Plan TL
IBERLEGAL

Ontology Engineering Group

Extracting terminologies in the legal domain: a syntactic pattern-based approach for Spanish

Elena Montiel-Ponsoda
Profesor Contratado Doctor
Universidad Politécnica de Madrid

11 de diciembre de 2019

Índice

Website http://nlp.linkeddata.es

Agradecimientos

Plan TL
IBERLEGAL

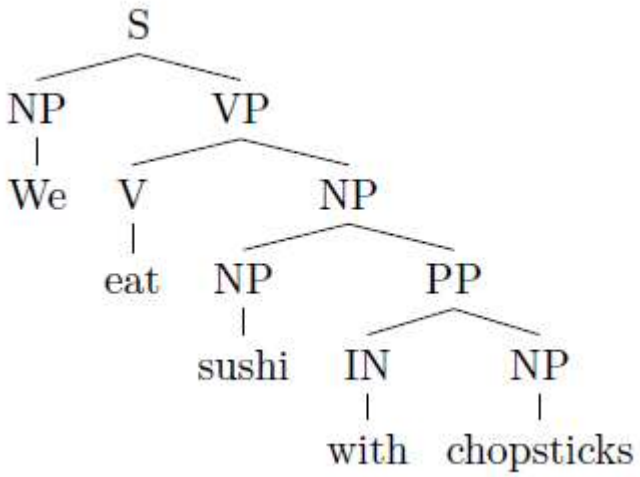
Lynx

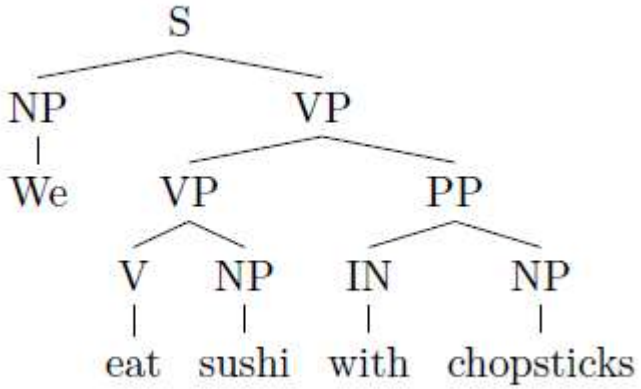
Prêt-à-LLOD

Ontology Engineering Group

1. Patrones sintácticos

Ambigüedad: dos interpretaciones de la misma frase





1. Patrones sintácticos. Novedad

Plan TL IBERLEGAL

Las técnicas de *deep learning* logran la interpretación más habitual

Announcing SyntaxNet: The World's Most Accurate Parser Goes Open Source

Thursday, May 12, 2016

Posted by Slav Petrov, Senior Staff Research Scientist

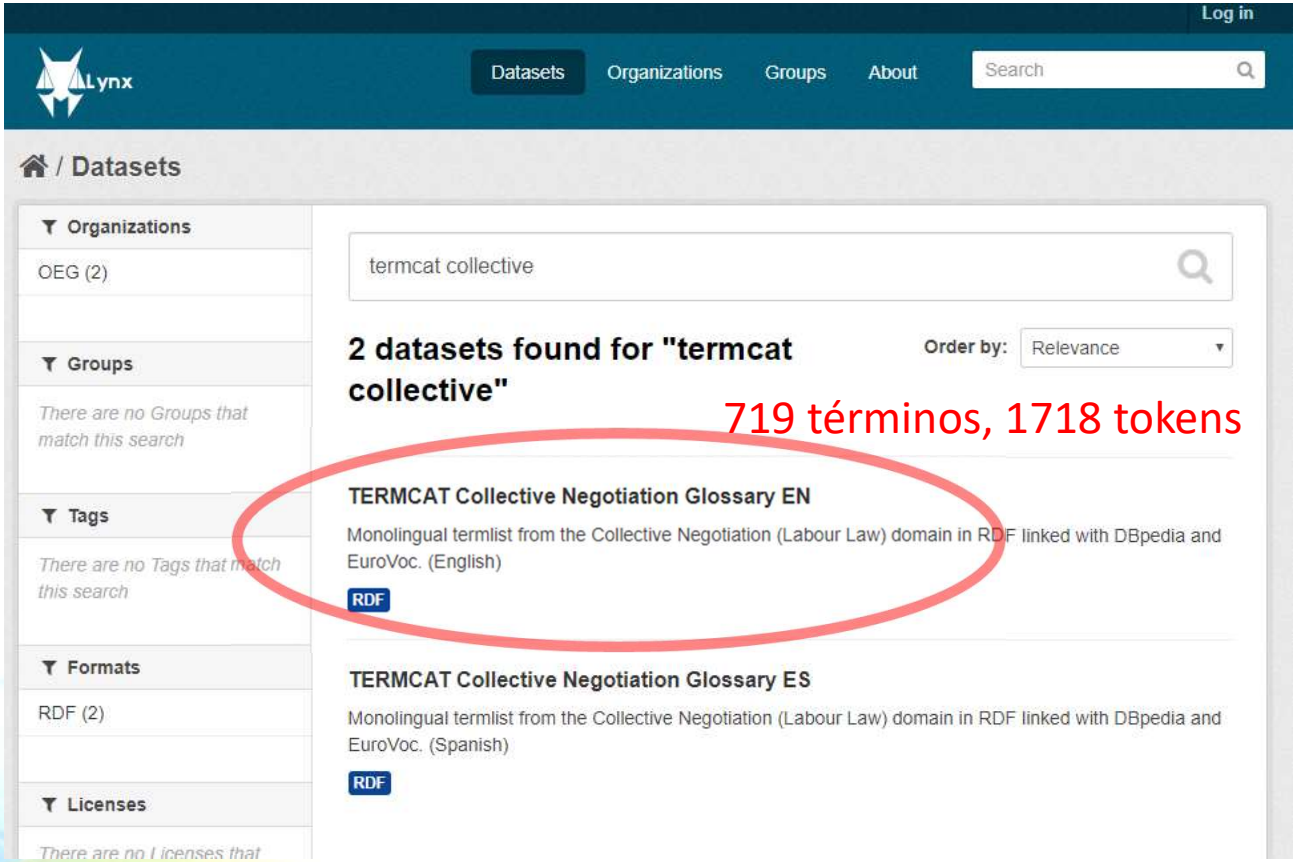
At Google, we spend a lot of time thinking about how computer systems can read and understand human language in order to process it in intelligent ways. Today, we are excited to share the fruits of our research with the broader community by releasing SyntaxNet, an open-source neural network framework implemented in TensorFlow that provides a foundation for Natural Language Understanding (NLU) systems. Our release includes all the code needed to train new SyntaxNet models on your own data, as well as *Parsey McParseface*, an English parser that we have trained for you and that you can use to analyze English text.

