este recurso adicional la hipótesis preferida por el reconocedor de voz será aquella cuyo producto de su *score* acústico y la probabilidad que le asigne el modelo del lenguaje sea mayor.

La creación de sistemas de reconocimiento de voz requiere de dos grandes esfuerzos de investigación: i) desarrollar los algoritmos de reconocimiento propiamente, incluyendo los algoritmos para crear los modelos del lenguaje y ii) crear los lexicones o diccionarios fonéticos de la lengua en cuestión para lo cual se requiere crear los modelos acústicos de todas las realizaciones de los fonemas del dialecto para el que se construye el reconocedor. Para esto último es necesario contar con un alfabeto fonético específico para el dialecto en el que se incluya un símbolo para cada alófono de cada fonema. Asimismo, se requiere contar con una gran cantidad de instancias de la realización acústica de cada alófono, recolectadas en los contextos acústicos en los que pueda ocurrir, para crear su modelo a nivel de la señal. También es necesario crear los diccionarios fonéticos propiamente, los cuales deben incluir todas las palabras del dialecto, normalmente en las decenas de miles, así como las diferentes formas en que cada palabra se puede pronunciar. Además, es necesario contar con corpus de grandes dimensiones en los dominios lingüísticos en los que se utilizará el reconocedor para crear los modelos del lenguaje correspondientes. La creación de estos recursos lingüísticos es también parte de la lingüística de corpus.

Con el fin de contar con un recurso lingüístico de esta naturaleza para el español de la Ciudad de México se desarrolló el alfabeto *Mexbet*,[32] el cual hizo posible diseñar, colectar y transcribir el corpus DIMEx100.[33] Asimismo, este corpus se utilizó para construir un diccionario fonético que incluye las

[32] Cuétara, J. **Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla**, Tesis de Maestría, Universidad Nacional Autónoma de México, México, 2004.

[33] Pineda, et al. (2010). **The Corpus DIMEx100: Transcription and Evaluation**. *Language Resources and Evaluation*, 44:347–370. DOI: 10.1007/s10579-009-9109-9

realizaciones de cada palabra, transcritas adicionalmente en tres niveles de granularidad, y se validó con la construcción de un número muy significativo de reconocedores de voz utilizando los algoritmos del sistema Sphinx producido por la Universidad de Carnegie Mellon.[34]

3.3.3. Sistemas conversacionales en lenguaje natural

Desde la propuesta inicial de Turing el gran reto de la lingüística computacional ha sido la construcción de sistemas que puedan conversar con los seres humanos a través del lenguaje, especialmente hablado. Los esfuerzos en esta línea de investigación se iniciaron a mediados de la década de los sesenta con el programa Eliza[35] que simulaba ser un psicoterapeuta Rogeriano, y un poco más tarde por el programa SHRDLU,[36] ambos desarrollados en el MIT. Esta tradición se continuó en la década de los noventa con sistemas interactivos capaces ya de interactuar en inglés hablado, apoyados por máquinas inferenciales, principalmente para hacer planes, como los sistemas de la serie TRIPS y TRAINS desarrollados en la Universidad de Rochester.[37]

La construcción de sistemas conversacionales o de diálogo en lenguaje natural se ha abordado en México en el contexto del Proyecto Diálogos Inteligentes Multimodales en Español (DIME).[38] Este proyecto se enfocó originalmente en la creación del Corpus DIME[39] consistente en un conjunto de diálogos colaborativos para la solución de tareas, los cuales se etiquetaron ortográficamente, así como en los niveles fonético, prosódico y entonativo, léxico, sintáctico y pragmático. Para el nivel sintáctico se desarrolló una gramática del español de México con énfasis en el sistema de clíticos dentro de

[34] https://es.wikipedia.org/wiki/CMU_Sphinx
[35] https://en.wikipedia.org/wiki/ELIZA
[36] https://en.wikipedia.org/wiki/SHRDLU
[37] Allen, J., Ferguson, G. **TRIPS: An Integrated Intelligent Problem-Solving Assistant**. *Proceedings of AAI*, 1998.
[38] http://turing.iimas.unam.mx/~luis/DIME/
[39] Idem

la perífrasis.[40] Para el nivel pragmático se desarrolló el esquema de análisis DIME-DAMLS[41] mediante el cual se obtuvo la estructura de los actos del habla, directos e indirectos, relativos a la tarea propiamente, a la administración de la tarea y a la administración de la comunicación.

