

# Computational linguistics

## In brief



 SPA [Lingüística computacional](#)

### other names


*Natural Language Processing* (NLP) is commonly used as a synonym of *Computational Linguistics*, although historically they have been used differently, as explained in the introduction of this entry.

### abstract


Computational Linguistics (CL) is an interdisciplinary field at the intersection of Artificial Intelligence, Linguistics, Cognitive Science, and Computer Science, devoted to the study of natural language from a computational perspective. The main theoretical and engineering goals are the understanding of natural language and the consequent development of tools and systems that can process and generate natural language, respectively. This entry focuses on CL applied to the written modality and covers the following aspects:

- (i) the most common CL methods applied to different levels of analysis of natural language (morphology, syntax, semantics and pragmatics);
- (ii) an historical account of CL through the main computational approaches used in this field and the impact of these approaches on the relation between CL and other areas of Linguistics;
- (iii) the different types of language resources used in CL and their role; (iv) an overview of the most prominent applications of CL;
- (iv) an outline of the current trends in CL and its areas of application.

### record

 Antonio Toral Ruiz & Tommaso Caselli

 2022

 Toral Ruiz, Antonio & Tommaso Caselli. 2022. "Computational linguistics" @ *ENTI (Encyclopedia of translation & interpreting)*. AIETI

 <https://doi.org/10.5281/zenodo.6366254>

 [https://www.aieti.eu/enti/computational\\_linguistics\\_ENG/](https://www.aieti.eu/enti/computational_linguistics_ENG/)

# Entry



 SPA [Lingüística computacional](#)

## contents

[Introduction](#) | [Levels of analysis](#) | [Computational approaches](#) | [Language resources](#) | [Applications](#) | [Research potential](#)

## Introduction

Computational Linguistics (CL) is an interdisciplinary field at the intersection of Artificial Intelligence, Linguistics, Cognitive Science, and Computer Science, devoted to the study of natural language from a computational perspective. The main theoretical and engineering goals are the understanding of natural language and the consequent development of tools and systems that can process and generate natural language, respectively.

Historically, the term CL is commonly used to refer to theoretical goals, i.e., developing computational models of natural language and understanding how computation of meaning is performed by humans. In its turn, the term *Natural Language Processing* (NLP) tends to be used to refer to the engineering aspects, i.e., the development of computer programs that process natural language. Often, however, these terms are used as synonyms in interchangeable contexts.



**Figure 1.** *Developing systems to generate natural language.*

This entry focuses on CL applied to the written modality and it is structured as follows. First, we will cover CL methods applied to different levels of analysis of natural language. Next, we describe the main computational approaches used in CL. Subsequently, we present Language Resources (LRs) and their role in the larger context of this discipline. We also provide an overview of the most prominent applications of CL. Finally, we outline the current trends in CL and its areas of application.

[back to top](#)

## Levels of analysis

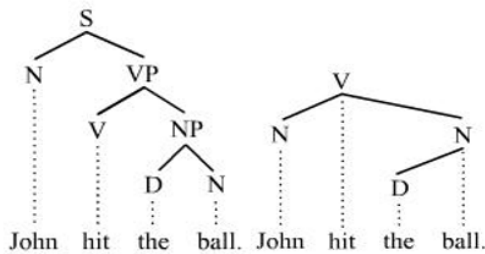
Natural languages are complex systems. In fact, the scientific study of language (Linguistics) has identified different sub-fields including the study of form (phonology and morphology), structures (syntax), meaning (semantics), and use (pragmatics). The remainder of this section illustrates how CL connects to each of these sub-fields (except phonology) by means of different levels of analysis, or tasks. Phonology is mostly used when dealing with spoken text and is hence excluded since this entry focuses on written text. Furthermore, we only discuss the main tasks, or levels of analysis, due to space limitations. For a more exhaustive description of CL tasks, interested readers are referred to Jurafsky & Martin (2020) or Eisenstein (2019).

The first task concerns the identification of which strings compose a token. Tokens can be normally understood as corresponding to words and they are considered the basic unit of analysis in CL. This layer of analysis is known as *tokenisation*. What this entails depends on the language, e.g., for Indo-European languages this corresponds to introducing spaces between words and punctuation symbols, while for Sino-Tibetan languages, in which there is no explicit separation (whitespace) between words, this comes to the [identification of word boundaries](#).