https://ai.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

2. Patrones sintácticos del español jurídico

Plan TL
IBERLEGAL

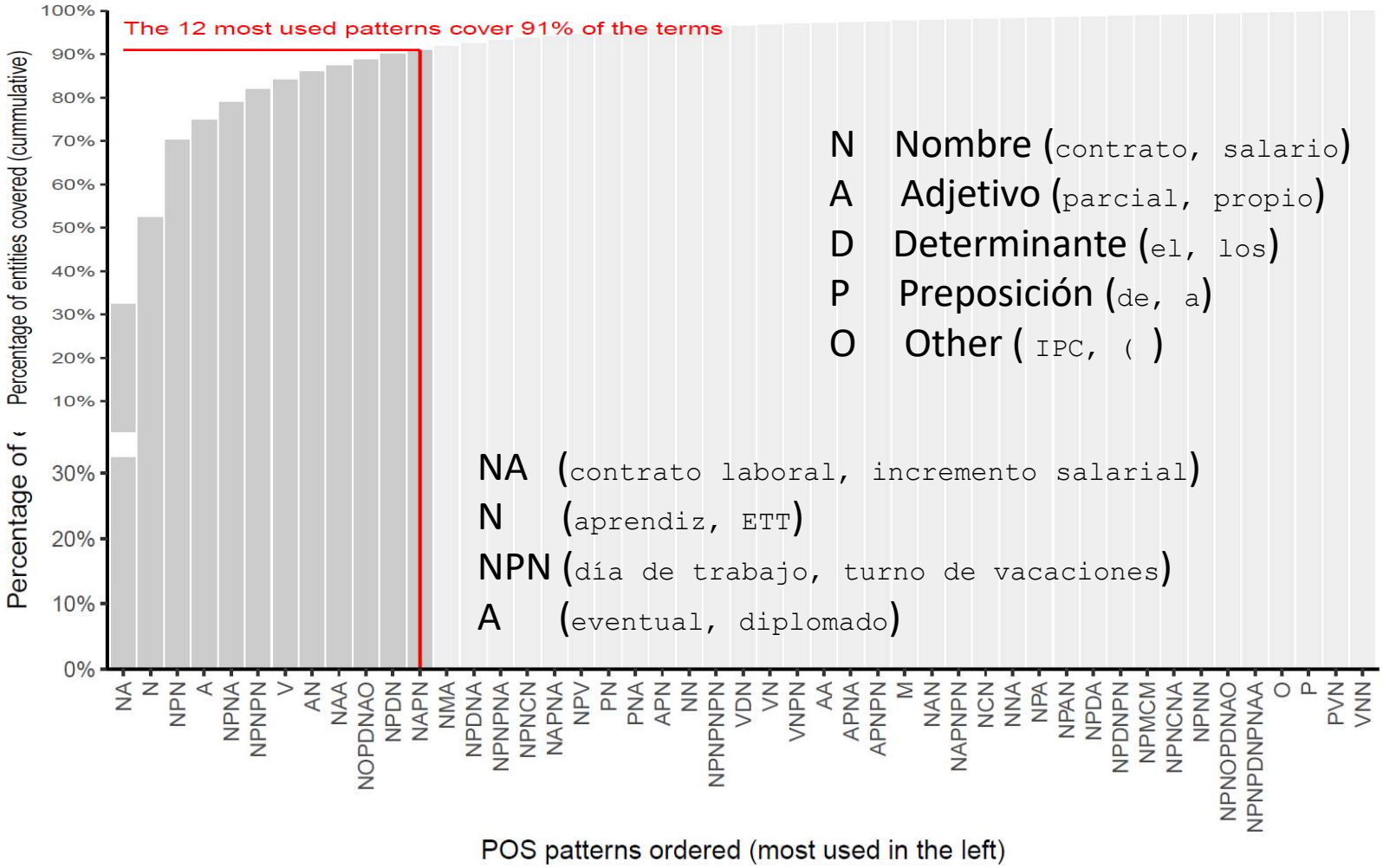
Análisis de terminología jurídica de termcat
centre de terminologia