Estas ideas se desarrollaron en paralelo con un modelo para la administración de diálogos computacionales enfocado a la interpretación de los actos del habla en relación a un contexto, el cual se caracteriza como una gráfica recursiva de situaciones llamada "modelo de diálogo", donde cada situación se define en términos de las expectativas y acciones intencionales del agente computacional. En este modelo una situación puede también embeber a un modelo de diálogo subordinado, de tal forma que la estructura del grafo corresponde a la estructura de la conversación. Dichas ideas dieron lugar a la creación del lenguaje de programación SitLog,[42] para la especificación e interpretación de modelos de diálogo en sistemas conversacionales y se incorporaron al Proyecto Golem,[43] para el desarrollo de robots de servicio capaces de comunicarse con los seres humanos en lenguaje hablado, tanto en español como en inglés.

3.3.4. Procesamiento de textos

El procesamiento de textos en México, especialmente para el análisis de los niveles léxico y sintáctico, ha tenido una influencia muy significativa del formalismo de dependencias; esto se debe a la correspondencia directa con los componentes de roles semánticos que conforman la oración, para un rango de

[40] Ver, por ejemplo: Pineda, L. A., Meza, I. (2005). **The Spanish Pronominal Clitic System**. *Procesamiento del Lenguaje Natural*, 34:67–103.

[41] Pineda et al. (2006). **Balancing Transactions in Practical Dialogues.** *Proceedings of CICLing 2006, LNCS 3878*, pp. 331–342. Ver también: Pineda, L. A., Estrada, V., Coria, S., Allen, J. (2007). **The obligations and common ground structure of practical dialogues**. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 11(36): 9–17.

[42] Pineda, L. A. et al. (2013). **SitLog: A Programming Language for Service Robot Tasks**. *Int J Adv Robot Syst*, 10:358. doi: 10.5772/56906

[43] http://golem.iimas.unam.mx/

aplicaciones desde recuperación de información,[44] búsqueda de respuestas[45,46] y minería de texto[47,48] hasta especificaciones de software.[49] Una gran variedad de enfoques semánticos, como grafos conceptuales,[50] Recursión de Semántica Mínima (MRS)[51] o redes semánticas, tienen rasgos similares a un conjunto de predicados.

El trabajo que se ha realizado con el enfoque de dependencias (especialmente para el idioma español) ha sido relativamente reciente, principalmente a partir del año 2000. En particular se ha utilizado *Connexor*,[52] un analizador sintáctico de dependencias que está disponible comercialmente en varios idiomas (inglés, español, francés, alemán, sueco, finlandés, y posiblemente otros). Este analizador se basa en el formalismo de *gramáticas de dependencia funcional* (FDG).[53] En México se desarrolló el analizador sintáctico para el es-

[44] Villatoro-Tello, E., Chavéz-García, O., Montes-y-Gómez, M., Villaseñor-Pineda, L., Sucar, L.. (2010). **A Probabilistic Method for Ranking Refinement in Geographic Information Retrieval**. *Procesamiento de Lenguaje Natural*, 44:123–130.

[45] Montes-y-Gómez, M., Villaseñor-Pineda, L., López-López, A. (2008). **Mexican Experience in Spanish Question Answering**. *Computación y Sistemas*, 12(1):40–60.

[46] Téllez-Valero, A., Montes-y-Gómez, M., Villaseñor-Pineda, L., Peñas-Padilla, A. (2011). **Learning to Select the Correct Answer in Multi-Stream Question Answering**. *Information Processing and Management,* 47(6):859–869.

