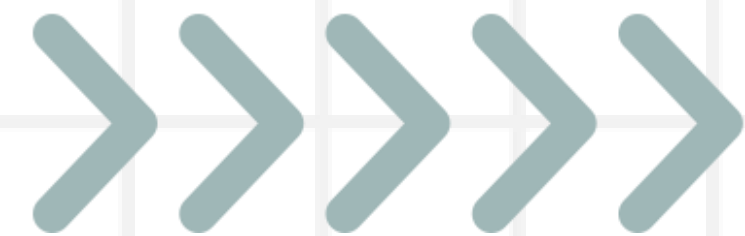




EVALUACIÓN DE ETIQUETADORES MORFOSINTÁCTICOS PARA EL ESPAÑOL





Índice

1. Introducción
2. HMM, en función del corpus
3. HMM
4. TnT: cómo afectan los métodos de suavizado en el tagger
5. Otros etiquetadores
6. Evaluación de otras herramientas: SpaCy y Stanza
7. Conclusiones



Introducción

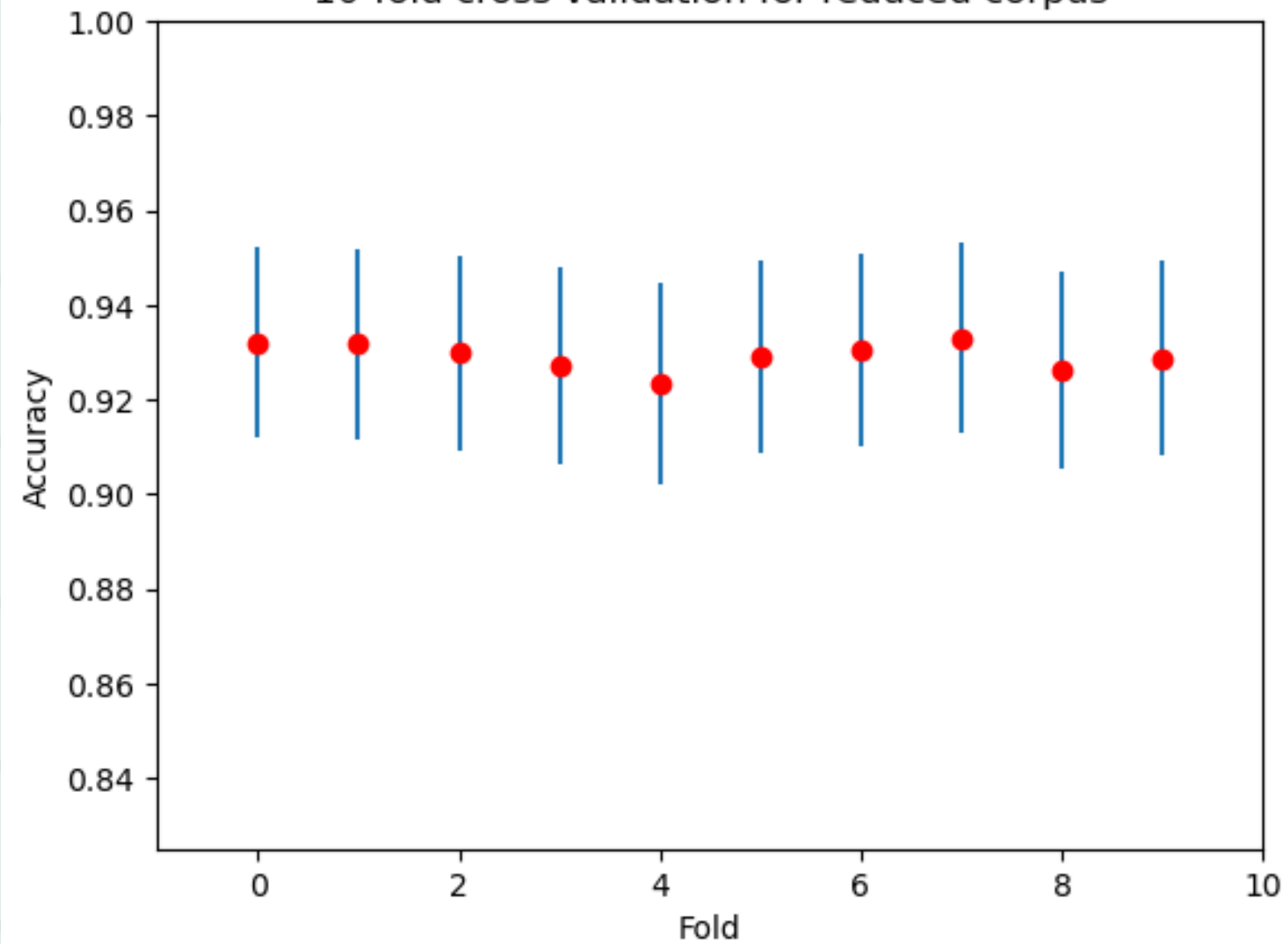
- ❑ **Objetivo:** Evaluación diferentes etiquetadores morfosintácticos en español, en función de diversos parámetros.
- ❑ **Corpus principal: cess_esp**
 - ❑ Reducido (66 etiquetas)
 - ❑ Completo (289 etiquetas)
- ❑ **Corpus Alicia** para evaluar el rendimiento de otras dos herramientas: Staza y SpaCy