2. Patrones sintácticos del español jurídico

Plan TL
IBERLEGAL

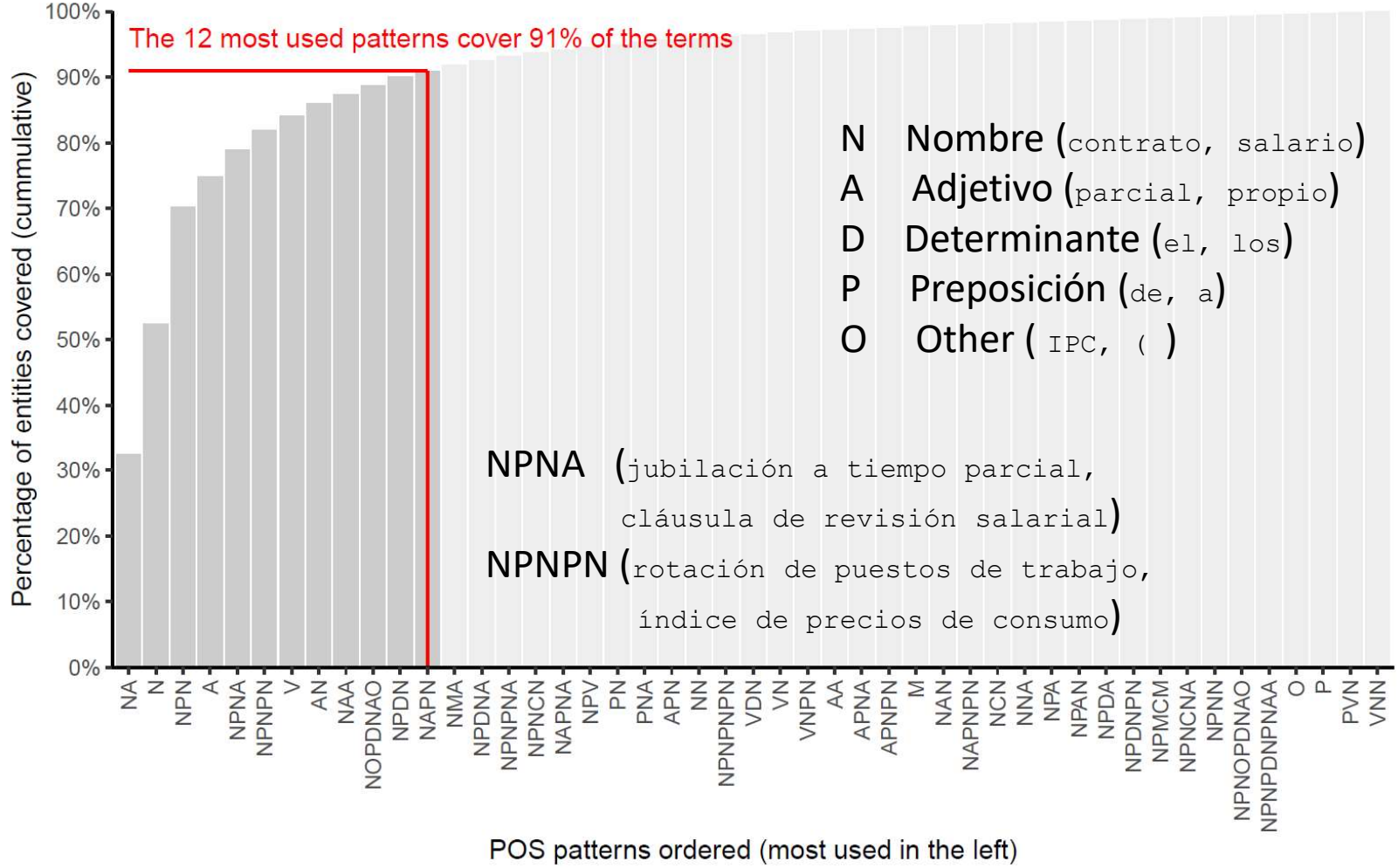
Análisis de terminología jurídica de termcat centre de terminologia