[47] Ramírez-de-la-Rosa, G., Montes-y-Gómez, M., Solorio, T., Villaseñor-Pineda, L. (2013). **A document is known by the company it keeps: Neighborhood consensus for short text categorization**. *Journal of Language Resources and Evaluation*, 47(1): 127–149.

[48] Montes-y-Gómez, M., Gelbukh, A., López-López, A. (2002). **Text Mining at Detail Level Using Conceptual Graphs**. *Uta Priss et al. (Eds.): Conceptual Structures: Integration and Interfaces, 10th Intern. Conf. on Conceptual Structures, ICCS-2002, Bulgaria, Lecture Notes in Computer Science, N 2393*, Springer-Verlag, pp. 122–136.

[49] Isabel, D., Moreno, L., Fuentes, I., Pastor, O. (2005). **Integrating Natural Language Techniques in OO-Method**. *Gelbukh, A. (ed.), Computational Linguistics and Intelligent Text Processing (CICLing-2005), Lecture Notes in Computer Science, 3406*, Springer-Verlag, pp. 560–571.

[50] Sowa, J. F. **Conceptual Structures: Information Processing in Mind and Machine**, Addison-Wesley, 1984.

[51] Copestake, A., Flickinger, D., Sag, I. **Minimal Recursion Semantics. An introduction**, CSLI, Stanford University, 1997.

[52] http://www.connexor.eu/technology/machinese/demo/syntax

[53] Tapanainen, P. y Järvinen, T. (1997). **A non-projective dependency parser**. *Proceedings of the 5th Conference on Applied Natural Language Processing, Washington, D.C.*, pp. 64–71.

pañol con el enfoque de dependencias DILUCT.[54] Su algoritmo utiliza reglas heurísticas para descubrir relaciones entre palabras, además de estadísticas de coocurrencias de palabras, las cuales aprende de una manera no supervisada para resolver ambigüedades, como la de adjunción de frase preposicional.[55] En las siguientes secciones se presentan varias aplicaciones desarrolladas con este enfoque.

3.3.4.1. Análisis de opinión

El análisis de opinión, también conocido como extracción de opiniones, minería de opiniones, análisis subjetivo o análisis de sentimiento (*sentiment analysis*), ayuda a conocer la percepción de la comunidad acerca de productos o servicios. Este análisis se basa frecuentemente en la información textual disponible en las redes sociales, como Facebook y *twitter*, incluyendo los llamados "emoticones".

Para este análisis se consideran la fuente o emisor, el objetivo o receptor y el tipo de actitud o polaridad, que puede ser positiva o negativa. El análisis de opinión puede ser simple, complejo o avanzado, dependiendo respectivamente de si sólo se reporta la polaridad, de si además se reporta el grado de la actitud, o si también se reporta la fuente y el objetivo, posiblemente con otras características, como teléfonos y fechas, así como los llamados *hashtags* en *twitter*.

La metodología consiste en crear corpus anotados a diferentes niveles de granularidad con los parámetros de opinión y utilizar estos recursos para crear clasificadores mediante toda la gama de algoritmos de aprendizaje au-

[54] Calvo H., Gelbukh A. (2006). **DILUCT: An Open-Source Spanish Dependency Parser Based on Rules, Heuristics, and Selectional Preferences**. *En Kop C., Fliedl G., Mayr H.C., Métais E. (eds) Natural Language Processing and Information Systems. NLDB 2006. Lecture Notes in Computer Science, vol 3999*. Springer, Berlin, Heidelberg, pp. 164–175.

