# Propuestas de Temas

Arturo Curiel
me@arturocuriel.com

1 de octubre de 2018

# Contenidos

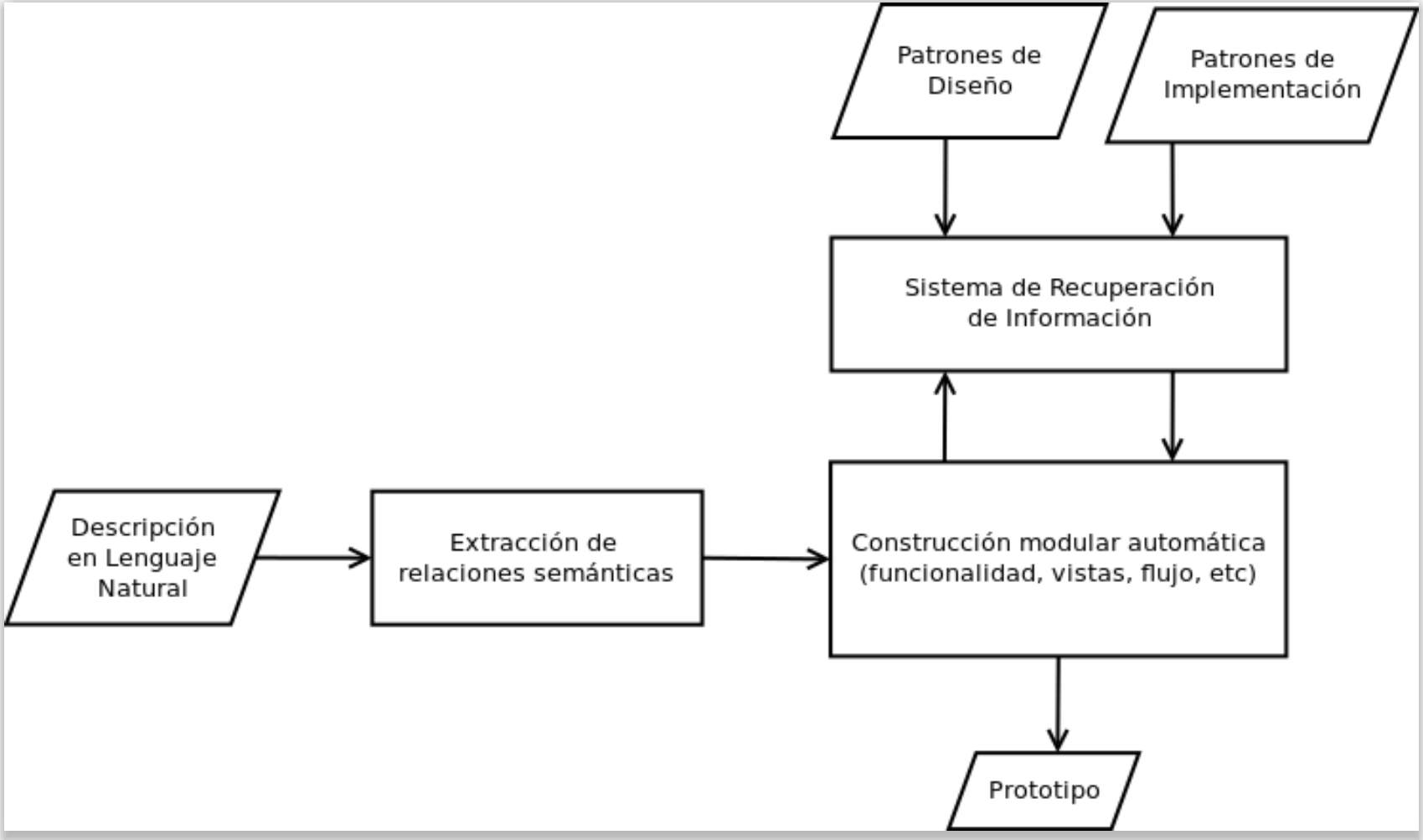
# Contenidos

# Especificación de Software en Lenguaje Natural

## Problema

Desarrollar métodos de análisis de requerimientos basados en lenguaje natural.

# Especificación de Software en Lenguaje Natural




Arquitectura Propuesta

# Especificación en Lenguaje Natural

1. Extracción de Relaciones Semánticas
   - Análisis discursivo
   - Construcción no supervisada de ontologías
   - Extracción de términos
2. Sistema de IR
   - Representación e indexado
   - Criterios de relevancia
   - Algoritmos de búsqueda eficaz/eficiente
3. Construcción Modular
   - Interpretación semántica
   - Tests unitarios automáticos
   - Visualización

# Contenidos

# Análisis contextual de interfaz

## Problema

Inferir información contextual a partir del análisis pragmático de los participantes de una conversación alrededor de una interfaz.

# Análisis contextual de interfaz

- Tenemos un dispositivo que viola privacidad.
  - Siempre ve, siempre escucha.
- Usar información lingüística para acotar contexto de interacción.
  - Lo que hacen Alexa y Google para venderte anuncios.
  - *e.g.* estas hablando de comprar computadoras y te aparecen publicidad de Apple o Dell

# Análisis contextual de interfaz

En una primera instancia:

- Recuperación de transcripciones durante la interacción con una interfaz.
- Análisis del lenguaje especializado
  - Entidades nombradas (elementos de la interfaz).
  - Algoritmos para la identificación de relaciones lexico-sintácticas.
- Aplicarlo a interfaces tangibles.
  - Se espera un demo: recuperación de elementos dinámicos de la interfaz.

# Contenidos

# Interfaces de Voz

## Problema

Desarrollar un *framework* para la interpretación de comandos de voz en español mexicano.

# Interfaces de Voz

- Análisis fonético/fonológico.
  - Colectar y etiquetar un corpus con información fonética.
  - Aplicar algoritmos de reconocimiento en señales.
- Descripción de comandos
  - Estructura de los comandos
  - Dominios de aplicación
  - Cuestiones de usabilidad

# Interfaces de Voz

Se espera:

- Una versión del framework con una buena tasa de reconocimiento.
- Una interfaz de ejemplo.
  - Interacción con una interfaz tangible.