Tokens need to be further analysed and distinguished into the different parts-of-speech (e.g. nouns, verbs, adjectives, articles, among others). This task is known as *part-of-speech tagging*. The aim is to assign to every token, or meaningful combination of tokens (e.g., multiword expressions), a label that expresses basic morpho-syntactic information distinguishing whether it represents, e.g., a noun or a verb, whether it is, e.g., singular or plural, whether it expresses a finite tense, among other features. Different language-specific repositories of part-of-speech tags have been developed in the course of the years. A language-independent proposal is the [UD tagset](#).

Another common task connected with morphology is *lemmatisation*. The aim is to assign its lemma to each token, i.e., its canonical form. In English, for instance, the words "funded", "funding" and "funds" would be assigned the same lemma: "fund". A motivation for this task is to reduce sparsity due to the presence of different inflected forms of the same lemma, especially during training (see Computational Approaches for more details). A related task is morphological analysis. Its goal is to identify and make explicit the morphemes that make up each token and their morphological features, e.g., gender, number, case, tense. Morphological analysis is very useful when working with languages with productive morphology (e.g., Turkic and Finno-Ugric languages).

The identification of how words can combine together to form grammatical structures is addressed by *syntactic parsing*. Following theoretical frameworks and formalisms from Linguistics, there are two common ways to address this task. The first, constituency parsing, is inspired by Chomsky's paradigm of [Context-Free Grammars](#) (CFGs) and is based on the identification of constituents (e.g., noun phrases and verb phrases) and rules for their combination. The second, dependency parsing, inspired by the work of the French linguist [Tesnière](#), informs syntactic structure by means of asymmetric relationships, called dependencies, between tokens. In both formalisms, each syntactic relationship holds between a pair of tokens. In CL dependency parsing is now the most used paradigm to represent syntactic structure as it has shown some advantages (Jurafsky & Martin 2020 - chapter 15) among which the fact that it can be easily applied to different language families, ease in annotation, ease of representation of the grammatical function, and ease of computation (in terms of representation format). Figure 2 graphically illustrates the differences between these two syntactic representation paradigms for the same sentence: "*John hit the ball*".



**Figure 2.** Examples of constituency (left) and dependency (right) parses for one sentence. S stands for sentence, NP for noun phrase, VP for verb phrase, D for determiner, N for noun and V for verb. Source [Wikipedia](#) CC-BY-SA 3.0.

While syntax can be conceived as the study of *formal* relations between words (or constituents), the representation of what a word denotes in the world as well as how words combine together to give rise to meaningful expressions is addressed by *semantics*. The study of semantics is inherently linked to the subject of meaning representations, i.e., the relationship between signs and signifiers, reference and denotation. In CL there are different tasks that can be related to this area of Linguistics. In this entry we will shortly describe two of them: Word Sense Disambiguation (WSD) and Semantic Role Labelling (SRL).

WSD is the task of identifying which sense a word has in context and is commonly used as a tool to evaluate natural language understanding models. WSD has been traditionally investigated by means of manually curated computational lexical resources (e.g., [WordNet](#)) that list the senses of the words as well as their possible relations (e.g., synonymy, hypernymy/hyponymy, etc.). More recently, a new methodology inspired by [the Distributional Hypothesis](#) (Firth 1957), called *Distributional Semantics*, has been successfully applied (Baroni & Lenci 2010). In this case, word meanings and relations are inferred from large amounts of textual data from corpora. Further developments of the distributional semantic approach are word embeddings (Mikolov, Sutskever, Chen *et al.*, 2013), a set of learning techniques that map word meanings to a vector space of real numbers.

SRL contributes to formalise so-called *shallow semantic representations* at sentence level. In particular, SRL makes explicit what is the *semantic role* of the different arguments associated with a predicate expression (whether a verb or a noun). For instance, in the sentence "Mary ate an apple", "Mary" is understood to be the **agent** of the action, while "the apple" is the **theme** of the action, i.e., the argument that is affected. Systems for SRL assign the roles on the basis of a list codified in dedicated resources. Two of the most used for English are [PropBank](#) and [FrameNet](#), the latter being available for one Iberian language: [Spanish](#). Semantic roles are also referred to as *thematic roles* (Fillmore 1968).