[55] Calvo, H., Gelbukh, A. (2003). **Improving prepositional phrase attachment disambiguation using the web as corpus**. En *Proceedings of the 8th Iberoamerican Congress on Pattern Recognition (CIARP 2003),* Springer Berlin, Heidelberg, pp. 604–610.

tomático, para clasificar o anotar los textos analizados. Esta tarea se puede realizar también a través de lexicones de opinión[56] o mediante la medición de distancias semánticas a diversos conceptos.[57] Éstos se pueden crear de manera automática a partir de corpus anotados.[58]

3.3.4.2. Detección de ironía

Una tarea complementaria al análisis de opinión es el análisis de ironía, ya que esta figura retórica consiste precisamente en expresar una proposición mediante su negación, pero esperando que el interlocutor haga la interpretación correcta. El problema es que si una proposición se expresa irónicamente el análisis de opinión anotará las opiniones positivas como negativas y viceversa. La detección de ironía se ha abordado dentro del procesamiento de textos mediante modelos basado en *n*-gramas simples de categorías gramaticales, de palabras frecuentemente utilizadas en textos con tono humorístico, tales como aquellas relacionadas con sexualidad, relaciones humanas, parentesco, así como de medidas de palabras usualmente clasificadas como positivas o negativas.[59] También se han considerado características de afectividad usando el recurso de WordNet-affect,[60] y una medida de complacencia basada en diccionarios de términos afectivos.[61]

---

[56] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[57] Calvo, H. (2015). **Opinion analysis in social networks using antonym concepts on graphs**. En Proc. 2nd. Int. Conf. on Future Data and Security Engineering (FDSE 2015), L NCS, Springer 9446:109–120.

[58] Minqing, H., Liu, B. (2004). **Mining and summarizing customer reviews**. En *Proceedings of the tenth ACM SIGKDD Int. Conf. on Knowledge discovery and data mining, ACM*, *(KDD 2004),* pp. 168-177.

[59] Reyes, A., Rosso, P., Veale, T. (2013). **A multidimensional approach for detecting irony in Twitter**. *Language Resources and Evaluation*, 47(1):239–268.

[60] Strapparava, C., Valitutti, A. (2004). **WordNet-Affect: an Affective Extension of WordNet**. *Proceedings of the 4th* Int. Conf. on Language Resources and Evaluation *(LREC 2004)*, Lisbon, pp. 1083–1086.

[61] Antonio, R., Rosso, P. (2012). **Making objective decisions from subjective data: Detecting irony in customer reviews**. *Journal on Decision Support Systems*, 53(4):754–760.

3.3.4.3. Semejanza entre palabras y diccionarios de ideas afines

Los diccionarios de ideas afines o tesauros (*thesaurus* en inglés) son recursos lingüísticos que organizan las palabras de acuerdo a la relación que guardan entre sí, tales como sinonimia, antonimia, hiperonimia, hiponimia, meronimia, holonimia, etc. Mediante estas relaciones es posible establecer "distancias" entre palabras;[62] por ejemplo, las palabras más cercanas a *cebra* son *jirafa*, *búfalo*, *hipopótamo*, *rinoceronte*, *gacela* y *antílope*; las próximas a *excepción* son *exención*, *limitación*, *exclusión*, *instancia*, *modificación*; y a *jarrón* son *tazón*, *sartén*, *jarra*, *contenedor*, *platillo* y *taza*.

Los tesauros tienen muchas aplicaciones dentro del procesamiento de lenguaje natural, tales como el análisis sintáctico,[63] la interpretación de conjunciones, la resolución de anáforas, la medición de cohesión en textos, la desambiguación de sentidos de palabras (WSD), la corrección ortográfica y el reconocimiento de voz, entre muchas otras. Algunos de los tesauros más utilizados son: WordNet,[64] Roget,[65] WASPS,[66] Word Sketches[67] y Medical Subject Headings (*MeSH*).[68]

---

[62] Ortega, R. M., Aguilar, C., Villaseñor-Pineda, L., Montes, M., Sierra, G. (2011). **Hacia la identificación de relaciones de hiponimia/hiperonimia en Internet**. *Revista Signos*, 44(75):68–84.

[63] Calvo, H., Gelbukh, A. (2004). **Acquiring selectional preferences from untagged text for prepositional phrase attachment disambiguation**. En *Proc Int. Conf. on Application of Natural Language to Information Systems,* Springer, pp. 207–216.

