La industria del lenguaje en la era del dato

Léon-Paul Schaub

La industria del lenguaje en la era del dato

Schaub, Léon-Paul
Akio / Limsi-CNRS Université Paris-Saclay
schaub@limsi.fr

Resumen:

La tecnología del procesamiento del lenguaje natural (PLN, en inglés *NLP*), también llamada lingüística computacional, es hoy en día una industria digital de primer nivel. En este artículo intentamos dar una definición general pero concreta de esta ciencia híbrida a medio camino entre la lingüística y la informática. Indicamos también cuáles son las aplicaciones prácticas del PLN y cómo los gigantes de lo digital[1] se han convertido en referentes mundiales en varios aspectos de esta tecnología. Finalmente, damos algunas pistas sobre el futuro de la industria del lenguaje.

Palabras clave: *lingüística, informática, procesamiento del lenguaje natural, minería de texto, GAFAM*

Abstract:

Natural language processing technology (NLP), also called computational linguistics, is today a major digital industry. In this article we try to give a general but concrete definition of this hybrid science halfway between linguistics and computer science. We also look at the practical applications of NLP and how the numerical giants have become state-of-the-art in various aspects of this technology. Finally, we give hints about the future of the language industry.

Keywords: *linguistics, computer science, natural language processing, text mining, GAFAM*

1. Introducción.

Desde mediados de los años 2000 y de la globalización del acceso a Internet, el procesamiento del dato no estructurado se ha convertido en un reto importante de la ciencia de la información. El dato no estructurado [1] corresponde a todo contenido, escrito u oral, que no ha sido indexado en una base de datos o en un formato específico (XML, Json…), por ejemplo comentarios sobre un producto, o en Facebook, un tweet, mensajes en un foro, articulos de periodico... Al acto de procesar tales textos se le llama "minería de texto" (*text mining*). La minería de texto posee varias aplicaciones como por ejemplo traducción automática [2], corrector instantáneo [3], asistente vocal [4], filtrado de SPAM [5], predicción de palabras [6]…

1 tambien llamados GAFAM o GAFA: Google Apple Facebook Amazon (Microsoft)

En 2020, usamos a diario herramientas desarrolladas por ingenieros especializados en el PLN, como por ejemplo el traductor automático de Google, que mostró una mejoría notable en esta última década, o SIRI, el asistente inteligente de Apple. También usamos otras casi sin darnos cuenta, como el filtrador de SPAMS de nuestro correo electrónico, o el sistema de búsqueda de Google.

De la misma manera, las empresas necesitan las tecnologías PLN, tanto en lo interno (resumen de reuniones, automatización de tareas), como en la relación con sus clientes (respuesta automática a correos y llamadas, FAQ inteligente, bot conversacional (*chatbot) [7]*) o simplemente para ponerse al corriente de la opinión pública acerca de su actividad (e-reputación, análisis de sentimientos, busqueda de opinión [8]).

La pregunta que podemos hacer es la siguiente: ¿**cómo funciona esta tecnología y en qué medida se ha convertido en una industria de primer nivel para las empresas?**

En la sección 2, presentamos un breve historial del PLN y definimos los fundamentos teóricos de esta ciencia, y explicamos por qué se sitúa en el cruce de diferentes disciplinas. En la sección 3 describimos la evolución de esa tecnología en los últimos diez años, así como el progreso del mercado del PLN. En la sección 4, hablamos de la revolución del dato, y de la adaptación de la industria al RGPD. En la 5, describimos brevemente el estado de la técnica en varios campos del PLN y lo que representa a nivel económico en varios países punteros. Y en la 6 proponemos algunas conclusiones.

2. Historia del PLN: lingüística e informática, y temas concomitantes.

2.1 Definición

Al PLN o procesamiento del lenguaje natural se le llama también lingüística computacional, que podemos definir como el estudio del lenguaje humano desde el punto de vista del informático que intenta formalizarlo para que sea interpretable por una máquina. El PLN es sin lugar a duda un campo de la lingüística.

2.2 Orígenes

Los orígenes de la lingüística computacional se encuentran en la teoría de la gramática generativa de Chomsky [9]: *“Una gramática generativa, en el sentido en que Noam Chomsky utiliza el término, es un sistema de reglas formalizado con precisión matemática que, sin necesidad de información ajena al sistema, genera*

las oraciones gramaticales de la lengua que describe o caracteriza y asigna a cada oración una descripción estructural o análisis gramatical." Eso significa que el lenguaje natural puede ser explicado por completo, al menos en teoría, a través de reglas matemáticas.