2. Patrones sintácticos del español jurídico

Plan TL
IBERLEGAL

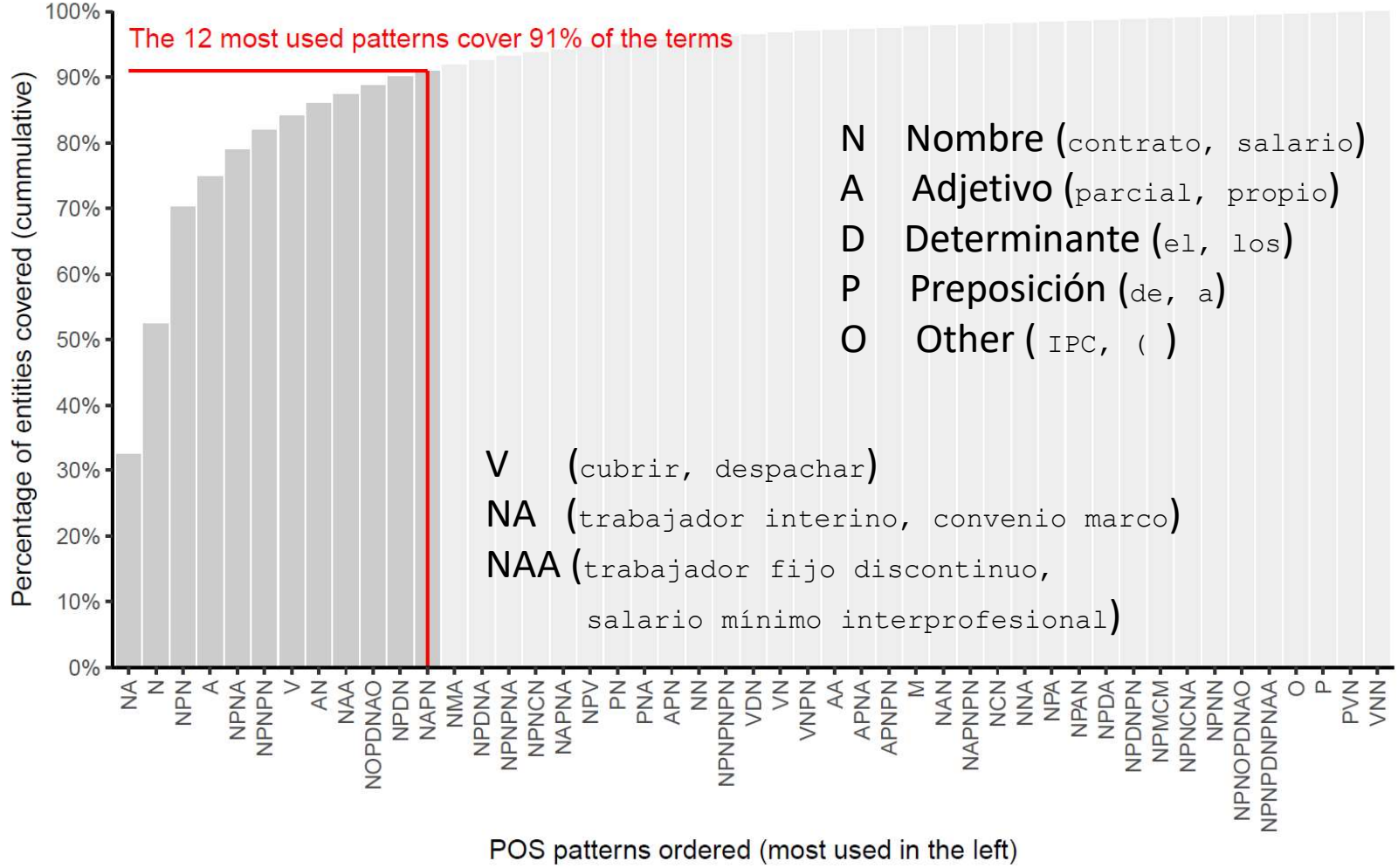
Análisis de terminología jurídica de termcat centre de terminologia