# Contenidos

# Extracción de Información Legal

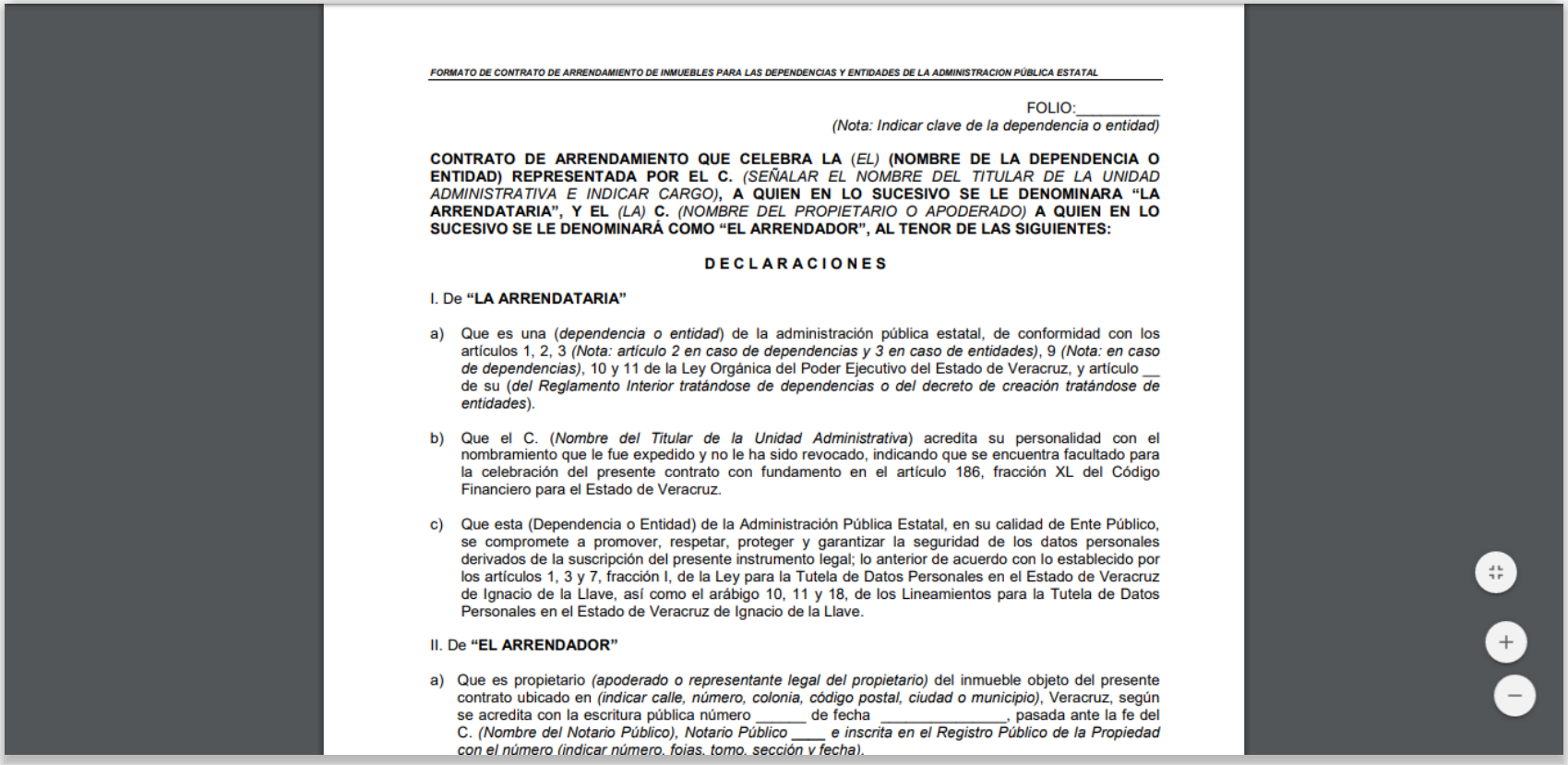
## Problema

Análisis de lenguaje legal con aplicaciones.

# Extracción de Información Legal

1. Búsqueda de información en documentos especializados.
   - Ayudar a un abogado a hacer investigación/llenar formatos.
2. Inferencia lógica de casos.
   - Analizar casos e inferir que leyes aplican para construir una demanda.

# Extracción de Información Legal



*FORMATO DE CONTRATO DE ARRENDAMIENTO DE INMUEBLES PARA LAS DEPENDENCIAS Y ENTIDADES DE LA ADMINISTRACION PÚBLICA ESTATAL*

FOLIO:_____
*(Nota: Indicar clave de la dependencia o entidad)*

**CONTRATO DE ARRENDAMIENTO QUE CELEBRA LA** *(EL)* **(NOMBRE DE LA DEPENDENCIA O ENTIDAD) REPRESENTADA POR EL C.** *(SEÑALAR EL NOMBRE DEL TITULAR DE LA UNIDAD ADMINISTRATIVA E INDICAR CARGO)***, A QUIEN EN LO SUCESIVO SE LE DENOMINARA "LA ARRENDATARIA", Y EL** *(LA)* **C.** *(NOMBRE DEL PROPIETARIO O APODERADO)* **A QUIEN EN LO SUCESIVO SE LE DENOMINARÁ COMO "EL ARRENDADOR", AL TENOR DE LAS SIGUIENTES:**

**D E C L A R A C I O N E S**

I. De **"LA ARRENDATARIA"**

a) Que es una (*dependencia o entidad*) de la administración pública estatal, de conformidad con los artículos 1, 2, 3 *(Nota: artículo 2 en caso de dependencias y 3 en caso de entidades)*, 9 *(Nota: en caso de dependencias)*, 10 y 11 de la Ley Orgánica del Poder Ejecutivo del Estado de Veracruz, y artículo __ de su (*del Reglamento Interior tratándose de dependencias o del decreto de creación tratándose de entidades*).

b) Que el C. (*Nombre del Titular de la Unidad Administrativa*) acredita su personalidad con el nombramiento que le fue expedido y no le ha sido revocado, indicando que se encuentra facultado para la celebración del presente contrato con fundamento en el artículo 186, fracción XL del Código Financiero para el Estado de Veracruz.

c) Que esta (Dependencia o Entidad) de la Administración Pública Estatal, en su calidad de Ente Público, se compromete a promover, respetar, proteger y garantizar la seguridad de los datos personales derivados de la suscripción del presente instrumento legal; lo anterior de acuerdo con lo establecido por los artículos 1, 3 y 7, fracción I, de la Ley para la Tutela de Datos Personales en el Estado de Veracruz de Ignacio de la Llave, así como el arábigo 10, 11 y 18, de los Lineamientos para la Tutela de Datos Personales en el Estado de Veracruz de Ignacio de la Llave.