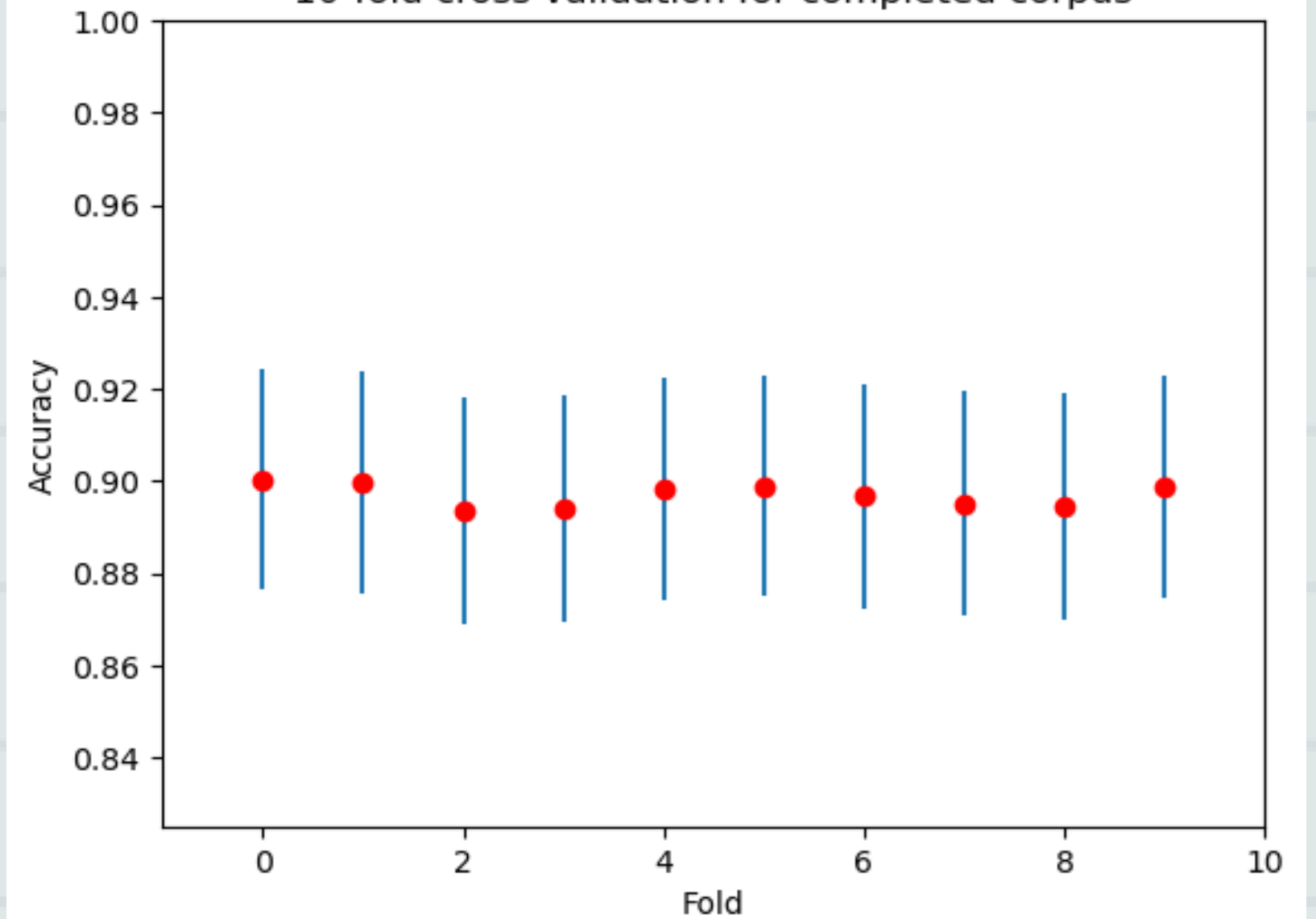
HMM

En función del número de etiquetas del corpus

