

Normalización de textos y análisis morfológico

Raúl E Gutierrez de Piñerez R. Ph.D

raul.gutierrez@correounivalle.edu.co

Escuela de Ingeniería de Sistemas y Computación
Universidad del Valle-Colombia

October 17, 2016



¿Qué es un corpus?

Una colección sistemática de textos ocurridos naturalmente (escritos o hablados).

Tipos de anotación

- Marcado de corpus
 - Información de procesamiento e información de formatos.
 - Clasificación de metadatos y textos
 - Representación estructural
- Tagging
 - Categorización de las palabras
- Parsing
 - Constituyentes y análisis de alto nivel.
 - Chunking/shallow vs análisis sintáctico completo
 - Análisis sintáctico superficial.
- Anotadores de lenguajes de marcado
 - Anotación en XML, LateXML, MathML, OMDoc.



Es un conjunto de corpus o corpora lingüísticas que puede ser procesados desde NLTK-PYTHON, (alrededor de 100)

- Descargar con **nltk.download** después de entrar a python
- Iniciar con **import nltk**
- Para llamar un corpus **from nltk.corpus import [id del corpus]**
- id de los corpus:
 - AnCora; *cess_esp*
 - PTB; *treebank*
 - CoNLL 2007 Dependency Treebanks: *conll2007*
 - Conll2000; *conll2000* y *conll2002*
 - Twitter Streaming AP: *twitter_samples*
 - Sentence Polarity: *sentence_polarity*
 - Subjectivity dataset: *subjectivityg*
 - Sentiment Polarity Dataset Version 2.0: *movie_reviews*
 - TIMIT Acoustic-Phonetic Continuous Speech Corpus; *timit*
 - Experimental Data for Question Classification: *qc*



- *Características principales*
 - Lema y categoría morfológica
 - Constituyentes y funciones sintácticas
 - Estructura argumental y papeles temáticos
 - Clase semántica verbal
 - Tipo denotativo de los nombres deverbales
 - Sentidos de WordNet nominales
 - Entidades nombradas
 - Relaciones de correferencia
- *Cada sentido verbal de los léxicos AnCora-Verb tiene asociada la siguiente información:*
 - La clase semántica, la estructura léxico-semántica, los esquemas sintácticos, los papeles temáticos y las restricciones selectivas de VerbNet. El esquema conceptual de [FrameNet](#).
 - Los sentidos verbales [PropBank](#), con sus argumentos y papeles temáticos correspondientes.
 - Los sentidos verbales de [WordNet 3.0](#).
 - Los sentidos verbales de la ontología [OntoNotes](#), donde la agrupación de sentidos es más fina que en WordNet.



Anotación morfológica en Ancora

| Word | lemma | PoS |
|---------|----------|---------|
| Si | si | CS |
| trabajo | trabajar | VMIP1SO |
| bajo | bajo | SPS00 |
| presión | presión | NCFS000 |
| bajo | bajar | VMIP1SO |
| el | el | DA0MS0 |
| interés | interés | NCMS000 |
| . | . | Fp |

(S
 (S.F.C.co-CD
 (S.F.C
 (sn-SUJ
 (espec.fp
 (da0fp0 Las el))
 (grup.nom.fp
 (ncfp000 reservas reserva)
 (sp
 (prep
 (sps00 de de))
 (sn.x
 (grup.nom.co
 (grup.nom.ms
 (ncms000 oro oro))
 (coord
 (cc y y))
 (grup.nom.fp
 (ncfp000 divisas divisa))))

lema

etiqueta morfoló.

AnCoraVerb_ES

abolir (verb, es) "abolir.lex.xml" in "AnCoraVerb_ES"

sense: 1 not lexicalized
verb.abolir.1.default Iss: A21.transitive-agentive-patient

AncoraNet

| PropBank | VerbNet | FrameNet | WordNet | Grouping |
|------------|-------------|----------|-------------|------------|
| abolish.01 | remove-10.1 | | abolish (1) | abolish.01 |

Arguments

| Function | Argument | Theme | Constituents | abolish.01(pb) | 10.1(vn) |
|----------|----------|-------|--------------|----------------|----------|
| suj | arg0 | agt | | arg0 | Agent |
| cd | arg1 | pat | | arg1 | Theme |

| AnCora-Verb | | |
|-------------|------------|---------|
| | Castellano | Catalán |
| Lemas | 2.820 | 2.248 |
| Sentidos | 3.938 | 3.118 |
| Frames | 5.117 | 4.642 |

| AnCora-Nom | | |
|------------|------------|---------|
| | Castellano | Catalán |
| Lemas | 1.658 | |
| Sentidos | 3.098 | |
| Frames | 3.208 | |