II. De **"EL ARRENDADOR"**

a) Que es propietario *(apoderado o representante legal del propietario)* del inmueble objeto del presente contrato ubicado en *(indicar calle, número, colonia, código postal, ciudad o municipio)*, Veracruz, según se acredita con la escritura pública número _____ de fecha _____, pasada ante la fe del C. *(Nombre del Notario Público), Notario Público ____ e inscrita en el Registro Público de la Propiedad con el número (indicar número, foias, tomo, sección y fecha).*

Búsqueda de información en documentos especializados.

# Extracción de Información Legal

1. Extracción de información en contratos/demandas.
   - Entrenamiento de un módulo de OCR (la mayoría son scans).
   - Etiquetado de parrafos de documentos.
   - Representación en un espacio vectorial.
     * Entrenamiento de un clasificador supervisado.
     * Búsqueda de información.
   - Se espera:
     * Un demo que pueda etiquetar los parrafos de un documento legal con un resumen de su contenido.

# Extracción de Información Legal

2. Representación lógica de información legal.[a]
   - Análisis e indexación de leyes y sus reglamentos.
   - Representación lógica de casos.
     * Identificación de las partes.
     * Extracción de términos.
     * Extracción de relaciones léxico-sintácticas.
     * Búsqueda de información.
   - Se espera:
     * Obtener un demo que reciba una descripción y retorne que reglamentos aplican (ámbito de lo familiar).

[a] Ya hay un candidato.

# Contenidos

# Instrumentos de detección de depresión

## Problema

Queremos diseñar un instrumento de detección de depresión basado en preguntas abiertas.

# Instrumentos de detección de depresión

Se tiene:

- Las baterias de tests pueden ser largas y dificiles de aplicar. Si son niños es peor.
  - Pueden simplemente dejar de contestar.
  - Hay un adulto presente que puede sesgar los resultados.
- La mayoría no estan normalizadas para población mexicana.
- Las personas dan indicios en conversación.
  - Un buen entrevistador puede obtener mucha información.
  - Queremos estructurar esa información para obtener correlaciones.

# Instrumentos de detección de depresión

Se tiene:

- Acceso a una escuela secundaria en el Estado de México.
  - Se están aplicando tests y dando pláticas sobre detección de la depresión.
- Algunas preguntas de prueba basadas en entrevistas de control de confianza y baterias normalizadas.

# Instrumentos de detección de depresión

Se espera:

- Hacer el análisis de los datos.
  - Hacer análisis de patrones sobre transcripciones de entrevistas para correlacionarlos con los resultados de baterias normalizadas.
- Obtener un clasificador estadístico con esos datos para medir el riesgo de depresión desde una transcripción.
- Construir un demo de prueba.

# Contenidos

# Simplificación lectora

## Problema

Tomar un texto en español y simplificarlo sin perder información semántica.

# Simplificación lectora

- La comunidad sorda tiene bajo nivel de lectura bilingüe.
  - Hacer como hacen en "Simple english Wikipedia".
  - Algoritmos de compresión de frases.
  - Análisis de lecturabilidad de textos.
    * Frecuencias lexicales
    * Paráfrasis simple
    * Relaciones en *synsets* de WordNet.

# Simplificación lectora

- Se espera:
  - Un sistema resumidor de texto (abstractivo).
    * Simplificar estructura de frase.
    * Reducir variabilidad del vocabulario.
    * Conservar contenido semántico.
  - Demostración empírica.

# Contenidos

# Descripción textual parametros articulatorios

## Problema

Queremos hacer búsqueda paramétrica de señas en LSM desde el castellano.

# Descripción textual parametros articulatorios

Se requiere:

- Una representación "semasiológica" en español.
  - ¿Cómo representan las personas una seña en palabras?
  - Análisis cuantitativo de producciones escritas.
- ¿Cómo se indexa en un sistema de recuperación de la información?
  - Identificación de *parámetros* relevantes.

# Descripción textual parametros articulatorios

Se espera:

- Un análisis cuantitativo de producciones escritas describiendo señas.
- Un sistema de IR que reciba descripciones y recupere señas.
- Pruebas empíricas.

# Contenidos

# Interfaz en LSM para colecciones museisticas

## Problema

Diseño de una interfaz móvil para presentar colecciones del MUNAL en LSM.

# Interfaz en LSM para colecciones museisticas

Se requiere:

- Una interfaz basada en elementos visuales que sea fácilmente navegable por los Sordos.
  - ¿Cómo navegar los cuadros?
  - ¿Cómo dar información de contexto? (nivel educativo bajo)
  - ¿Cómo educar sobre el comportamiento en el museo? (experiencia nueva)

# Interfaz en LSM para colecciones museisticas

Se espera:

- Diseño e implementación de una aplicación de despliegue (en el teléfono).
  - Cuadros de José María Velasco (se tiene un rendering de alta resolución, datos de LSM y un demo salvaje).
- Diseño e implementación de una aplicación de edición.
  - ¿Cómo se introducen los materiales y la interpretación?
- Análisis de usabilidad.

