# *Introducción al PLN | ProgPLN*

*Víctor Peinado*

*30 de septiembre - 6 de octubre de 2016*

## *Referencias*

- *Intro to NLP*.[1]
- *What is Computational Linguistics?*[2]
- *Perspectives in Computational Linguistics*.[3]
- *The Stupidity of Computers*.[4]
- *An Inside Update on NLP*.[5]

[1] Introduction to NLP `http://futurewavewebdevelopment.com/wp/2016/08/brucemwhealton/introduction-to-natural-language-processing-nlp-20`

[2] What is Computational Linguistics `http://www.coli.uni-saarland.de/~hansu/what_is_cl.html`

[3] Perspectives in Computational Linguistics `http://www.linguisticsociety.org/content/computers-and-languages`

[4] The Stupidity of Computers `https://nplusonemag.com/issue-13/essays/stupidity-of-computers/`

[5] An Inside Update on NLP `https://breakthroughanalysis.com/2016/06/23/jbnlp/`

## *¿Qué es el PLN?*

El Procesamiento del Lenguaje Natural (PLN)[6] es el estudio científico del lenguaje desde un punto de vista computacional. Es un área claramente multidisciplinar que aglutina lingüística, ingeniería, inteligencia artificial, informática, psicología, etc.

El PLN se interesa en proporcionar modelos computacionales para describir, modelar o reproducir distintos fenómenos lingüísticos. Tradicionalmente, estos modelos han tenido dos aproximaciones diferentes:

1. sistemas basados en conocimiento: en problemas que podemos modelar, proporcionamos conocimiento lingüístico formalizado y las máquinas actuán aplicando reglas.
2. sistemas basados en estadística: en problemas que son costosos o no podemos modelar, proporcionamos ingentes cantidades de datos (colecciones de documentos) y dejamos que la máquina cree el modelo a partir del cálculo de probabilidades y la detección de patrones de uso.

El objetivo último del PLN reside en conseguir que los ordenadores analicen el lenguaje natural, lo comprendan y sean capaces de extraer significado de manera (que parezca) inteligente y resulte útil.

[6] En inglés, *Natural Language Processing* (NLP) o mejor #NLProc. La disciplina recibe otros nombres, como *Human Language Technologies* (HLT), tecnologías de la lengua, ingeniería lingüística, lingüística computacional, etc.

### *Tareas típicas del PLN*

Una buena manera de conocer los temas que trata un área de investigación es revisar el calendario de los congresos más importantes:[7]

[7] NLP Conferences Calendar `http://cs.rochester.edu/~omidb/nlpcalendar/`

- ACL 2016: *call for papers*[8] y programa[9].
- EMNLP 2016: *call for papers*[10] y programa[11].
- COLING 2016: *call for papers*[12] y programa[13]
- SEPLN 2016: *call for papers*[14]y programa[15]

De este modo, podemos identificar algunas de las tareas más comunes del área:

- Desambiguación semántica (*word sense disambiguation*) y reconocimiento de entidades (*named entities recognition*).
- Análisis morfo-sintáctico (*PoS tagging/parsing*)
- Traducción automática (*machine translation*): Google Translate
- Extracción de información (*information extraction*): TripIt y los bundles de Inbox
- Reconocimiento del habla (*automatic speech reconition*) y síntesis de voz (*speech synthesis*): Google Voice Search
- Recuperación de información (*information retrieval*): Google Search, Bing y Wolfram | Alpha
- Resumen automático (*automatic summarization*) y generación automática de textos: Quakebot y Automated Insights
- Búsqueda de respuestas (*question answering*): tímidos intentos de Google o Bing y, sbre todo, Watson
- Análisis de opiniones (*sentiment analysis*): Bitext y Atribus
- Comprensión del lenguaje natural (*natural language understanding*): Siri, Google Now y Cortana

*Problemas resueltos y cuestiones abiertas*

## *¿Por qué es tan difícil el PLN?*

El lenguaje natural es eminentemente **ambiguo**. Esta es la principal diferencia entre lenguas naturales y lenguajes artificiales.

Esta ambigüedad existe a varios niveles:

- ambigüedad fonética y fonológica: *vaca/baca*, *casa/caza*, *has sido tú/has ido tú*
- ambigüedad morfológica: *casa, beso, río, bajo*
- ambigüedad sintáctica: *Ayer me encontré a tu padre corriendo*

[8] ACL 2016 CFP http://acl2016.org/index.php?article%20id=9
[9] ACL 2016 CFP http://acl2016.org/index.php?article%20id=9
[10] EMNLP 2016 CFP http://www.emnlp2016.net/call.html
[11] EMNLP 2016 Program http://www.emnlp2016.net
[12] COLING 2016 CFP http://coling2016.anlp.jp/cfp/
[13] COLING 2016 Program http://coling2016.anlp.jp/
[14] SEPLN 2016 CFP http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=51713&copyownerid=85257
[15] SEPLN 2016 Program http://www.congresocedi.es/es/sepln#tabs7