Moving away from single words and isolated sentences, the study of meaning of larger chunks of text has seen a rich array of approaches and theories in Linguistics and Philosophy, such as [Formal Semantics](#), [Truth-conditional Semantics](#), and [Cognitive Semantics](#). CL focuses on the development of automatic procedures, or algorithms, to construct meaning representations of natural language expressions, which can then be used for automatic reasoning. Some of the methods used to develop these automatic approaches rely on the theories above, while others follow data-driven approaches. Traditional approaches, linked to compositional theories of meaning, see the sentence as the basic unit of meaning and discourse as the intersection among the semantic values of the constituent sentences. Such a vision has been challenged by dynamic theories of semantics (e.g., [DRT](#), [SDRT](#), [Dynamic Predicate Logic](#)) that promote a vision where sentences are dynamic elements which instantiate relations from one discourse context to the other. The notion of discourse context varies and it can be a representational structure (e.g., DRT), a set of assignment functions, or a set of model assignment pairs. While the study of semantics can be traced back to [Aristotle](#), discourse semantics and computational discourse modeling represent quite recent developments

both in the area of Linguistics (Mann & Thomson, 1988) and CL (Grosz & Sidner [1986](#); Webber [1988](#); Gardent & Webber [1988](#)).

To conclude this section, we report on a task that connects the semantic and pragmatic sub-fields: *anaphora resolution*. In linguistics, reference in a text to an entity that has been previously introduced in the discourse is called anaphora, while the referring expression is called anaphor. For instance, in the sentence *Mary loves ice-cream, but John hates it*, the pronoun *it* is an anaphor for the noun *ice-cream*, that is called antecedent. CL approaches to anaphora resolution aim at automatically identifying mentions of anaphors and linking them to their antecedent. This requires for a system to be "aware" of the discourse structure as well as pragmatic factors that may influence the resolution of the task.

Overall, in CL these tasks have been traditionally seen as following a hierarchical structure where the information flows from one task to the other, as in the following two contemporary examples, which follow sequential pipelines:

- [UD Pipe](#) (Straka & Straková [2017](#)), in which the input text is first tokenised (surface level), subsequently PoS tagged and lemmatised (morphology level) and finally dependency parsed (syntactic level).
- [The Parallel Meaning Bank](#) (Abzianidze, Bjerva, Evang et al., [2017](#)): given an input text, the following processing pipeline is run: tokenisation (surface level), then constituency parsing (syntax level), followed by semantic tagging (semantic level), symbolisation, semantic role labelling (semantic level), word sense disambiguation (semantic level), anaphora resolution (semantic/pragmatic level), and discourse representation (semantic level).

An example of how these tasks, linked to levels of linguistic analysis, play a role in a CL application is in Machine Translation (MT), where the translation may be performed e.g., (i) at surface level, (ii) using surface and morphological information or (iii) surface, morphological and syntactic information.

[back to top](#)

## Computational approaches

In general, we can identify three big waves of computational approaches that have informed the development of tools for natural language understanding, generation and translation. The evolution of each of the different waves has been highly influenced by the technological advancements and computing power of hardware. Given that these waves took place sequentially, a historical account of the discipline emerges naturally from their overview, which we now present.

1. **Rule-based**, also referred to as knowledge-based. In this approach experts in the linguistic aspects targeted write up rules manually. Approaches vary from simple pattern matching rules to knowledge grounded in theoretical frameworks from formal linguistics (e.g. generative grammars, DRT).
2. **Statistical**. This is a data-driven approach, in which a computational system induces knowledge from data (e.g. texts). This methodology has been attributed to have originated in the field of speech processing and was later applied to written text. The following quotes provide an example of the *feeling* at the time of their introduction in CL: "Anytime a linguist

leaves the group the [speech] recognition rate goes up" (Jelinek [1988](#)) and "There is no data like more data" (unpublished comment by Mercier in 1985, cit. in Jelinek [2004](#)).

3. **Neural.** This approach uses artificial neural networks and like the statistical approach it is also data-driven, a key difference being the type of representations typically used (Jurafsky & Manning [2020](#) - Chapter 6): discrete (sparse) in the statistical approach and continuous (dense) in the neural one.