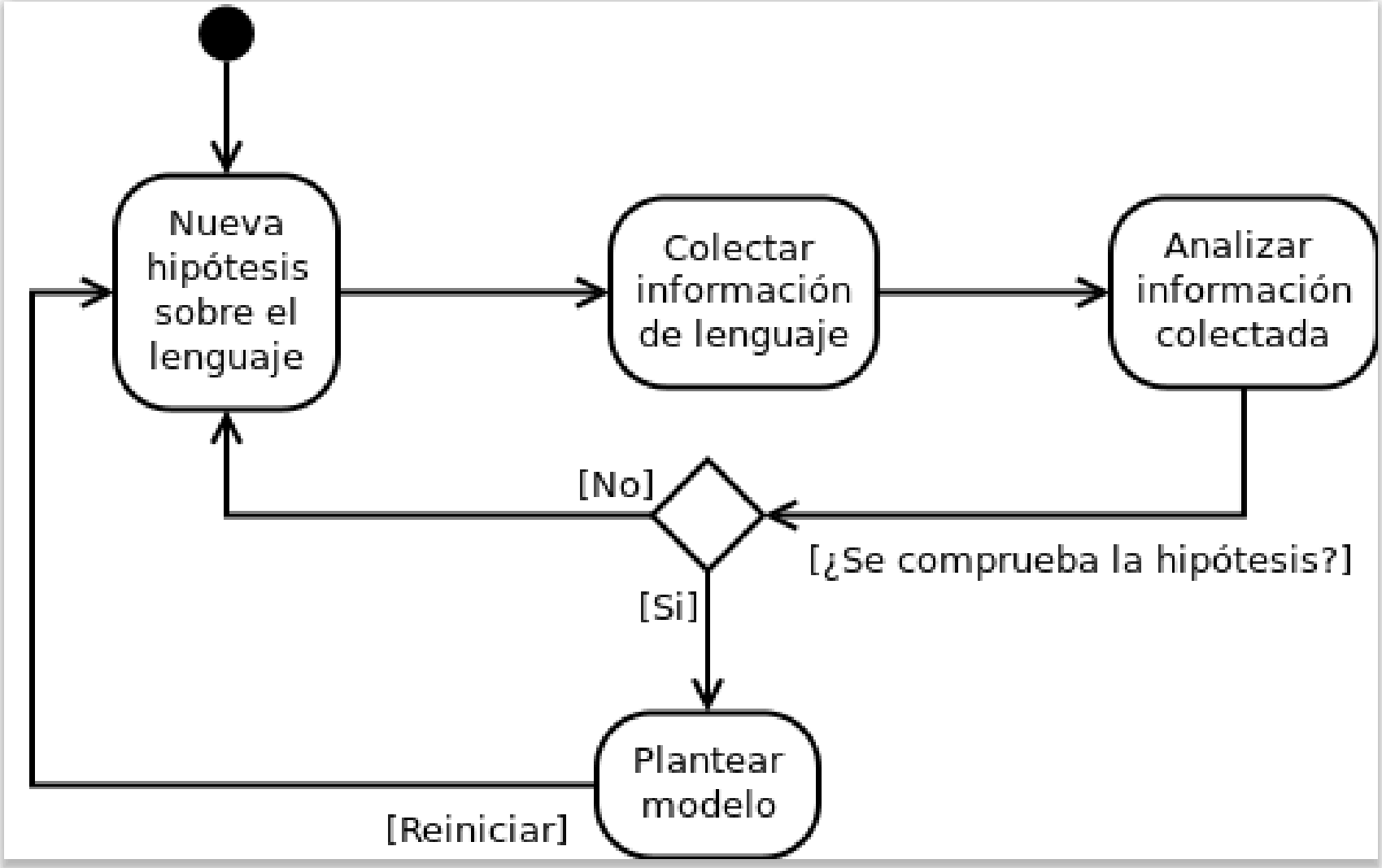
# Contenidos

# Módulos de anotación de video de LS

## Problema

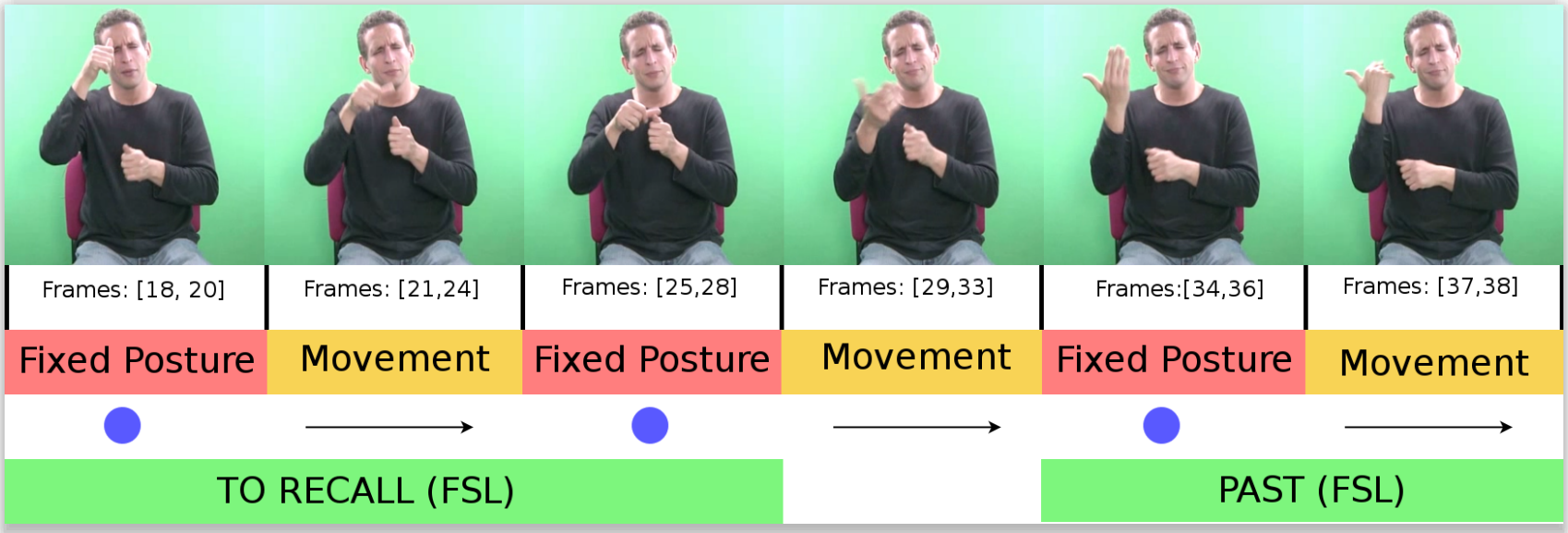
Crear tres módulos de anotación temporal de parámetros de LS.