10-fold cross validation for reduced corpus

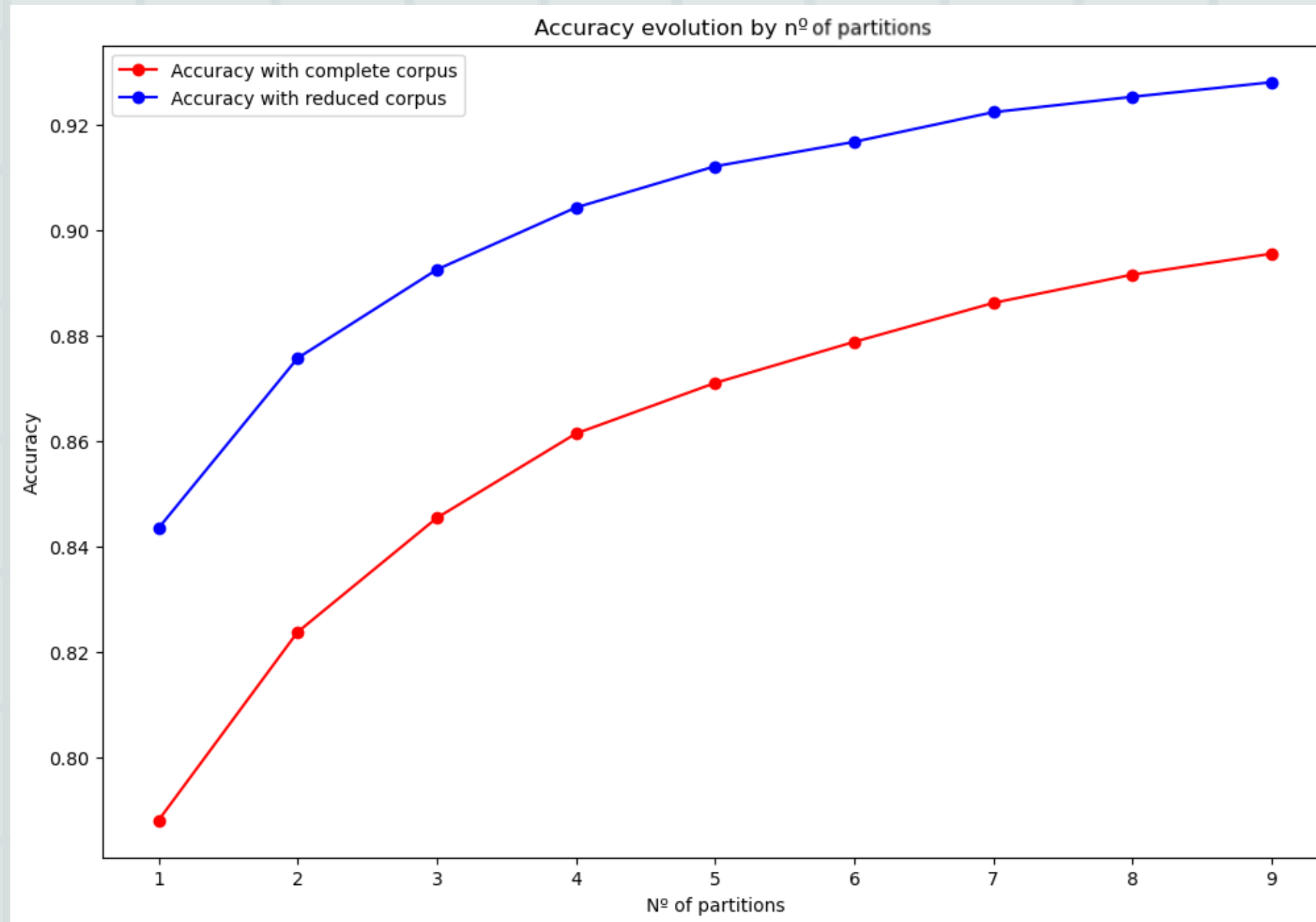