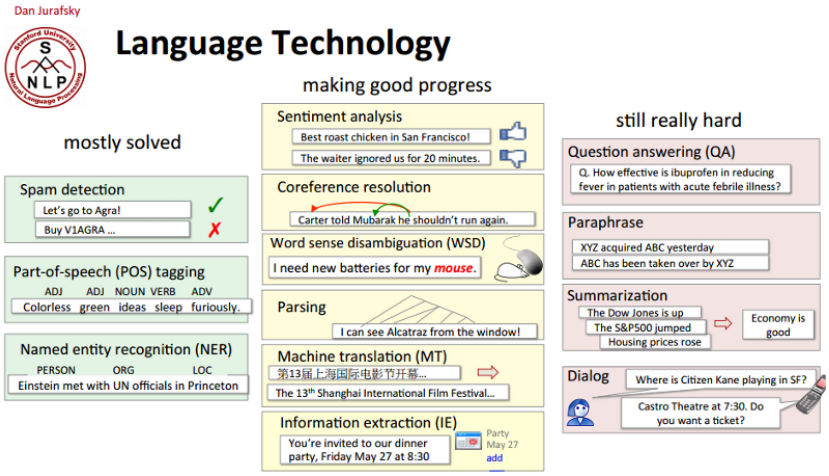


Figure 1: Language Technologies Progress, according to Stanford NLP

- ambigüedad semántica: *banco*, *pie*, etc.
- ambigüedad de discurso: correferencia, resolución de anáforas.

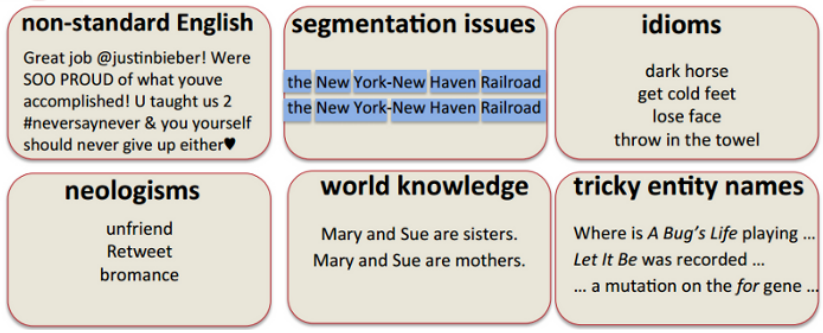


Figure 2: Language Technologies Difficulties, according to Stanford NLP

Según la ACL (*Association for Computational Linguistics*): *Computational Linguistics, or Natural Language Processing (NLP), is not a new field*[16], sin embargo no es sencillo definir los límites de la disciplina. Así que podemos considerarla como un conjunto de problemas relacionados con fenómenos lingüísticos y una amalgama de soluciones computacionales, de distinto tipo dependiendo del origen del investigador.

Según xkcd,[17] los lingüistas computacionales han vivido muy bien hasta ahora vendiendo la moto, así que no se merecen más que nos metamos con ellos :-) [18]

[16] ACL FAQ `http://www.aclweb.org/aclwiki/index.php?title=Frequently_asked_questions_about_Computational_Linguistics`

[17] http://www.xkcd.org/114/

[18] http://www.explainxkcd.com/wiki/index.php/114:_Computational_Linguists

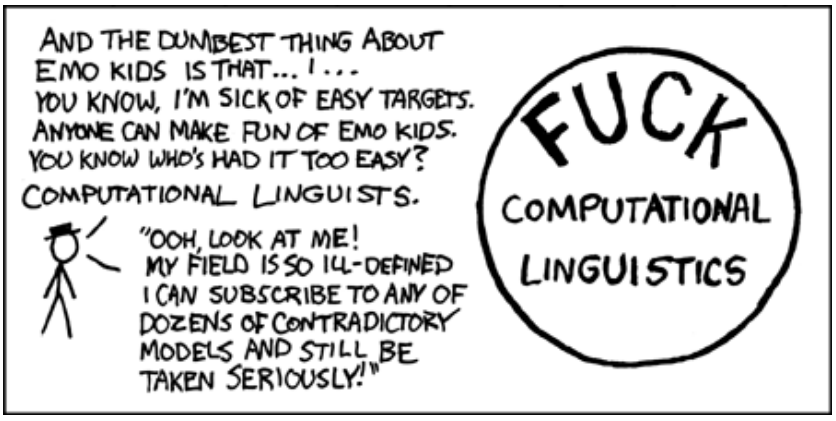


Figure 3: Computational Linguistics