Unos veinte años antes de las **Syntactic Structures** de Chomsky, en 1936, Alan Turing presenta la teoría de la máquina computacional, un modelo matemático supuestamente capaz de formalizar cualquier operación lógica, más conocida como la máquina de Turing [10], que aún sirve de modelo para cualquier CPU[2] de ordenador de hoy en dia. En 1950, Turing publicó un artículo en el que describe el famoso Test de Turing [11]: se trata de un test al que se somete un sistema informático de interacción hombre-máquina (HMI). El sistema pasa el test si un humano que lo usa no sabe si está interactuando con una máquina o con otro humano. Es el artículo fundador de lo que hoy llamamos inteligencia artificial. Este test sirvió de evaluador para los primeros sistemas conversacionales como ELIZA [12]. Hoy en día el test de Turing sigue siendo una referencia para la evaluación de los sistemas de diálogo hombre-máquina [13]. En 2011, IBM Watson [14], una tecnología de interacción hombre-máquina, ganó el juego televisivo Jeopardy![3] contra dos humanos. Es un acontecimiento importante puesto que Jeopardy! es un juego que requiere cultura general y agilidad intelectual, pero también sentido del humor y del uso de la lengua por la presencia de juegos de palabras en las preguntas del presentador.

2.3 Intersección de Varias ciencias

El PLN es el campo de la lingüística que utiliza la teoría computacional para describir el lenguaje humano, llamado lenguaje natural, en oposición al lenguaje formal (matemático, programación), y al lenguaje binario (máquina de Turing, compilador). Pero según Chomsky, el estudio del lenguaje natural no se puede hacer sin entender la psicología humana a nivel individual (personal) y colectivo (sociedad). En resumen, el PLN es un campo de la lingüística, de la informática, de las matemáticas y también de las ciencias cognitivas.

2.4 Dificultad de la formalización del lenguaje.

En este artículo no tratamos las dificultades cognitivas, sociológicas y geográficas del lenguaje, pero sí de las ambigüedades endógenas de la lengua humana, y por ello nos referiremos exclusivamente a la variedad del español que se ha constituido como estándar "oficial" en España para las finalidades que nos atañen. Lo haremos desde el punto de vista del informático que desarrolla un sistema de comprensión del lenguaje natural (*NLU).*

2 Central processing unit
3 https://fr.wikipedia.org/wiki/Jeopardy!

La lengua natural se puede dividir en cinco niveles de expresión, cada uno de los cuales usa el precedente para definirse.

- **Fonología**: es el estudio de los sonidos del lenguaje y de cómo se interpretan para que tengan sentido (fonemas). La dificultad a este nivel está en el hecho de que un mismo símbolo gráfico, una letra, se pronuncie de dos maneras diferentes y, viceversa, de que un solo fonema se pueda escribir de dos formas distintas:
    - *len**g**ua**j**e / **G**ijón*
    - ***Y**o/ caba**ll**o.*
- **Morfología**: es el estudio de la mínima forma autónoma de la lengua: la palabra. Se descompone en morfemas (la más pequeña unidad lingüística con significado), lemas (asociación genérica de morfemas que definen una palabra sin sus derivados; generalmente es una entrada en el diccionario) y paradigmas (conjunto de un lema y sus derivados (p.e. declinación, conjugación..)). La dificultad a este nivel está en el hecho de que una misma letra tenga varias funciones:
    - la letra /s/ significa a la vez el plural de los substantivos y la 2a persona del singular de los verbos.

Sintaxis: es el estudio de la frase, conjunto de palabras con orden y jerarquía dictada por la reglas gramaticales de relaciones entre dichas palabras. Por ejemplo: *el árbol conduce mejor que ayer*. La frase carece de sentido pero gramaticalmente es correcta. La dificultad está en el hecho que las mismas palabras tengan diferentes funciones en frases idénticas.

- *Juan ve a María con el telescopio,* ¿quien tiene el telescopio, Juan o María?
- *Persigo al ladrón en bici,* ¿quien va en bici, el ladrón o yo?