# Módulos de anotación de video de LS



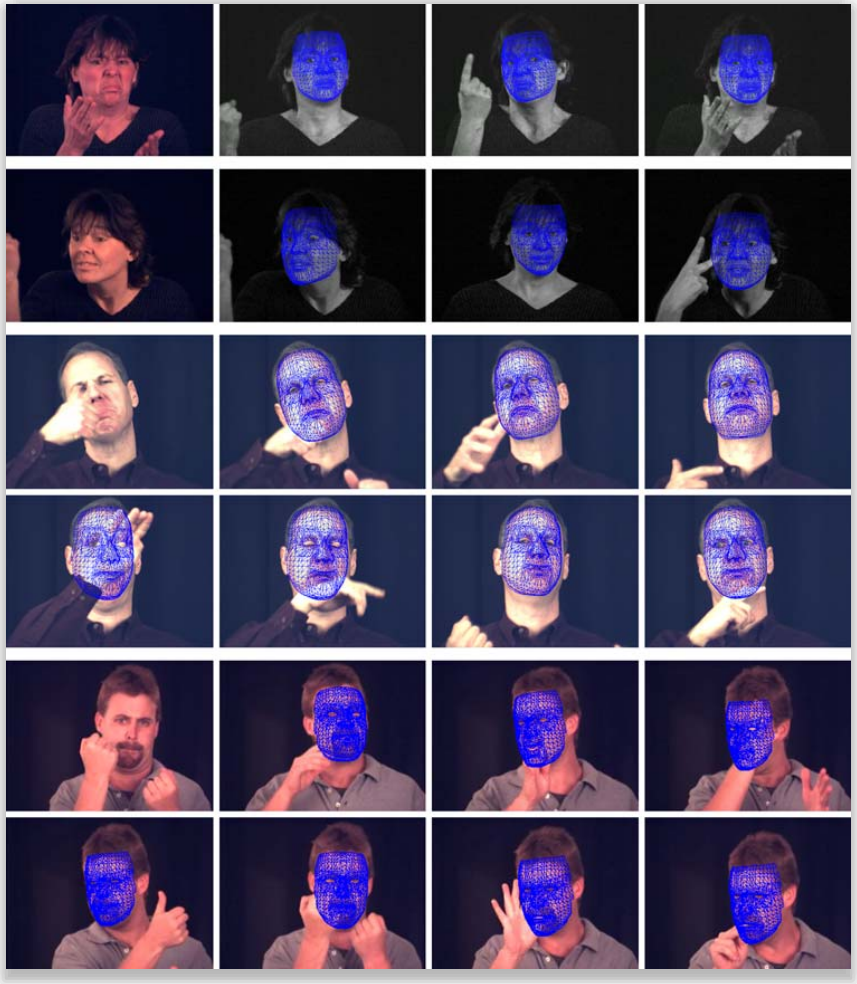

El proceso de investigación lingüística.