AnCora lexicón y AnCora corpus en español en xml

| | | | | | |
|------------------------|------------------------|------------------------|------------------------|----------------------|----------------------|
| abandono.lex.xml | abaratamiento.lex.xml | abastecimiento.lex.xml | abolicion.lex.xml | abonado.lex.xml | aborto.lex.xml |
| absorcion.lex.xml | abstencion.lex.xml | abstinencia.lex.xml | abstraccion.lex.xml | abundancia.lex.xml | aburrimiento.lex.xml |
| acabado.lex.xml | acatamiento.lex.xml | acceso.lex.xml | accidente.lex.xml | accion.lex.xml | aceleracion.lex.xml |
| acentuacion.lex.xml | aceptacion.lex.xml | acercamiento.lex.xml | achuchon.lex.xml | acierto.lex.xml | acogida.lex.xml |
| acompanamiento.lex.xml | acontecimiento.lex.xml | acopio.lex.xml | acorralamiento.lex.xml | acortamiento.lex.xml | acoso.lex.xml |

Anotación morfológica en xml

```
<!/lem="de" pos="sp00" postype="preposition" wd="de">
</prep>
<sn complex="yes">
  <grup.nom coord="yes">
    <grup.nom gen="m" num="s">
      <n gen="m" lem="oro" num="s" pos="ncms000" postype="common" sense="16:10487505" wd="oro">
```

<http://clic.ub.edu/corpus/es/ancora-descarregues>



| ETIQUETAS | | | |
|-----------|-----------|-----------|-----------|
| Posición | Atributo | Valor | Código |
| Columna 1 | Columna 2 | Columna 3 | Columna 4 |

| NOMBRES | | | |
|---------|-------------------------|--------------|--------|
| Pos. | Atributo | Valor | Código |
| 1 | Categoría | Nombre | N |
| 2 | Tipo | Común | C |
| | | Propio | P |
| 3 | Género | Masculino | M |
| | | Femenino | F |
| | | Común | C |
| 4 | Número | Singular | S |
| | | Plural | P |
| | | Invariable | N |
| 5-6 | Clasificación semántica | Persona | SP |
| | | Lugar | G0 |
| | | Organización | O0 |
| | | Otros | V0 |
| 7 | Grado | Aumentativo | A |
| | | Diminutivo | D |

tesis tesis NCFN000

Barcelona barcelona NP000G0

COI coi NP000O0

Pedro pedro NP000P0

- **Wikicorpus** es un corpus de tres lenguas (Catalan, Español, Inglés) que contiene grandes porciones de Wikipedia.
- El corpus ha sido anotado con el lema y POS usando **Freeling**
- También ha sido anotado el sentido usando UKB (Algoritmo de desambiguación de palabras). UKB asigna los sentidos de Wordnet.

```
<doc id="20540" title="658" nonfiltered="1"
      processed="1" dbindex="10000">
Acontecimientos acontecimientos NP00000 0
. . Fp 0
Nacimientos nacimientos NP00000 0
. . Fp 0
Fallecimientos fallecimientos NP00000 0
. . Fp 0
</doc>
```



Corpus de entrenamiento para POS tagging (Inglés) sobre PTB

```
( (S
  (NP-SBJ -1
    (NP (NNP Rudolph) (NNP Agnew))
    (, ,)
    (UCP
      (ADJP
        (NP (CD 55) (NNS years) )
        (JJ old) )
      (CC and)
      (NP
        (NP (JJ former) (NN chairman) )
        (PP (IN of)
          (NP (NNP Consolidated) (NNP Gold) (NNP Fields) (NNP PLC) )))
      (, ,))
    (VP (VBD was)
    (VP (VBN named)
    (S
      (NP-SBJ (-NONE- *-1) )
      (NP-PRD
        (NP (DT a) (JJ nonexecutive) (NN director) )
        (PP (IN of)
          (NP (DT this) (JJ British) (JJ industrial) (NN conglomerate) )))))
    (. .) ))
```

- El tagset de PTB tiene 45 tags <http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html>
- En `nltk:print(treebank.tagged_words('wsj_0001.mrg'))`



Anotación en Stanford Log-linear Part-Of-Speech Tagger

Corpus_NNP annotation_NN is_VBZ the_DT practice_NN of_IN
adding_VBG interpretative_JJ linguistic_JJ information_NN to_TO a_DT
corpus_NN

NN singular noun

VBZ form of the verb “BE”