2. Patrones sintácticos del español jurídico

Plan TL
IBERLEGAL

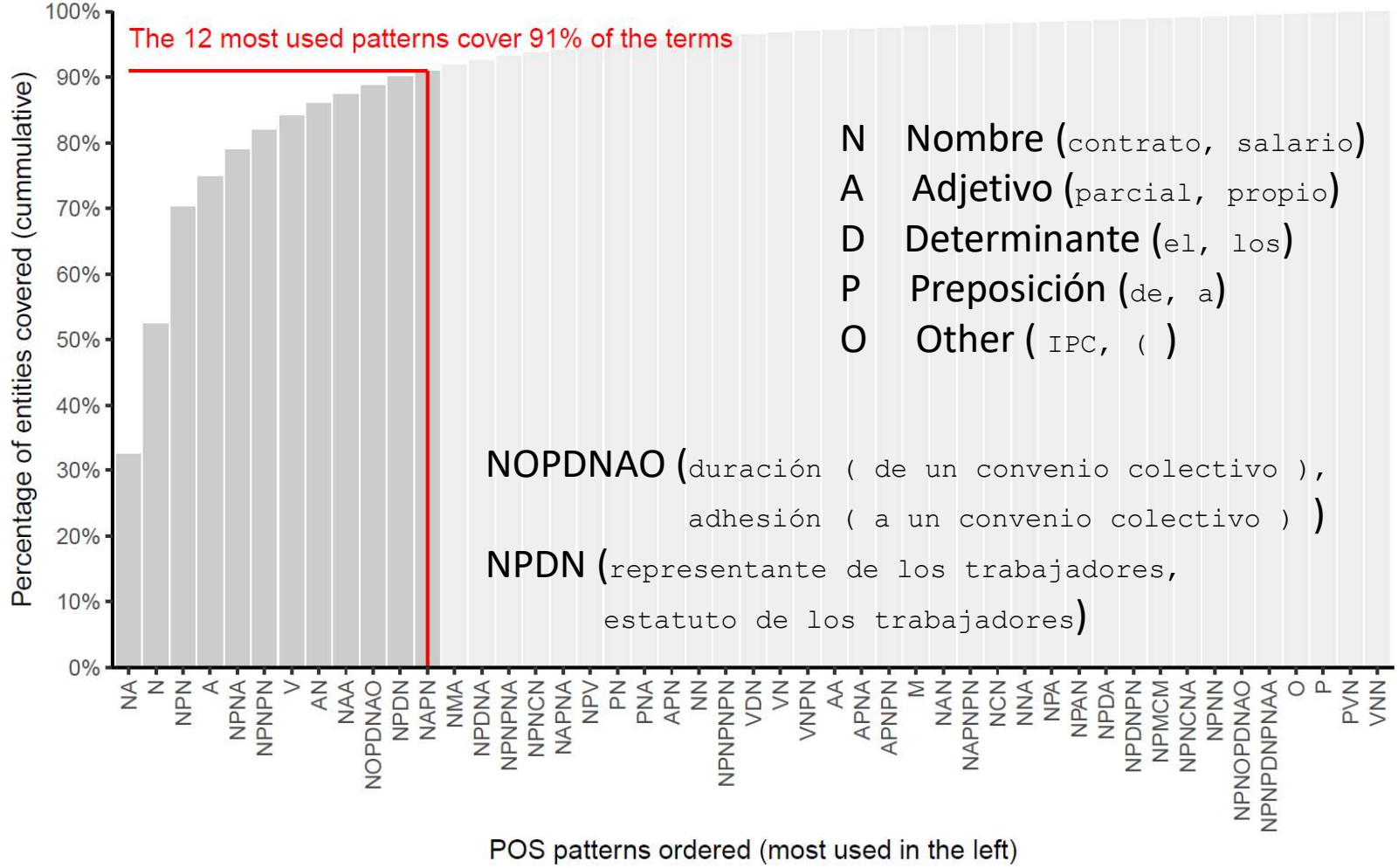
Análisis de terminología jurídica de termcat centre de terminologia