# Módulos de anotación de video de LS



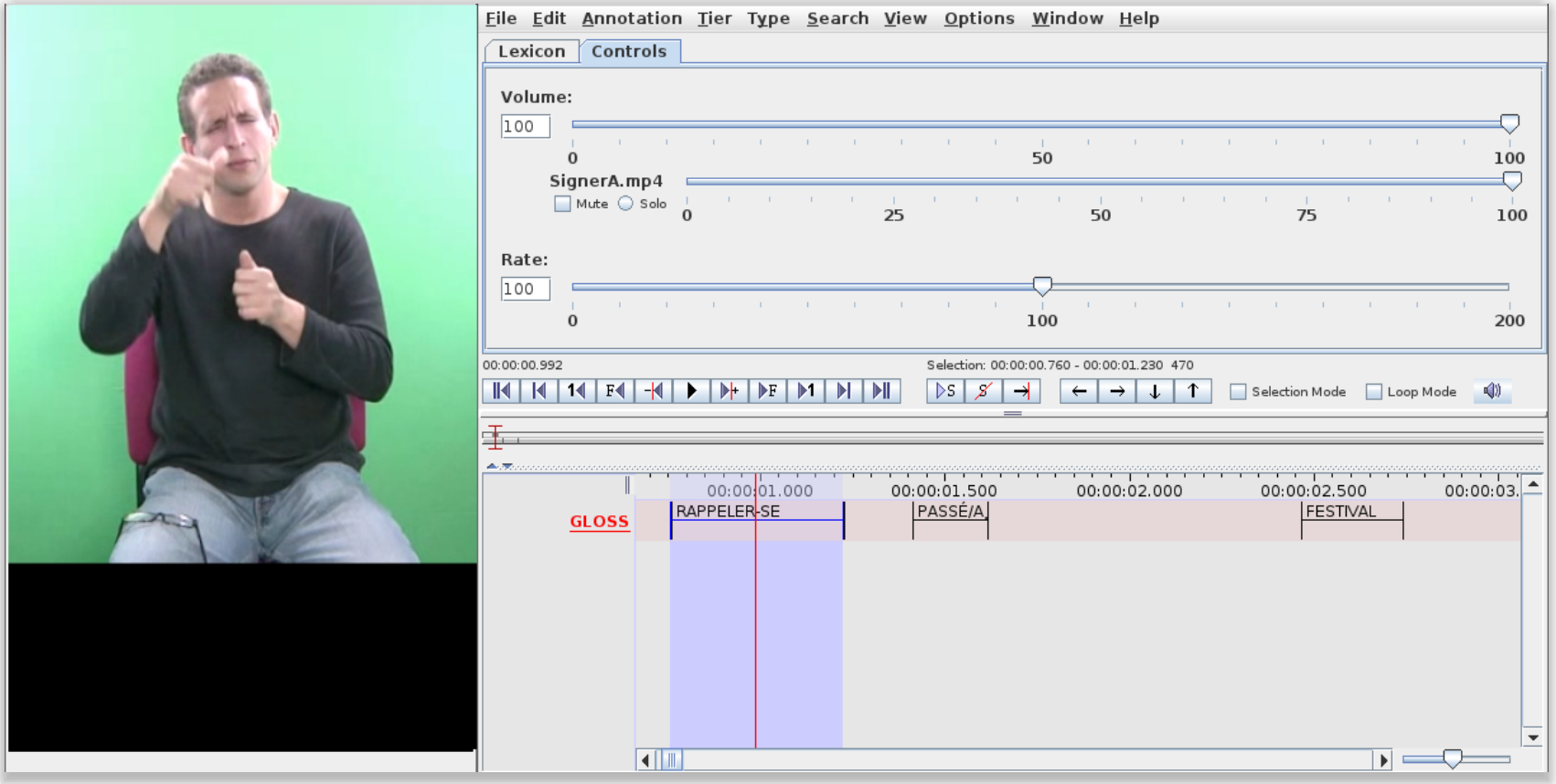

Anotación de video de LSF en dos niveles.

# Módulos de anotación de video de LS



Reconocimiento de expresión facial para ASL.

# Módulos de anotación de video de LS




Herramienta de anotación multimodal (ELAN)

# Módulos de anotación de video de LS

Se espera:

- Crear módulos de ELAN para el análisis de lenguas de Señas.
- Tres módulos específicos:
  1. Módulo de transcripción manual.
  2. Módulo de anotación de expresiones faciales.
  3. Módulo de anotación de movimientos con características fonéticas.
- Pruebas sobre corpus de 3 lenguas de señas (LSF, LSM y ASL).

# Contenidos

# Agente Conversacional Español-LSM (salud)

## Problema

Construir una interfaz de conversación Español simple-LSM que responda preguntas sobre salud.

# Agente Conversacional Español-LSM (salud)

- Cuando un sordo busca información sobre salud la mayor parte está en Español.
- No hay términos en LSM para buscar con señas:
  - Términos especializados en medicina no tienen traducción.
  - A veces ni al castellano, hacemos el prestamo lexical.
- Hay preguntas que se van a que hacer en lengua vocal.

# Agente Conversacional Español-LSM (salud)

Se espera:

- Crear un agente conversacional que responda:
  - Preguntas hechas en español simple con LSM.
  - Introduzca elementos visuales para pedir más información desde la interfaz (minimizar el uso de español).
- Obtener un corpus de video indexado con datos de salud.
  - Juntar el corpus.
  - Establecer los criterios para la indexación.

# Contenidos

# Morfología

## Definición

En lingüística la morfología se refiere al estudio de los *morfemas*, las unidades mínimas de significado.

# Segmentación morfológica automática

## Problema

Encontrar una representación óptima que permita obtener información sobre la formación de palabras en lenguaje humano.

# Morfema vs. Palabra

Morfema $\neq$ Palabra:

- Una palabra puede funcionar por si sola como unidad; un morfema no todos.
- Morfemas pueden ser libres o ligados:

  Ligados: ob-ten-er
  Libres: determinantes, preposiciones, conjunciones, pronombres, etc.

# Morfología

Se consideran dos tipos:

- Morfología concatenativa.
- Morfología no-concatenativa.

# Morfología concatenativa

Formación de lexemas a partir de "pegar" morfemas.

- Morfología Sintética (pocos morfemas libres)
  - Aglutinante (alto número de morfemas por palabra)
  - Fusionante (bajo número de morfemas por palabra)
- Morfología Analítica o Aislante (muchos morfemas libres)

Es un espectro

# Morfología concatenativa

## Morfología Sintética (Aglutinante)

- Los morfemas suelen pegarse sin (poca) modificación.
- Las reglas suelen ser muy regulares.
- El orden de las palabras tiende a tener menos importancia.
  - Las palabras contienen mucha información.
  - Los morfemas son distinguibles.

# Ejemplo: Lengua Nahuatl

## Nahuatl

Lengua indígena mesoamericana.

- Aglutinante.
- Sujeto Objeto Verbo.
  - No hay infinitivo, el agente y el paciente se aglutinan (prefijo) con la raíz del verbo en una sola palabra.
  - *agente* + *paciente* + *raíz* + *modificadores*
    - Modificadores: aspecto, tiempo, número, etc.

# Nahuatl

## Ejemplo (verbos)

Raíz: *xka* (asar/tostar con calor seco)

## Ejemplo (verbos)

Raíz: *xka* (asar/tostar con calor seco)

No puede aparecer sólo

| *nikixka* | | |
|---|---|---|
| *ni* | *ki* | *xka* |
| P. tónico | P. átono | raíz |
| "yo" | "lo" | "asar/tostar" |
| "Yo lo tuesto" | | |

## Ejemplo (verbos)

Raíz: *xka* (asar/tostar con calor seco)

No puede aparecer sólo

| *nimoxka* | | |
|---|---|---|
| *ni* | *mo* | *xka* |
| P. tónico | P. átono | raíz |
| "yo" | "me" | "asar/tostar" |
| "Yo me aso" | | |

# Nahuatl

## Ejemplo (verbos)

Raíz: *miki* (morir)

No puede aparecer sólo

| *nimomiki* | | |
|---|---|---|
| *ni* | *mo* | *miki* |
| P. tónico | P. átono | raíz |
| “yo” | “me” | “muerte” |
| “Me suicido” | | |

## Ejemplo (verbos)

Raíz: *xka* (asar/tostar con calor seco)