- **Semántica**: es el estudio del sentido de las palabras. La dificultad a este nivel está en el hecho que la misma palabra tenga varios significados o al revés.
    - *pez / pescado*
    - *sobre (substantivo) / sobre (preposición).*
- **Pragmática**: es el estudio de las frases en contexto. Es la mayor dificultad para un sistema automático pues necesita tener en cuenta el contexto del uso de ciertas palabras.
    - pregunta:“*Quieres café ?*” respuesta: “*Lleva cafeína…*”

La cuestión es ahora averiguar cómo crear una máquina capaz de entender estas ambigüedades como lo haría un humano.

3. Evolución de la tecnología. Diferentes aspectos y aplicaciones.

3.1 Métodos

Desde los inicios de la informática en la década de los 50, los métodos de análisis textual, también llamada análisis de corpus[4], en lingüística computacional se dividen en dos categorías: los métodos simbólicos y los métodos estadísticos.

- **simbólico**: modelo de reglas manuales deterministas basadas en la teoría de los autómatas de estado finito.

  Este método posee dos ventajas. La primera es que las reglas son un sistema muy preciso, por lo tanto se observan pocos errores en el sistema. De hecho, es el método vigente en la mayoría de empresas pequeñas porque tienes pocos datos que analizar. La creación de reglas específicas para un corpus de tamaño pequeño no requiere mucho tiempo. La segunda ventaja es que los ingenieros que crean las reglas no necesitan grandes conocimientos en informática o en matemáticas, sino en lingüística y en el negocio describido por el corpus.

  Sin embargo, el método conlleva algunos inconvenientes: tiene una flexibilidad nula: un modelo de reglas funciona únicamente con los textos usados para su creación, si se cambia ligeramente el corpus, el rendimiento decaerá inevitablemente, porque las reglas son deterministas, no se adaptan a formas textuales que nunca analizaron. Por otra parte, el coste humano es importante: para ser potente, el modelo de reglas necesita permanentemente un humano que añada o actualice reglas para no correr el riesgo de que baje su rendimiento como acabamos de explicar.

- **estadístico**: modelo basado en las probabilidades matemáticas en el que la máquina "aprende" (*machine learning*) a analizar textos.

  Este método posee también dos ventajas. Primero, la intervención humana es menos necesaria, las reglas probabilistas son creadas por la máquina, a través de fórmulas heurísticas [15,16]. Segundo, un modelo e alto rendimiento puede ser aplicado para analizar textos diferentes de los que se usaron para crear el modelo.

  El gran inconveniente es que para que un modelo estadístico sea potente se necesita una gran cantidad de textos porque tiene que generalizar sobre todo un conjunto de textos, a veces difíciles de reunir. Se traduce en un coste computacional cada vez más importante [17]: El estado de la técnica hoy en dia es un modelo llamado Transformer [18] cuyo coste de aprendizaje representa en energía lo que consume un coche durante toda su vida.

Hasta los años 1980 [19] los métodos simbólicos eran los más usados por la comunidad científica, sobre todo porque hasta entonces la potencia de los ordenadores no permitía generar modelos estadísticos en un tiempo humano razonable [20], y porque los textos disponibles eran infinitamente menos que los que

4 conjunto de documentos (textos, imágenes..)

tenemos ahora. Hoy en día, los modelos más populares (y más exigentes en términos de potencia computacional) son los métodos de aprendizaje profundo (*deep learning*) [21], en los cuales se superponen varias capas de cálculos en el modelo de tal manera que la entrada de una capa es la salida de la precedente, lo que permite un alto nivel de abstracción por parte de la máquina y una mejor generalización.

3.2 Evaluación

Para evaluar un sistema de análisis textual, el protocolo suele ser el mismo. Primero se crea un corpus de textos y varias personas enriquecen cada texto añadiendo anotaciones. Después, se mide el acuerdo inter-anotador para saber si las personas están de acuerdo sobre las anotaciones. Esta etapa es importante cuando los textos son difíciles de analizar cuando hay por ejemplo ironía o sarcasmo, o simplemente ambigüedades de las que hablamos en la sección 2 y que pueden confundir a los anotadores. Puesto que el ser humano no es infalible, las anotaciones de referencia son llamadas “verdad de campo”. Cuanto más alto es el acuerdo inter-anotador, más seguras son las anotaciones. El objetivo es crear un sistema que sepa crear estas anotaciones sin intervención del humano. Finalmente se divide el corpus en dos : el corpus de entrenamiento que contiene el 80% de documentos del corpus original y que sirve para el aprendizaje (bien sea para que un humano cree reglas o que un algoritmo estadístico genere un modelo); y el corpus de test, que son los 20% de documentos restantes, desconocidos para el sistema y que sirven para evaluar el modelo generado.