2. Patrones sintácticos del español jurídico

Plan TL
IBERLEGAL

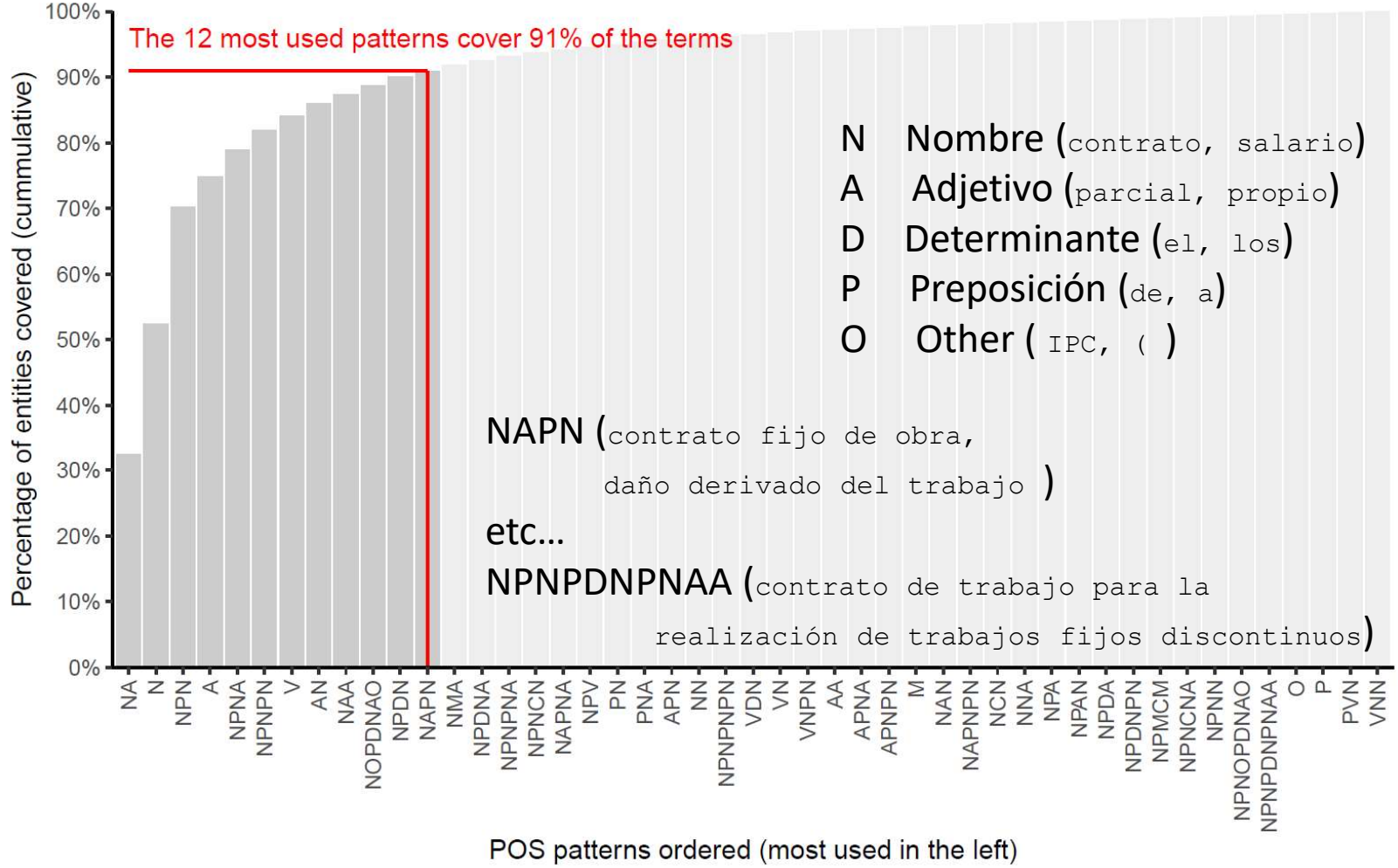
Análisis de terminología jurídica de termcat centre de terminologia