It should be noted that the distinction between these approaches is in practice not so clear-cut, since there are interactions and combinations across them, resulting in what can be referred to as *hybrid* systems. The term hybrid is normally used in this context to refer to a combination of rule-based and data-driven approaches (see Klavans & Resnik 1996 for an overview and discussion of the problems and solutions). A recent trend is on hybrid systems that combine neural networks and knowledge-based approaches (Gómez, Denaux & García 2020).

Within data-driven approaches, there is a further distinction depending on whether the data used to train a model (i.e., training data) is labelled or not. In the first case the systems are referred to as **supervised** while in the second they are called **unsupervised**. An example of a label could be the part-of-speech (PoS) of each word; a supervised system for automatically predicting the PoS (i.e., commonly referred to as a PoS-tagger) would be trained on data where each word has been manually annotated with its PoS label. On the other hand, unsupervised systems rely on large quantities of non-labelled data. A common technique for unsupervised learning is **clustering**, i.e., the aggregation of data points that are similar according to some criteria. For instance, PoS-tagging has been addressed in the unsupervised paradigm, being re-framed as PoS Induction (Biemann 2011; Stratos, Collins and Hsu [2016](#); Cardenas, Lin, Ji *et al.* [2019](#)). The unsupervised paradigm has seen applications in other areas of CL such as Word Sense Induction ([Navigli 2009](#)) and MT (Artetxe, Labaka, Agirre *et al.* [2018](#)).



**Figure 3.** *Artificial neural networks are revolutionising natural language processing.*

It is often the case that small amounts of labelled data are available for a given task (i.e., data that could be used to train a supervised system) while there are large amounts of unlabelled data (i.e., data that could be used to build an unsupervised system). In order to make the most of both types of data, one can use **semi-supervised** approaches.

It is the perceived wisdom in the CL community that the wide adoption of data-driven methods, together with the use of machine learning techniques, have contributed to change the field profoundly. First, data-driven models have proven more cost-effective to build when compared to rule-based systems, and have obtained competitive or even better results. Second, the use of supervised data-driven approaches has boosted the development of annotated data with corresponding annotation schemes, most of which are informed by linguistic theories or frameworks. Finally, there is an increasing loosening of the dialogue between CL and other areas of Linguistics (Baldwin & Kordoni [2011](#)).

## ¶ Language resources

The term *language resource* is defined by the [European Language Resources Association](#) as “a set of speech or language data and descriptions in machine readable form”. Language Resources occupy a central role in Computational Linguistics since any of its approaches (see section Computational Approaches) requires some form of data. In the following, we establish a distinction between three types of Language Resources - corpora, structured linguistic datasets and benchmark datasets - on which we elaborate further and for which we provide examples for Iberian languages.

### Corpora

A corpus is "a collection of pieces of language that are selected and ordered according to explicit criteria in order to be used as a representative sample of the language" (Sinclair [1996](#)). Several further distinctions can be made, e.g. depending on the modality (written versus spoken; where the second may be provided as audio and/or transcribed), on the annotation (plain text, i.e. without any annotations, versus annotated; where the annotations can concern many different aspects, such as different levels of analysis or metadata), or on the languages covered (monolingual versus multilingual; where multilingual corpora may be parallel or comparable).

A well-known annotated corpus for Spanish and Catalan is [AnCora](#), which contains 500,000 words for each language with several annotations at the morphological, syntactic and semantic levels. [CORILGA](#) is a spoken corpus of Galician annotated with speaker metadata. [TweetMT](#) is a parallel corpus of tweets for Spanish paired with Catalan, Galician, Basque and Portuguese. [caWaC](#) is a plain-text corpus of Catalan crawled from the web. Several other corpora for Iberian languages can be found in [Sketch Engine](#), e.g. for [Portuguese](#).

### Structured linguistic datasets

We use this term to refer to a varied set of basic language resources that, unlike corpora, have some degree of structure. These include computational lexicons, ontologies (both [foundational](#) and [domain-specific](#)), and terminologies. These resources have become popular in CL with the development of machine-readable dictionaries. Although their use was initially limited to the study of the lexicon, methods for their application to text analysis have since gained popularity. These datasets can be described as structured repositories of lexicographic information and/or world knowledge. A popular such repository for Iberian languages is the [Multilingual Central Repository](#) (MCR), which integrates wordnets for Spanish, Catalan, Basque, Galician, Portuguese and English as well as the [Adimen-SUMO](#) ontology. [BabelNet](#) is a multilingual encyclopedic dictionary and a semantic network that was built by linking Wikipedia to WordNet. It contains linguistic data for all the Iberian official languages plus Aragonese, Asturian and Extremaduran.