[64] WordNet puede consultarse en línea en wordnet.princeton.edu

[65] Roget, P. M., Dutch, R. A., ed., **The Original Roget's Thesaurus of English Words and Phrases** (Americanized ed.), New York: Longmans, Green & Co./Dell Publishing Co., Inc., 1962.

[66] Kilgarriff, A, (2003). **WASPS: Thesauruses for natural language processing**. *Proceedings of Natural Language Processing and Knowledge Engineering*. DOI: 10.1109/NLP-KE.2003.1275859

[67] Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). **Itri-04-08 the sketch engine**. *Information Technology*, pp. 105–116.

[68] Rogers, F. B. (1963). **Medical subject headings**. *Bull Med Libr Assoc.* 51:114–6.

WordNet es una base de datos organizada de manera jerárquica que contiene un tesauro en línea junto con un diccionario para el idioma inglés. A través del proyecto EuroWordNet también está disponible para otros idiomas, como el español, italiano, alemán, francés, sueco, checo y estonio. WordNet define los sentidos utilizando un concepto llamado *synset* (conjunto de sinónimos). El *synset* contiene un conjunto de palabras que están relacionadas de manera cercana con la definición de la palabra, como se ilustra en los ejemplos mencionados.

3.3.4.4. Desambiguación del sentido de las palabras

La desambiguación del sentido de las palabras se aborda desde tres enfoques diferentes: i) la desambiguación basada en conocimiento que usa fuentes léxicas externas tales como diccionarios y tesauros,[69, 70] aunado al uso de propiedades del discurso; ii) la desambiguación supervisada, la cual utiliza algoritmos de aprendizaje de máquina y se requiere contar con un corpus en el que se etiquete el sentido apropiado de las palabras[71] y iii) la desambiguación no supervisada, que requiere un conjunto de entradas codificadas, pero no necesariamente el etiquetado del sentido apropiado.[72] También hay enfoques mínimamente supervisados y el sentido más frecuentemente utilizado.[73] En

[69] Calvo, H., Gelbukh, A., Kilgarriff, A. (2005). **Distributional thesaurus versus WordNet: A comparison of backoff techniques for unsupervised PP attachment**. En *International Conference on Intelligent Text Processing and Computational Linguistics,* Springer, Heidelberg, pp. 177–188.

[70] Tejada-Cárcamo, J., Calvo, H., Gelbukh, A. (2008). **Improving unsupervised WSD with a dynamic thesaurus**. En *International Conference on Text, Speech and Dialogue*, Springer Berlin, Heidelberg, pp. 201–210.

[71] Rosso, P., Montes-y-Gómez, M., Buscaldi, D., Pancardo-Rodríguez, A., Villaseñor-Pineda, L. (2005). **Two Web-based approaches for Noun Sense Disambiguation**. *International Conference on Intelligent Text Processing and Computational Linguistics, CICLing-2005. Mexico City,. Lecture Notes in Computer Science 3406*, Springer, pp. 267–279.

[72] Calvo, H. (2008). **Augmenting word space models for Word Sense Discrimination using an automatic thesaurus**. En *Advances in Natural Language Processing.* Springer Berlin Heidelberg, pp. 100–107.

[73] Cárcamo, J. T., Gelbukh, A., Calvo, H. (2008). **An innovative two-stage WSD unsupervised method**. *Procesamiento del lenguaje natural*, 40:99–105.

los enfoques basados en aprendizaje de máquina se construye un vector de características que representan al contexto.[74] En particular para el español, se evalúan los trabajos a través de concursos internacionales como SENSEVAL-2, SENSEVAL-3, que después dieron lugar a SEMEVAL.[75]