2. Patrones sintácticos del español jurídico

Plan TL
IBERLEGAL

Análisis de terminología jurídica de termcat centre de terminologia



N Nombre (`contrato, salario`)
A Adjetivo (`parcial, propio`)
D Determinante (`el, los`)
P Preposición (`de, a`)
O Other (`IPC, (`)

NAPN (`contrato fijo de obra, daño derivado del trabajo`)

etc…

NPNPDNPNAA (`contrato de trabajo para la realización de trabajos fijos discontinuos`)

3. Extracción de terminología jurídica

Aplicamos los patrones identificados en termcat centre de terminologia a un corpus jurídico (convenios colectivos)

- Patrones generados por una ***expresión regular***. Ejemplo clásico:

  Phrasal Noun (sintagma nominal) en inglés: (A|N)*N(PD*(A|N)*N)*

  Proviene de esta gramática:

$$BaseNP \rightarrow (Adj|Noun)^* \ \ Noun$$
$$PrepPhrase \rightarrow Prep \ \ Det^* \ \ BaseNP$$
$$NounPhrase \rightarrow BaseNP \ \ PrepPhrase^*$$

  N, AN, NN, AAN, NNN,…
  PDN, PDAN, PDDN, PDDAN,…
  NPDN, ANPDN, ANPDAN,…

  Crece muy rápido:
  - Patrón de 1 elemento: N
  - Patrones de 2 elementos: AN y NN (2 patrones)
  - Patrones de 3 elementos: AAN, ANN, NAN, NNN, NPN (5 patrones)
  - Patrones de hasta 5 elementos: … ¡¡55 patrones!!

3. Extracción de terminología jurídica

Plan TL
IBERLEGAL

- Patrones generados por una *expresión regular*

  Sintagma Nominal en español: N(A|N)*(PD*N(A|N)*)*

  Resultados: identifica patrones muy largos, pero genera casos inválidos (e.g. `quince días, viernes 18 de enero de 2008`)

- Patrones más importantes (12 primeros de TERMCAT)

  Resultados: pierde patrones muy largos (e.g. `boletín oficial de la comunidad de madrid`)

- Todos los patrones de TERMCAT

  Resultados: pierde patrones muy largos y genera "ruido" (e.g. `de prevención, a partir, en http://www.boe.es`)

4. Conclusiones

- Generación de terminologías jurídicas y enriquecimiento de las existentes
  - Generación de un asistente para proponer candidatos (términos en español), que serán evaluados por humanos
    - Añadir nuevas terminologías (similares a TERMCAT) para “entrenar” el asistente
    - Añadir nuevos corpus para identificar nuevos candidatos
- Líneas de trabajo futuras: uso de técnicas de *deep learning*
    - Explicación de los resultados usando gramáticas formales

Plan TL
IBERLEGAL

Muchas gracias

www.PlanTL.es

PlanTecnologiasLenguaje@mineco.es