A popular research topic in this area concerns the interoperability and re-use of these resources, given the time and human efforts it takes in creating them. Different standardisation initiatives have been promoted (e.g., the [ISO TC37 Working Group](#)). The promotion of the Semantic Web has boosted the Linked Open Data community which has found its proponents, methods, and



applications in the area of lexicons and ontologies by giving rise to the [Linguistic Linked Open Data](#) movement.

## Benchmark datasets

Under this term fall a variety of datasets specifically developed to evaluate the performance of multiple NLP systems. The major distinction between a benchmark dataset and a corpus is that a benchmark is focused on a specific language phenomenon against which performances of systems are evaluated.

The use of benchmark data has gained popularity through evaluation campaigns which can be described as series of focused workshops or conferences centered around one or more tasks (i.e., the phenomenon of interest) and a common timeline for the release of the training and test distributions, and submission of the system predictions by the participants. A distinguishing feature of evaluation campaigns is that participants do not have access to the gold labels of the test data until the end of the so-called "evaluation period".



**Figure 4.** *Releasing code is important to ensure reproducibility.*

One of the most important evaluation campaigns in CL is [SemEval](#), which focuses on semantic aspects. Evaluation campaigns are nowadays also organised at national level both as a way to assess the state of the art of systems in specific languages. Examples are [IberLEF](#) for Iberian languages, [EVALITA](#) for Italian and [GermEval](#) for German.

Thus far we have followed a strict definition of Language Resources, limited to datasets. However, this term can have a broader interpretation that also includes **tools**. Nowadays it is very common in CL to release not only the data but the code used to run experiments, often under [free and open-source licenses](#). This is deemed very important to ensure reproducibility (Pedersen [2008](#)).

Disparity in the availability of language resources across languages is a known issue and a bottleneck for CL. An overview of the status of the coverage of language resources for 32 EU languages has been conducted during the META-NET Project resulting in the publication of the White Paper series. Each of these 32 EU languages was assessed for language technology support in four different areas: machine translation, speech interaction, text analysis and availability of language resources. As we have illustrated, the availability of annotated corpora or datasets plays a crucial role for the development of systems. Low- and under-resourced languages (e.g., Basque and Catalan, among many others) suffer from this disparity. This may also have negative effects in society: for instance, the lack of adequate MT systems is representing a problem to guarantee [access to correct information during the COVID-19 pandemic](#).

[back to top](#)

## ¶ Applications

Many **tasks** are researched in the field of CL. These are commonly classified according to the level of linguistic analysis at which they operate (see Section Levels of Analysis). Besides tasks, there are

also *applications*. While the main focus in tasks is to address a particular phenomenon of a natural language computationally, the main focus in an application is to tackle a real-world problem that involves natural language, and which is normally broader than a single task. In the following we describe briefly a set of popular applications from CL alphabetically:

- **Authorship attribution** aims to determine who the author is for a text whose authorship is not known.
- **Machine translation** tackles the automatic translation of texts between different natural languages.
- **Optical character recognition** (commonly abbreviated as OCR) aims to convert images of text, which may be typed or handwritten, into machine-encoded text.
- **Question answering** aims to find answers to questions posed by the user in large collections of text.
- **Sentiment analysis** aims to determine affective states in text. A common example is the classification of posts in social media as either positive, negative or neutral.
- **Text summarization** aims to shorten input texts preserving the most important information.
- **Writing assistants** aim to improve the quality of writing of the user, e.g. by correcting spelling and grammatical errors and by providing auto-complete suggestions.

[back to top](#)

## **Research potential**

The application of CL methods and systems can impact many different areas, ranging from education technologies to everyday activities (e.g., virtual assistants such as Siri or Alexa), to mention just two. In addition, new fields are emerging and gaining increasing popularity, such as social media analytics, where CL techniques are applied to social media messages to, e.g., monitor the popularity of products or people, or computational sociolinguistics. The increasing reliability of CL models has seen their applications in domains such as Humanities and Medicine. A very successful application of CL in these areas is the analysis of massive amounts of text data to identify potentially relevant novel information that may need further investigation. An example is the [BiographyNet](#) project, which has seen the use of text analysis technologies to historical documents in coordination with historians. More recently, the outbreak of the COVID-19 pandemic has pushed the development of information extraction applications from [COVID-related scholarly publications](#). All in all, the availability of NLP technologies is radically and rapidly transforming the market, and [further growth is forecasted](#).