3.3.4.5. Detección de engaño

El problema de detección de engaño se ha estudiado ampliamente, para determinar si una opinión es verdadera o falsa, es decir que tiene por fin engañar al lector. En esta tarea, el texto se representa también como vectores de características que se utilizan para entrenar clasificadores mediante técnicas de aprendizaje de máquina.[76] En la mayoría de los trabajos se usan herramientas que requieren información específica del dominio, tales como *n*-gramas sintácticos y diccionarios léxicos de emociones. Sin embargo, otros trabajos consideran únicamente aspectos distribucionales del texto,[77] por ejemplo mediante el uso de algoritmos de modelado de tópicos dentro de un entorno probabilístico. Estos trabajos combinan sus características con otras obtenidas a partir de diversas fuentes, como representación de espacios de palabras, categorías gramaticales o diccionarios de aspectos psicológicos de las palabras. Para efectos de evaluación existen diversos conjuntos de datos enfocados a dominios específicos. Aunque se ha tratado de encontrar un detector universal de engaño con alto desempeño, esto todavía no se ha logrado.

---

[74] Tejada-Cárcamo, J., Calvo, H., Gelbukh, A., Hara, K. (2010). **Unsupervised WSD by finding the predominant sense using context as a dynamic thesaurus**. *Journal of Computer Science and Technology*, 25(5):1030–1039.

[75] http://www.senseval.org

[76] Hernández Fusilier, D., Guzmán Cabrera, R., Montes-y-Gómez, M, Rosso, P. (2015). **Detection of Positive and Negative Deceptive Opinions with PU-learning**. *Information Processing and Management*, 51:433–443.

[77] Hernández-Castañeda, A., Calvo, H., Gelbukh, A., Flores, J. J. G. (2017). **Cross-domain deception detection using support vector networks**. *Soft Computing*, 21:585.

3.3.4.6. Detección de autoría

La identificación de autor es otro problema que se aborda dentro del procesamiento de textos. Por ejemplo, es necesario saber qué autor escribió un libro anónimo, identificar la autoría de las notas de un posible criminal, etc. Esta tarea puede ser muy compleja cuando se realiza en un entorno abierto, es decir, cuando no se tiene un conjunto predefinido de posibles autores. Es por ello que se han creado tareas más específicas donde se incluyen varios documentos de un autor conocido,[78,79] y un documento y un autor desconocido. Entre más cortos son los textos, la tarea se dificulta más, pues existen menos pistas que permitan distinguir el estilo de un autor. En casos reales como el campo forense, difícilmente se cuenta con textos largos. Otra tarea asociada con esta problemática es la identificación del perfil del autor, cuyo objetivo es inferir el género, lugar de origen, rango de edad e incluso rasgos de personalidad del autor a partir de los temas sobre los que versa el texto y su estilo.[80]

Se han considerado diversos enfoques para obtener características más informativas basadas en el estilo; también es posible generar características al extraer información léxica, sintáctica o semántica. La información léxica usualmente se limita a conteos de palabras y ocurrencias de palabras comunes. Por otra parte, mediante la información sintáctica es posible obtener, hasta cierto punto, el contexto de las palabras. Algunos trabajos usan información semántica léxica para encontrar características que permitan discriminar los textos mediante modelos probabilísticos de distribuciones latentes.

---

[78] Coyotl-Morales, R. M., Villaseñor-Pineda, L., Montes-y-Gómez, M., Rosso, P. (2006). **Authorship Attribution using Word Sequences**. *11th Iberoamerican Congress on Pattern Recognition, CIARP 2006. Lecture Notes in Artificial Intelligence 4225*, Springer, pp. 844–853.

[79] Escalante, H. J., Solorio, T., Montes-y-Gómez, M. (2011). **Local Histograms of Character n-grams for Authorship Attribution**. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Portland, Oregon, USA, pp. 19–24.

[80] López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Stamatatos, E. (2015). **Discriminative Subprofile-Specific Representations for Author Profiling in Social Media**. *Knowledge Based Systems*, 89:134–147.

3.3.4.7. Reconocimiento de paráfrasis e implicación textual

Otro reto del procesamiento de textos es el parafraseo, el cual consiste en reconocer si dos expresiones significan lo mismo, o reformular una expresión con otra que tenga el mismo significado. El reconocimiento de paráfrasis puede ser utilizado en aplicaciones como extracción de información, sistemas de pregunta y respuesta, generación de resúmenes de múltiples documentos, detección de plagio, etcétera.