Se usan diferentes medidas según la tarea para la que se crea dicho sistema. Por ejemplo, para la traducción automática, se usa la medida llamada BLEU, que permite, hasta cierto punto, conocer la calidad de la traducción del sistema comparándola con la traducción de referencia, que se encuentra el el corpus de test. Una de las medidas más usadas es la media armónica entre **precisión**y **exhaustividad**, llamada **f-medida**.

La precisión se mide al dividir los documentos correctamente anotados por el sistema por el total de documentos anotados por el sistema: es lo que llaman el **ruido**. La exhaustividad se mide al dividir los documentos correctamente anotados por el total de documentos que tenía que haber anotado: es lo que llaman el **silencio**. la f-medida es una media armónica entre ambas.

Ejemplo : creamos un sistema de análisis de 10 tweets que contienen el *hashtag* “CambioClimático”. Según el contenido del tweet, el sistema debe detectar si el autor cree en el cambio climático o no.

Tweets	humano	sistema	verdadero positivo (vp)	falso positivo (fp)	falso negativo (fn)
Tweet 1	Cree	Cree	X		
Tweet 2	Cree	No cree			X
Tweet 3	No Cree	Cree		X	
Tweet 4	Cree	Cree	X		
Tweet 5	No Cree	Cree		X	
Tweet 6	No Cree	No Cree	X		
Tweet 7	Cree	No Cree			X
Tweet 8	Cree	Cree	X		
Tweet 9	No Cree	No Cree	X		
Tweet 10	Cree	No Cree			X

Precision = $\frac{vp}{vp+fp} = \frac{5}{5+2} = 0.72$

Exhaustividad = $\frac{vp}{vp+fn} = \frac{5}{5+3} = 0.63$

F-medida = 2x $\frac{pr.ex}{pr+ex} = 0.67$

El sistema creado es más preciso que exhausto, por lo tanto cuantos más tweets tenga que analizar, más su f-medida decaerá.

Expliquemos ahora en qué medida el PLN se ha convertido en una industria de primer nivel.

4. Estado de la técnica, industria y economía.

Las GAFAM son las empresas tecnológicas más importantes del mundo. Todas han invertido en la investigación PLN desde hace veinte años para ser punteras no solo en el mercado del dato sino también en la innovación científica. Podemos clasificar el PLN en tres grandes grupos, y observamos que en prácticamente cada rama del PLN, una GAFAM ha desarrollado una tecnología operacional, mas o menos famosa:

- La búsqueda documental
  - motor de búsqueda (Google, Bing de Microsoft)
  - traducción automática (Google translate, Bing translator, Systran)

    - minería de texto (Fasttext-Facebook, AutoML-Google, IBM Watson, BERT-Google)
- La generación de texto
    - resumen automático (Google Brain)
    - Corrector automático (Microsoft Word, Google Android)
    - ayuda a la redacción
- La interacción hombre-máquina
    - sistema de pregunta-respuesta (BERT-Google, XLNet-Google)
    - bot conversacional (IBM Watson, DialogFlow-Google, Microsoft LUIS, ParlAI-Facebook)
    - asistente vocal (SIRI-Apple, Alexa-Amazon, Google Home).

Las GAFAM comparten el hecho de almacenar una cantidad inmensa de datos. Asimismo, Disponen de recursos suficientes como para invertir en las máquinas más potentes del momento, lo que les permite generar avanzados modelos de lenguaje estadísticos (BERT y XLNet de Google) pioneros en su campo. Las GAFAM también comparten la voluntad de querer “abrir” sus tecnologías, y vender únicamente el uso industrial de sus herramientas: Google Translate, por ejemplo, es gratis y sólo es pagando a partir de varios miles de usos diarios por una misma compañía.