No puede aparecer sólo

| *nitlaxka* | | |
|---|---|---|
| *ni* | *tla* | *xka* |
| P. tónico | P. indef. | raíz |
| "yo" | "algo" (contextual) | "tostar" |
| "Aso lo que es obvio que es tostable" | | |

lo que es obvio que es tostable = tortilla

## Ejemplo (adjetivos)

Raíz: *xka* (asar/tostar con calor seco)

| *tlaxkal* | | |
|---|---|---|
| *tla* | *xka* | *l* |
| P. indef. | Raíz | modificador |
| "algo" (contextual) | "tostar" | ADJ |
| "Propiedad de ser como tortilla" | | |

## Ejemplo (sustantivos)

Raíz: *xka* (asar/tostar con calor seco)

| *tlaxkalli* | | | |
|---|---|---|---|
| *tla* | *xka* | *l* | *li* |
| P. indef. | Raíz | modificador | modificador |
| "algo" (contextual) | "tostar" | ADJ | SUST |
| "Tortilla" | | | |

## Ejemplo (modificaciones)

Raíz: *xka* (asar/tostar con calor seco)

Modificador: *tlan* (lugar)

| *tlaxkallan* | | | |
|---|---|---|---|
| *tla* | *xka* | *l* | *lan* |
| P. indef. | Raíz | modificador | modificador |
| "algo" (contextual) | "tostar" | ADJ/SUST | LUGAR |
| "Lugar de las tortillas" | | | |

tlaxkal~~li~~~~t~~lan

# Nahuatl

## Ejemplo (verbificar)

| *tlaxkalowa* | | | |
|---|---|---|---|
| *tla* | *xka* | *l* | *owa* |
| P. indef. | Raíz | modificador | modificador |
| “algo” (contextual) | “tostar” | ADJ/SUST | VERB |
| “Tortillear” “Dar palmadas” | | | |

*tlaxkal~~l~~iowa*

# Nahuatl

## Ejemplo (verbificar)

| *nimotlaxkalowa* | | |
|---|---|---|
| ni | mo | *tlaxkalowa* |
| P. tónico | P. reflexivo | Raíz |
| “yo” | “me” | “dar palmadas” |
| “Me doy palmadas” | | |

# Nahuatl

## Ejemplo (verbificar)

| *nimitzixkotlaxkalowa* | | | |
|---|---|---|---|
| *ni* | *mitz* | *ixko* | *tlaxkalowa* |
| P. tónico | P. posesivo | Sust. | Raíz |
| “yo” | “tuyo” | “rostro” | “dar palmadas” |
| “Te doy cachetadas” | | | |

# Nahuatl

## Ejemplo (verbificar)

| *nimitzixkotlaxkalowa* | | | |
|---|---|---|---|
| *ni* | *mitz* | *ixko* | *tlaxkalowa* |
| P. tónico | P. posesivo | Sust. | Raíz |
| "yo" | "tuyo" | "rostro" | "dar palmadas" |
| "Te doy cachetadas" | | | |

Palabras muy grandes, muy variables

# Morfología Concatenativa

## Morfología Sintética (Fusionante)

Lenguajes que usualmente presentan:

- Morfemas posicionales modificados.

  Sufijos, prefijos: *e.g.* ob-ten-er, ob-ten-ía, ob-tuv-e, ob-tuv-iste

- Morfemas irregulares.

Común en lenguas Romances.

# Español

## Lengua Española

En Español flexionamos:

- Sustantivos (género, número)
- Verbos (aspecto, tiempo, modo, persona, numero)
- Adjetivos (género, número)
- Artículos (género, número)
- Pronombres (género, número, función)

# Español

## Ejemplo

Concatenamos posiciones con respecto de la raíz.

| | | |
|---|---|---|
| | | er |
| ob | | ido |
| man | ten | me |
| con | | go |

# Español

## Ejemplo

Concatenamos posiciones con respecto de la raíz.

| | | |
|---|---|---|
| | | er |
| ob | | ido |
| man | ten | me |
| con | | go |

Puedo añadir más posiciones: pre-ob-ten-er

# Español

## Ejemplo

Concatenamos posiciones con respecto de la raíz.

| | | |
|---|---|---|
| | | er |
| ob | | ido |
| man | ten | me |
| con | | go |

Puedo añadir más posiciones con límite: $\emptyset$-pre-ob-ten-er

## Español vs. Rumano

Hay lenguajes más flexivos:

El reloj de Gabriel es mas viejo que el de María.
*Ceasul Gabrielei este mai vechi decât al Mariei.*
*Ceasul Gabriel lui este mai vechi decât al Mariei.*

La carta de Gabriel es mas vieja que la de María.
*Scrisoarea Gabrielei este mai veche decât a Mariei.*

La carta de alguna persona es mas vieja que la de María.
*Scrisoarea unei persoane este mai veche decât a Mariei.*

Los nombres de algunos niños son mas viejos que los de los hombres.
*Numele unor copii sunt mai vechi decât ai bărbaților.*

* un, ceas, scrisoare, veche, copil, bărbat, persoană, nume

## Morfología Analítica (Aislante)

- Poca flexión.
  - Más importancia al orden (sintaxis).
  - Más palabras gramaticales.
- Palabras más cortas.
  - Menos morfemas por palabra
  - *e.g. Lexicon del Vietnamita es 80 % bisilábico*.

La gata persigue a los perros
*The female cat chases the dogs*

# Chino

## Morfemas libres (Chino hablado)

| *wǒ* | *de* | *péngyou* | *men* | *dōu* | *yào* | *chī* | *dàn* |
|---|---|---|---|---|---|---|---|
| Yo | POSESIVO | amigo | PL | todo | querer | comer | huevo |
| “Todos mis amigos quiere comer huevo” | | | | | | | |

Orden sintáctico estricto

## Morfología no concatenativa

- La raíz se modifica sin "pegar" cosas.
- Orden interno de los morfemas estricto.
  - Lenguas semíticas (transfijación con raíz tri-consonante)

# Árabe

## Ejemplo k-t-b

Raíz triconsonante k-t-b (pertinente a la escritura/lectura):

kitāb (libro)

kutub (libros)

kātib (escritor)

kuttāb (escritores)

kataba (él escribió)

yaktubu (él escribe)