Recently, the last wave of NLP systems based on fine-tuning pre-trained language models (LMs) together with the increasing presence of ready-to-use code has pushed for an increased democratisation of CL and NLP. These LMs use big neural networks and are pre-trained on vast amounts of data, entailing high computational requirements. The cost to train the most recent of such models at the time of writing, [GPT-3](#), is estimated at 12\$ million. This implies that only a handful of large companies can train the models the whole CL community depends on. That said, the success of such models is undeniable: in numerous and different NLP tasks fine-tuned LMs are now achieving new state-of-the-art results. Their success has pushed researchers to investigate what [kind of linguistic knowledge](#) is actually encoded in such models, and also what does it mean for these models to "[understand language](#)" or on their [generalisation capabilities](#). More generally,

the interpretability of neural network representations and computations has seen the interest of a [growing research community](#). Interests in this area are also pushed forward by the interpretability and explainability requirements of the EU GDPR regulation.

Additional aspects that are getting an increasing interest and attention also in the CL and NLP communities concern the relationship between data, algorithms and ethics. A prolific research area focuses on understanding biases in data that can be transferred to systems. For instance, Bolukbasi, Chang, Zou et al. ([2016](#)) have shown that word embeddings trained on Google News are not free from female/male gender stereotypes.

We close this last section with a brief account on the relevance of CL/NLP techniques in machine translation. While in the previous statistical approach, it was common for different NLP applications to use different computational techniques, there is a convergent tendency in the current neural approach, so that the core architecture, e.g. Transformer (Vaswani, Noam, Parmar *et al.* [2017](#)), is shared across different applications. This has brought researchers that work in different NLP applications (including machine translation) closer to each other, fostering cross-fertilisation of ideas. Most machine translation systems built nowadays use only plain text, thus using implicit linguistic knowledge. However, augmenting these systems with explicit linguistic knowledge is useful in some scenarios, e.g. morphological segmentation for agglutinative languages (Ataman, Negri, Turchi *et al.* [2017](#)) and syntactic information in low-resource settings (Li, Xiong, Tu *et al.* [2017](#)).

[back to top](#)

## References



Abzianidze, Lasha; Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen & Johan Bos. 2017. "The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations" @ Lapata, Mirella; Phil Blunsom & Alexander Koller (eds.) 2017. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 242-247. Valencia: ACL. [\[+info\]](#) [\[quod vide\]](#)

Artetxe, Mikel; Gorka Labaka, Eneko Agirre & Kyunghyun Cho. 2018. "Unsupervised neural machine translation" @ *Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018)*. Vancouver. [\[+info\]](#) [\[quod vide\]](#)

Ataman, Duygu; Matteo Negri, Marco Turchi & Marcello Federico. 2017. "Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English" @ *The Prague Bulletin of Mathematical Linguistics* 108, 331-342. [\[+info\]](#) [\[quod vide\]](#)

Baldwin, Timothy & Valia Kordoni. 2011. "The Interaction between Linguistics and Computational Linguistics" @ *Linguistic Issues in Language Technology* 6/1, 1-6. [\[+info\]](#) [\[quod vide\]](#)

Baroni, Marco & Alessandro Lenci. 2010. "Distributional Memory: A General Framework for Corpus-Based Semantics" @ *Computational Linguistics* 36/4, 673-721. [\[+info\]](#) [\[quod vide\]](#)

Biemann, Chris. 2011. *Structure Discovery in Natural Language* Berlin: Springer. [\[+info\]](#)

Bolukbasi, Tolga; Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama & Adam T. Kalai. 2016. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings" @ *Advances in Neural Information Processing Systems* 29. [\[+info\]](#) [\[quod vide\]](#)

\* Eisenstein, Jacob. 2019. *Introduction to Natural Language Processing*. Cambridge: The MIT Press. [\[+info\]](#)

Fillmore, Charles J. 1968. "The case for case" @ Bach, Emmon & Robert T. Harms (eds.) 1968. *Universals in Linguistic Theory*, 1-88. New York: Holt, Rinehart & Winston. [\[+info\]](#) [\[quod vide\]](#)