Otras empresas, que no tienen ni tanto dato ni tanto presupuesto para invertir en máquinas, necesitan usar las herramientas ofrecidas por las GAFAM, que se hacen omnipresentes puesto que son las más conocidas y las que ofrecen mejores garantías *a priori.* La pregunta que podemos hacer es: ¿por qué las empresas tienen necesidad de desarrollar tecnologías de PLN ? Nosotros planteamos dos respuestas :

1) Para aliviar el trabajo de los empleados:
    a) clasificación automática de correos electrónicos de clientes según su contenido para ganar tiempo.
    b) asistente escrito (*chatbot*) para responder a las preguntas simples del cliente.

2) Para analizar y clasificar los problemas que plantean los clientes:
    a) "minería" de opiniones en los mensajes de los clientes (email, sms…)
    b) analisis automatico de encuestas
    c) e-reputacion en redes sociales (análisis de twitter / facebook) para saber lo que la gente opina de la empresa.

Las empresas tienen entonces dos posibles opciones: o desarrollar sus propias herramientas de PLN, o comprar la tecnología de una empresa especializada. Muchas empresas son reacias a usar la tecnologías de las GAFAM por miedo a que éstas tengan acceso a datos privados y los exploten, o por no querer ser rehenes de la tecnología de un gran grupo. Es por eso por lo que muchas empresas pequeñas y start-ups se especializan en PLN, por ejemplo en traducción automática (Deepl), o en chatbots (RASA), y hacen competencia a las GAFAM.

Entre que la tecnología de PLN es cada vez más necesaria, y que las empresas cada vez se especializan más en PLN, el mercado de las nuevas tecnologías ha conocido un "boom" en los últimos diez años. En la industria, el PLN ha generado 3.000 millones de dólares en 2017, 13.000 en 2020 y las proyecciones calculan que pueden llegar a 40.000 en 2025 segun Statistica.com[5]. En la comunidad científica, el número de artículos sobre el PLN se ha duplicado en los últimos años[6].

5. La crisis del dato privado, escándalo Facebook y RGPD.

En marzo de 2018 varios periódicos revelan que desde 2014 datos personales de millones de usuarios de Facebook han sido utilizados sin consentimiento de dichos usuarios por la empresa Cambridge Analytica (CA), y han servido entre otras cosas para influir fraudulentamente en algunas elecciones favoreciendo el voto a ciertos políticos que habían contratado a dicha empresa. En abril 2018, Facebook admite haber utilizado datos de internautas, registrados en Facebook o no. Este escándalo dejó marcado a Facebook durante meses. En mayo de 2018 entra en aplicación el RGPD (reglamento general de protección de datos), que obliga a todas las empresas que manipulan información personal a respetar una serie de normas muy estrictas para evitar otro escándalo como el de CA y para proteger la vida privada de los usuarios [22]. De ahora en adelante, el RGPD obliga a todas las empresas, incluidas las GAFAM, a preservar la información que manipulan, y a asegurarse de que en los modelos de lenguaje que generan no se puedan encontrar datos personales.

Ahora bien, hay un daño colateral: el RGPD se convierte en el mayor obstáculo para la industria del lenguaje puesto que ciertas tecnologías, como la de asistentes personales, necesitan por esencia tener acceso a datos personales de usuarios para ser eficaces. Por ejemplo, en el PLN medical, el aspecto privado del dato hace que sea prácticamente imposible con el RGPD efectuar "minería" de datos o desarrollar un asistente virtual sin correr el riesgo de manipular información personal [23].

5 https://www.statista.com/statistics/607891/worldwide-natural-language-processing-market-revenues/
6 http://www.marekrei.com/blog/ml-and-nlp-publications-in-2018/

6. Conclusión

En este artículo hemos presentado el procesamiento del lenguaje natural (PLN) a la vez como un campo de investigación de actualidad y como una industria cada vez más importante, así como el difícil reto de formalizar algo tan complejo como el lenguaje humano. Nacido en los años cincuenta, el PLN evoluciona a la vez que evolucionan los procesadores (CPU) y los modelos estadísticos de aprendizaje automático (*machine learning, deep learning*). También hemos explicado las razones por la que el PLN es cada vez más importante a nivel económico para las empresas y su mercado se multiplica año tras año de tal modo que ya se puede hablar de una Industria del Lenguaje. Y, finalmente, hemos hablado del reciente RGPD (2018), el cual impone limitaciones a las empresas que desarrollan tecnología PLN al tener que proteger la información personal y no poder manipular los datos de los usuarios sin su consentimiento. Será interesante ver en el futuro cómo la industria del lenguaje conseguirá desarrollar tecnologías PLN que sean competitivas y a la vez respeten el RGPD.