10-fold cross validation for completed corpus



HMM

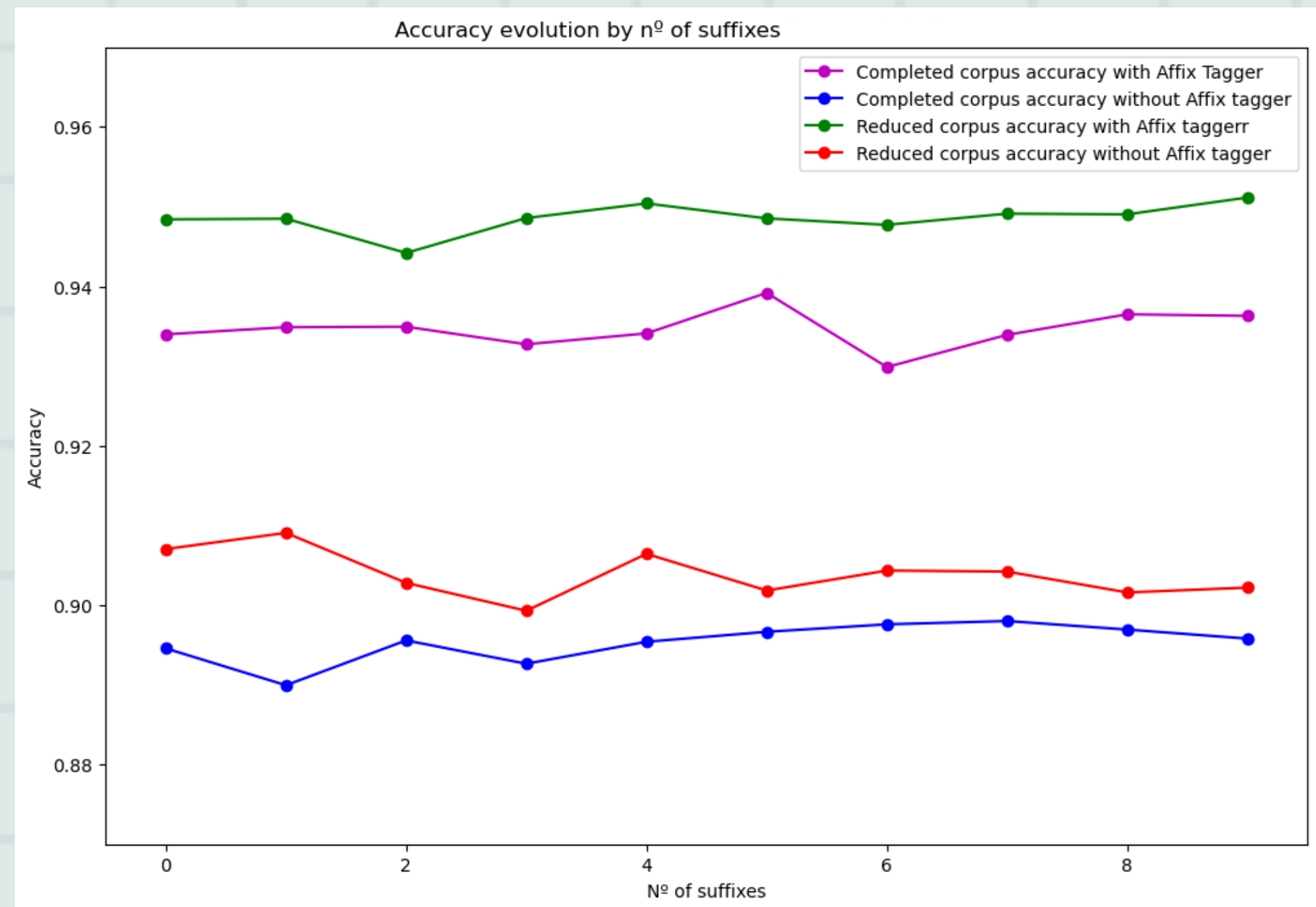