Firth, John R. 1957. "A synopsis of linguistic theory 1930-1955" @ *Studies in linguistic analysis*, 1-32. Oxford: Blackwell. [\[+info\]](#)

Gardent, Claire & Bonnie Webber. 1988. "Describing discourse semantics." @ *Fourth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+ 4)*, Philadelphia: ACL, 50-53. [\[+info\]](#) [\[quod vide\]](#)

\* Goldberg, Yoav. 2017. *Neural Network Methods in Natural Language*. San Rafael: Morgan & Claypool Publishers. [\[+info\]](#)

Gómez Pérez, José Manuel; Ronald Denaux & Andrés García Silva. 2020. *A Practical Guide to Hybrid Natural Language Processing*. Berlin: Springer. [\[+info\]](#)

Grosz, Barbara J & Candace L. Sidner. 1986. "Attention, Intentions, and the Structure of Discourse" @ *Computational Linguistics* 12/3, 175-204. [\[+info\]](#) [\[quod vide\]](#)

\* Jurafsky, Dan & James H. Martin. (in progress). *Speech and Language Processing*. 3rd edition. Stanford: Stanford University. [\[+info\]](#)

Klavans, Judith L. & Philip Resnik (eds.) 1996. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge: The MIT Press. [\[+info\]](#)

Li, Junhui; Deyi Xiong, Zhaopeng Tu, Muhua Zhu, Min Zhang & Guodong Zhou. 2017. "Modeling source syntax for neural machine translation" @ *Proceedings of ACL*. [\[+info\]](#) [\[quod vide\]](#)

Mann, William C. & Sandra A. Thompson. 1988. "Rhetorical structure theory: Toward a functional theory of text organization" @ *Text* 8/3, 243-281. [\[+info\]](#)

\* Manning, Christopher & Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press. [\[+info\]](#)

Mikolov, Tomas; Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality" @ *Advances in Neural Information Processing Systems* 26. [\[+info\]](#) [\[quod vide\]](#)

Navigli, Roberto. 2009. "Word sense disambiguation: A survey" @ *ACM computing surveys (CSUR)* 41/2. [\[+info\]](#) [\[quod vide\]](#)

Pedersen, Ted. 2008. "Last Words: Empiricism Is Not a Matter of Faith" @ *Computational Linguistics* 34/3, 465-470. [\[+info\]](#) [\[quod vide\]](#)

Schubert, Lenhart. 2014. "Computational Linguistics" @ *Stanford Encyclopedia of Philosophy*. Stanford: Stanford University. [\[+info\]](#)

Straka, Milan & Jana Straková. 2017. "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe" @ *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88-99. Vancouver: ACL. [\[+info\]](#) [\[quod vide\]](#)

Stratos, Karl; Michael Collins & Daniel Hsu. 2016. "Unsupervised part-of-speech tagging with anchor hidden markov models" @ *Transactions of the Association for Computational Linguistics* 4, 245-257. [\[quod vide\]](#)

Vaswani, Ashish; Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. "Attention is All you Need" @ *Advances in Neural Information Processing Systems* 30, 5998-6008. [\[+info\]](#) [\[quod vide\]](#)

Webber, Bonnie Lynn. 1988. "Tense as Discourse Anaphor" @ *Computational Linguistics* 14/2, 61-73. [\[+info\]](#) [\[quod vide\]](#)

## Credits



### **Antonio Toral Ruiz**

Assistant Professor in Language Technology at the University of Groningen. He holds a PhD in Computational Linguistics from the University of Alicante and has carried out research in the area of Machine Translation (MT) since 2010. His research interests include the application of MT to literary texts, MT for under-resourced languages and the analysis of translations produced by machines and humans.



### **Tommaso Caselli**

Assistant Professor in Computational Semantics at the University of Groningen. He received his PhD in Computational Linguistics from the University of Pisa. His main research areas include event extraction and representation, temporal processing, and storyline extraction. He is one of the initiators of the Event and Stories in the News workshop series, and is currently working on the development of computational models and NLP tools to extract plot structures from news.



Licensed under the [Creative Commons Attribution Non-commercial License 4.0](#)

[Asociación Ibérica de Estudios de Traducción e Interpretación \(AIETI\)](#)