Considere las oraciones E1: *Carlos aprecia la comida francesa*; E2: *A Carlos le gusta la cocina francesa* y E3: *Carlos aprecia la comida francesa picante.* A pesar de que E1 y E2 son oraciones compuestas por distintas palabras la idea que expresan es la misma y por lo tanto se deben considerar como paráfrasis. Por su parte, aunque E2 y E3 transmiten la idea de que a Carlos le gusta la comida francesa, no se pueden considerar como paráfrasis ya que E3 es más específica que E2.

De forma general el procesamiento de paráfrasis puede ser dividido en tres grandes tareas: i) extracción, ii) generación y iii) reconocimiento. La extracción tiene como objetivo obtener el conjunto más grande posible de pares de expresiones que conforman paráfrasis; esto se realiza a partir de un corpus de referencia. La generación, tiene como objetivo generar el mayor conjunto de expresiones en lenguaje natural que sean paráfrasis de la expresión objeto de análisis. Las expresiones generadas deben tener los menos errores posibles. Finalmente, el reconocimiento tiene por objetivo determinar si dos expresiones de entrada son o no paráfrasis.[81] Esta tarea se requiere a su vez para abordar otras, como la detección de plagio.[82]

---

[81] Por ejemplo, Calvo, H., Segura-Olivares, A., García, A. (2014). **Dependency vs. constituent based syntactic n-grams in text similarity measures for paraphrase recognition**. *Computación y Sistemas*, 18(3):517–554.

[82] Sánchez, F., Villatoro, E., Montes, M., Villaseñor, L., Rosso, P. (2013). **Determining and Characterizing the Reused Text for Plagiarism Detection**. *Expert Systems with Applications*. 40(5):1804–1813.

Una tarea relacionada es el reconocimiento de implicación textual. Ésta consiste en identificar si un determinado texto, denominado hipótesis, se implica o se puede inferir a partir de otro texto. Existen trabajos que se basan en enfoques de reconocimiento léxicos, sintácticos y semánticos, los cuales se complementan con diversas técnicas como lematización, eliminación de palabras sin contenido, manejo de negación, relaciones semánticas y semejanza entre palabras.[83] Estos métodos se basan en medir la razón o porcentaje de cobertura de la hipótesis con respecto al texto dado. Para efectos de evaluación se utiliza el marco de referencia PASCAL de reconocimiento de implicación textual.[84]

3.4. Perspectivas

La Lingüística Computacional se seguirá desarrollando a lo largo de todo el siglo XXI. El incremento de la capacidad de cómputo, memoria y conectividad a costos muy bajos, aunado a la proliferación de repositorios de información y dispositivos móviles, aumentarán la infraestructura para explotar la tecnología así como la demanda de servicios de información de carácter lingüístico. En particular hay dos tipos de demanda de gran escala, que además son ubicuos en todo el espectro de aplicaciones. El primero es la necesidad de acceder a textos en formatos digitales en una amplia gama de formatos: libros, revistas, periódicos, páginas personales, etc., además de la información disponible en las redes sociales, cuya diversidad, amplitud y alcance continuarán extendiéndose en los próximos años. En este rubro también se incluyen los recursos textuales relacionados con dominios específicos, como la salud, las leyes, la educación y la capacitación, etc. Es necesario considerar adicionalmente que estos recursos estarán disponibles en prácticamente todos los lenguajes y dialectos. Por lo mismo será cada vez más necesario

[83] Segura-Olivares, A., García, A., Calvo, H. (2013). **Feature Analysis for Paraphrase Recognition and Textual Entailment**. *Research in Computing Science*, 70:119–144.

[84] Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B. (2007). **The Third PASCAL Recognizing Textual Entailment**. RTE '07 *Proceedings of the ACL-PASCAL Workshop on Textual Entailmente and Paraphrasing,* pp. 1–9.

localizar información, extraerla, traducirla, resumirla y expresarla de forma relevante y accesible. El segundo tipo de demanda provendrá de la necesidad de comunicarse con dispositivos computacionales a través del lenguaje natural, especialmente hablado.