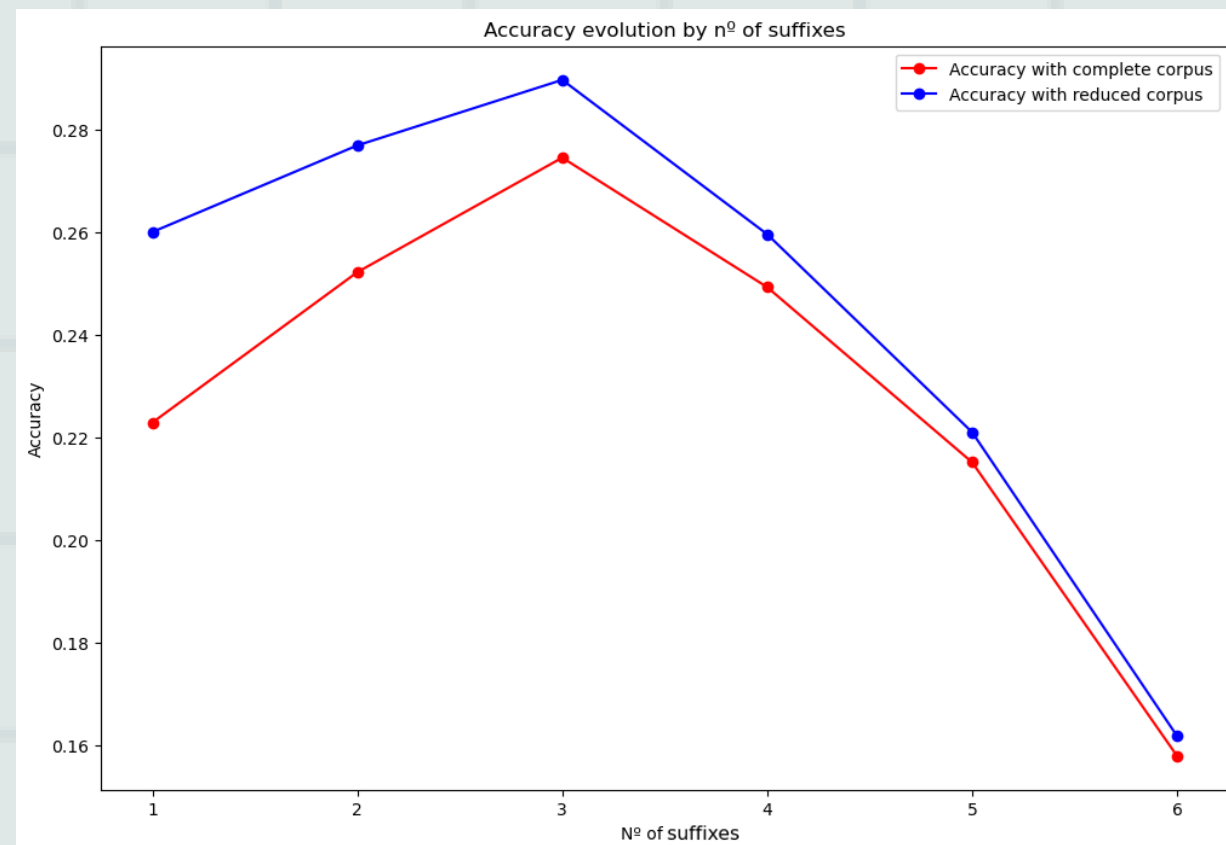
En función de la cantidad de datos de aprendizaje