# Árabe

## Ejemplo k-t-b

Raíz triconsonante k-t-b (pertinente a la escritura/lectura):

kitāb (libro)

kutub (libros)

kātib (escritor)

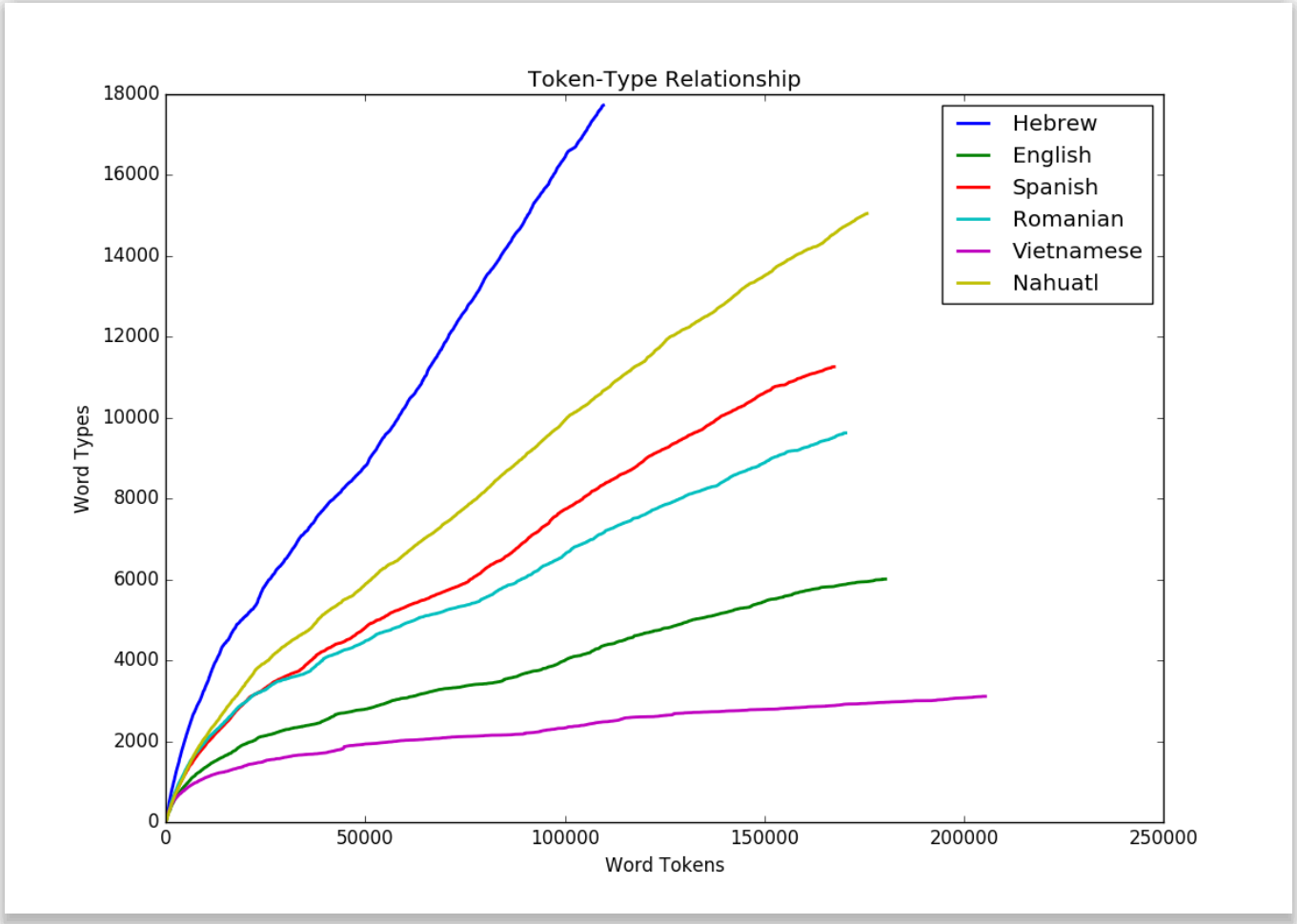
kuttāb (escritores)

kataba (él escribió)

yaktubu (él escribe)

Hay raices tetra-consonantes, penta-consonantes, hexa-consonantes ...

# Segmentación morfológica automática




Relación token-tipo en diferentes lenguas (complejidad morfológica)

# Segmentación morfológica automática

## Problema

Encontrar una representación óptima que permita obtener información sobre la formación de palabras en lenguaje humano.

# Segmentación morfológica automática

- Queremos encontrar una teoría del lenguaje bien fundada.
  - Basada en el estudio sistemático de fenómenos cuantificables.
- Mucho se ha hecho sobre el lexicon.
  - Estamos trabajando sobre elementos sub-lexicales de las palabras (en específico, morfemas).

# Segmentación morfológica automática

- Las palabras tienen una distribución Zipfiana, los morfemas no.
- Sin embargo, la relación token-tipo implica:
  - Existe una relación entre la distribución de las palabras y los morfemas.
  - Queremos inferir esa relación de un corpus multilingue.
- Exploramos técnicas de:
  - Topología algebraica y teoría de gráficas
  - Probabilidad
  - Simulación multi-agente (sistemas dinámicos)

# Segmentación morfológica automática

Espero obtener:

- Una representación de un corpus escrito en un espacio no métrico que me permita caracterizar su morfología.
  - Robusto a deformaciones (variación síncronica y diacrónica).
  - Induzca un álgebra que generalice la formación morfológica (*e.g.* un álgebra de Clifford).
- Un segmentador morfológico automático basado en la representación.
- Resultados empíricos sobre un corpus en Español y Nahuatl.
  - Fueron anotados morfológicamente.

# Otros temas

- Estoy abierto a otros proyectos a lo largo de estas lineas.
  - Contáctenme y lo planteamos.
- Si ya tienen datos lingüísticos disponibles y quieren trabajar con ellos, podemos hablarlo.

# Adopten uno

Los que no salen, los dormimos

# Adopten uno

Los que no salen, los dormimos