Desde el punto de vista de los enfoques teóricos y de la práctica de la especialidad, los métodos estadísticos y el aprendizaje de máquina, como el aprendizaje profundo o *Deep Learning* actualmente de moda, continuarán siendo populares en los próximos años, hasta que estas líneas se agoten y se abran o se retomen otras metáforas. En particular, los métodos actuales producen descripciones globales con información cualitativa que orientan acerca de las tendencias de los fenómenos estudiados y permiten tomar decisiones de carácter genérico. Por ejemplo, la extracción de información, el análisis de sentimiento e incluso la implicación textual producen respuestas cualitativas con un indicador de su grado de certeza o confiabilidad, lo cual permite saber, respectivamente, que hay un número de documentos en Google relacionados con cierta temática, que un producto comercial tiene cierto grado de aceptación en una comunidad o que posiblemente una proposición se siga de un texto.

Sin embargo, también es importante analizar al lenguaje de manera determinada y poder utilizar la información lingüística de manera causal, como cuando Juan cierra la ventana porque Pepe le dice "¿qué tal tu veranito?". Este nivel de comprensión, que es el que realizamos los seres humanos durante la conversación o en la lectura se puede abordar con las técnicas estadísticas sólo de manera limitada; por lo mismo, es posible que su estudio se retome con un grado mayor de madurez cuando se tenga una comprensión más profunda de la facultad lingüística y del cómputo natural, que es el que realizamos los humanos y también otros seres vivos, en oposición al cómputo artificial o de ingeniería enfocado a algoritmos, que es producto de la invención humana.

La meta propuesta por Turing de crear una máquina capaz de entender el lenguaje natural no se ha logrado todavía y no se alcanzará en el corto plazo. Programas como Eliza, que sostenía una conversación aparentemente inteligente simplemente reflejando el lenguaje del interlocutor humano iniciaron una tradición que busca "ganar la prueba de Turing" con trucos superficiales. Esta propuesta ha resultado tenaz y su manifestación actual son los llamados "Chatbots" integrados a sistemas de diálogo, que se utilizan en los *call centers* para comprar boletos de avión o para hacer reservaciones en restaurantes. Una aplicación más ubicua es conversar con un teléfono celular, como Siri,[85] que se adapta al usuario con el apoyo de algoritmos de aprendizaje. Sin embargo, por más inteligentes que estos sistemas parezcan su nivel de comprensión es superficial y son muy sensibles al tipo de información que reciben ya que más que comprender, su meta es engancharse con el usuario humano, a quien tienen que agradar. Un ejemplo reciente de las limitaciones de este tipo de "inteligencia por asociación" es Tay, un Chatbot producido por Microsoft que a las pocas horas de su lanzamiento empezó a enviar twitts racistas y misóginos, además de que su conducta se hizo repetitiva y predecible.[86]

Un escenario posible en el futuro es que los sistemas conversacionales con modelos superficiales se sigan desarrollando y que prácticamente todos los dispositivos con los que interactuamos los humanos, desde las licuadoras hasta los coches y aviones, tengan su sistema de diálogo que les permita conversar acerca de sus objetivos y funcionalidades. En este entorno habrá una percepción de que las máquinas hablan pero siempre de forma esquemática y superficial.

Por otra parte, para decir que las máquinas realmente entienden, éstas tendrían que interpretar y generar expresiones en relación al contexto, desplegando una amplia gama de actos del habla directos e indirectos que sean causales a su conducta. Esto a su vez dependerá de contar con una noción

[85] https://en.wikipedia.org/wiki/Siri
[86] https://en.wikipedia.org/wiki/Tay_(bot)

coherente y comprensiva del contexto, es decir, del nivel de representación pragmático y de su interacción con los demás niveles de representación lingüística. Estas máquinas podrían llegar a existir pero a la luz del estado actual del conocimiento del lenguaje y de la tecnología computacional no es posible afirmarlo todavía.