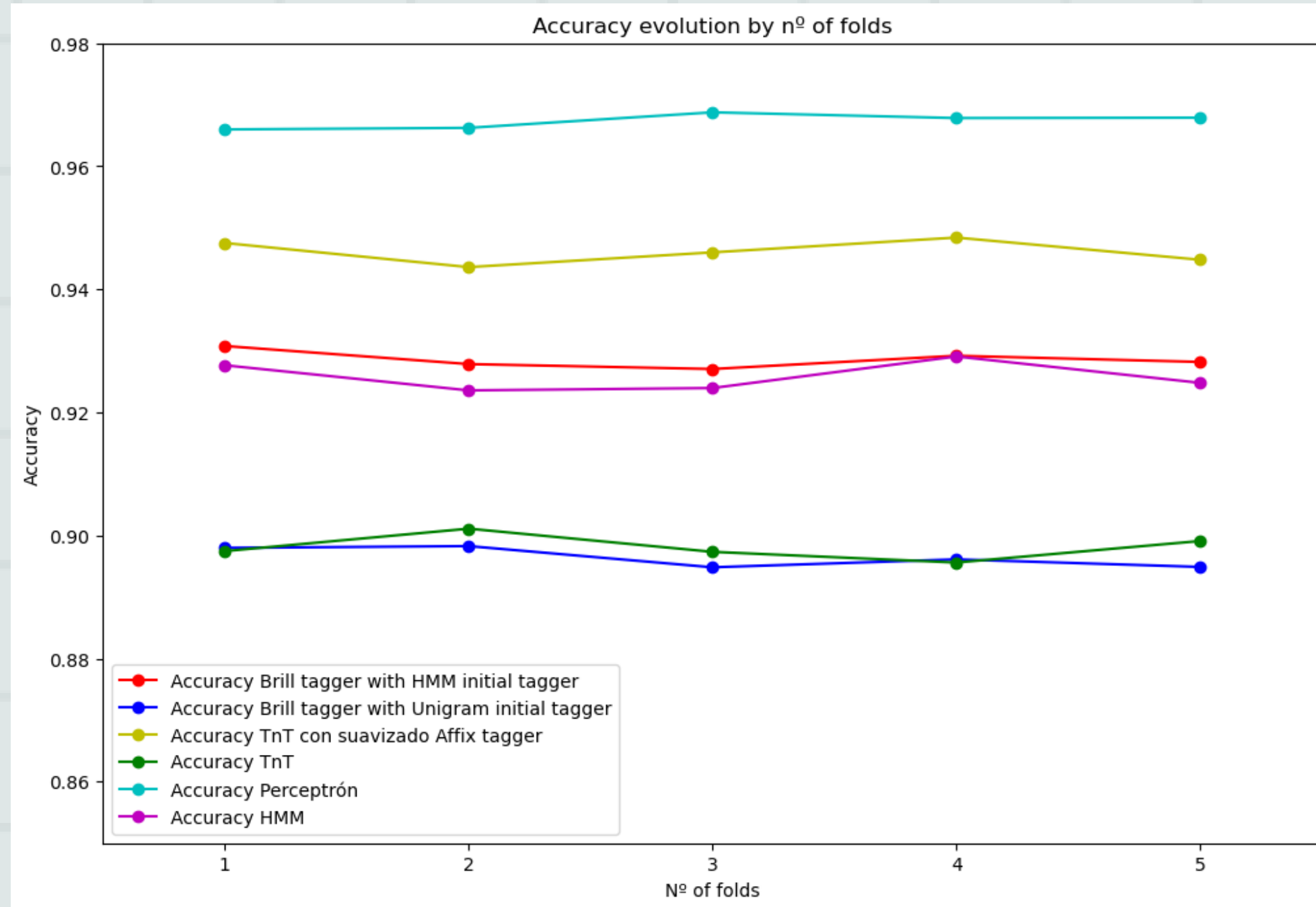
TnT

Incorporando Affix Tagger como método de suavizado

Longitud del sufijo óptimo = 3



Otros etiquetadores



Evaluación de otras herramientas



SpaCy

Stanza

- ❑ Ambas ofrecen **funcionalidades en español**
- ❑ **Instalación sencilla** (SpaCy más simple)
- ❑ **Documentación extensa** en línea y una **amplia comunidad** de soporte
- ❑ **Etiquetado** morfosintáctico con **resultados precisos**
- ❑ Método **NER** con **resultados muy pobres**

Conclusiones

- **Rendimiento mejora al aumentar la cantidad de datos de aprendizaje**
- **Reducción de etiquetas mejora los resultados**
- **No hay diferencias significativas entre los folds de un mismo tagger**
- **Perceptrón es el etiquetador más efectivo**
- **Métodos de suavizado mejoran el rendimiento del etiquetador**
- **SpaCy y Stanza son sencillas y eficientes para el etiquetado**

¡Gracias!

¿Alguna pregunta?