7. Agradecimientos

Sinceros agradecimientos a Andrea, Federico, Mari-Nieves y José Manuel por sus comentarios y consejos que ayudaron a mejorar el presente artículo. Un agradecimiento especial a Patrick Paroubek sin el que este trabajo no hubiera sido posible.

8. Bibliografía

BEVILACQUA Michele, et al. (2019) «Quasi Bidirectional Encoder Representations from Transformers for Word Sense Disambiguation». Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 122-131

CUN Yann Le, et al. (1987) *Modèles connexionnistes de l'apprentissage*. Intellectica. Revue de l'Association pour la Recherche Cognitive, vol. 2, n.º1, pp. 114-43

DÍAZVILLAAna María. (2005) «Tipología de errores gramaticales para un corrector automático" Sociedad Española para el Procesamiento del Lenguaje Natural, vol. 35

FELDMANRonen and SANGER James. (2006). *The Text Mining Handbook: Advanced*

Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press pp. 1-18

MENDEZ José R.,et al. (2007) *Sistemas inteligentes para la detección y filtrado de correo spam: una revisión.* Inteligencia Artificial,revista Iberoamericana De Inteligencia Artificial vol.*11*, pp. 63-81.

FRANCOPOULO Gil, et al. (2019) «Etude expérimentale de classification textuelle multi-étiquette pour la relation client». EGC 2019 Atelier fouille de textes Text Mine

GENKIN Alexander, et al. (2007)*Large-Scale Bayesian Logistic Regression for Text Categorization*. Technometrics vol. 49, n.º3 pp. 291-304

ARON Jacob (2011) *How innovative is Apple's new voice assistant, Siri?*New scientist , vol. 212, n.º2836, Reed Business Information, p. 24.

KAMOCKI Pawel, et al. (2018) «Data Management Plan (DMP) for Language Data under the New General Data Protection Regulation (GDPR)». Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)

KAY David Jon, et al. (2010) «Contextual prediction of user words and user actions». *US Patent*, 7679534

LEWIS, Burn L. (2012) *In the game: The interface between Watson and Jeopardy!*IBM Journal of Research and Development, vol. 56, n.º3.4, pp. 17:1-17:6

TURING Alan M. (1937), «On Computable Numbers, with an Application to the Entscheidungsproblem»Proceedings of the London Mathematical Society, s2-42: 230-265.

MOORE Gordon E. (1965) *Cramming more components onto integrated circuits,*Electronics, volume 38, number 8, April 19, pp.114 ff

SOCHER Richard, BENGIO Yoshua and MANNING Christopher (2012) . «Deep learning for NLP (without magic)»Tutorial Abstracts of ACL 2012 (ACL '12). Association for Computational Linguistics, USA, 5.

RADZIWILL Nicole M. and MORGAN C. Benton. (2017) *Evaluating Quality of Chatbots and Intelligent Conversational Agents*. arxiv, abs/1704.04579

SCHAUBLéon-Paul and VAUDAPIVIZ Cyndel. (2019) «Les systèmes de dialogue orientés-but : état de l'art et perspectives d'amélioration». RECITAL' 2019 SILVA Catarina and BERNADETE Ribeiro (2006). *On Text-Based Mining with Active Learning and Background Knowledge Using SVM*. *Soft Computing*, vol. 11, n.º6, Springer, pp. 519-30.

STRUBELL Emma, et al. (2019) «Energy and Policy Considerations for Deep Learning in NLP». Proceedings of the 57th Annual Meeting of the Association for Computational Linguisticsvol. 1, pp 19-1355.

SUAREZ Pedro, et al. (2019) «Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures». 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Leibniz-Institut für Deutsche Sprache

CHOMSKY Noam (1956) *Three models for the description of language*IRE Transactions on Information Theory, vol. 2, no. 3, pp. 113-124,

TURINGAlan. M (1950) *I.—COMPUTING MACHINERY AND INTELLIGENCE*. Mind, vol. LIX, n.º236, 1950, pp. 433-60

WEIZENBAUMJoseph (1983) «ELIZA --- a computer program for the study of natural language communication between man and machine». *Communications of the ACM*, vol. 26, n.º1, pp. 23-28

YAMADAKenji and KNIGHT Kevin (2001). «A syntax-based statistical translation model». *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*,2001