VBG -ing form of lexical verb

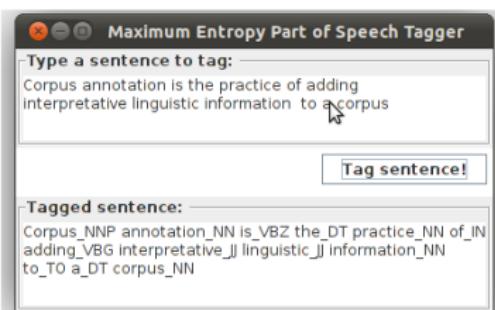
JJ adjective

IN the proposition

DT Determiner

TO the proposition of to

<http://nlp.stanford.edu/software/tagger.shtml>



Anotación sintáctica de AnCora y PTB

| PTB | ANCORA |
|--|--|
| ((S (NP-SBJ (DT That) (JJ cold) (, ,) (JJ empty) (NN sky)) (VP (VBD was) (ADJP-PRD (JJ full) (PP (IN of) (NP (NN fire) (CC and) (NN light))))) (. .))) | (S (sn-SUJ (espec.fs (da0fs0 La el)) (grup.nom.fs (ncfs000 declaración declaración))) (grup.verb (vmis3s0 propugnó propugnar)) (S.NF.C.co-CD (S.NF.C (infinitiu (vmn000 trabajar trabajar)) (sp-CC (prep (sps00 por por)) (sn (espec.fs (da0fs0 a el)) (grup.nom.fs (ncfs000 igualdad igualdad) (s.a.fs (grup.a.fs (aq0cs0 social social))))))) (Fp . .)) |

- En nltk: print (treebank.parsed_sents('wsj_0001.mrg')[0])
- <http://nlp.stanford.edu/software/tregex.shtml>



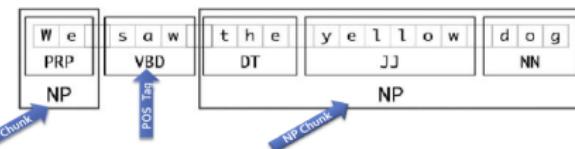
Ejemplo de anotación sintáctica PARSING ANNOTATION (M-TreeBank)

((S
 (VP (VBP-PS Suponga)
 (NP
 (DeD (NPU (DA-FS la) (NN-FS unión))
 (PP (IN de)
 (COORD
 (DeD (NPC (DA-MS el) (NN-MS conjunto)
 (NN-MS complemento))
 (PP (IN de) (VAR A)))
 (CC con)
 (DeD (DA-MS el) (NN-MS conjunto) (VAR A))))
 (SBAR (PRR que)
 (S
 (VP (VBZ-SI contiene)
 (PP (IN a)
 (DeD (DA-MS el) (NN-MS conjunto)
 (JJ-CS universal) (VAR U))))
 (. .)))

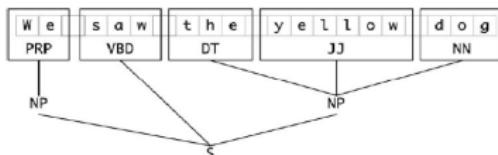


{NP,VP,PP, ADJP, ADVP} chunking IOB

Chunking consiste en dividir un texto en partes sintácticamente correlacionadas



Trees



Tags

- IOB tags (I-inside, O-outside, B-begin)
- label O for tokens outside a chunk

| | | | | |
|------|-----|------|--------|------|
| We | saw | the | yellow | dog |
| PRP | VBD | DT | JJ | NN |
| B-NP | O | B-NP | I-NP | I-NP |
| | | | | |
| | | | | |

| | | |
|-----------|-----|------|
| He | PRP | B-NP |
| reckons | VBD | B-VP |
| the | DT | B-NP |
| current | JJ | I-NP |
| account | NN | I-NP |
| deficit | NN | I-NP |
| will | MD | B-VP |
| narrow | VB | I-VP |
| to | TO | B-PP |
| only | RB | B-NP |
| # | # | I-NP |
| 1.8 | CD | I-NP |
| billion | CD | I-NP |
| in | IN | B-PP |
| September | NNP | B-NP |
| . | O | |

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

<http://text-processing.com/demo/tag/>

Ejemplo de anotación de Parsing: Chunking

- I-NP o I-VP es para cada palabra en el chunk
- B-NP o BPP es la primera palabra del chunk
- El tag O es para los tokens que no hacen parte del chunk
- Estructura de los datos de testeо y entrenamiento de CoNLL-2000

| | | |
|---------|-----|------|
| He | PRP | B-NP |
| reckons | VBZ | B-VP |
| the | DT | B-NP |
| current | JJ | I-NP |
| account | NN | I-NP |
| deficit | NN | I-NP |
| will | MD | B-VP |
| narrow | VB | I-VP |
| to | TO | B-PP |
| only | RB | B-NP |
| \# | \# | I-NP |
| 1.8 | CD | I-NP |
| billion | CD | I-NP |
| . | . | O |



A very important sub-task: **find** and **classify** names in text, for example:

- The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie, Rob Oakeshott, Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.
- The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie, Rob Oakeshott, Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organization

NER

El reconocimiento de entidades etiqueta secuencias de palabras en un texto las cuales son nombres de cosas, tales como personas y nombres de compañías, o nombres de genes y proteínas. Una buena tarea se define como un NER que reconoce las siguientes tres clases (PERSONAS, ORGANIZACIONES y LUGARES)

- **Resolución de anáforas:** *United Airline* y *united* usos de los pronombre personales sobre entidades
- **Detección de relaciones y clasificación:** Relaciones entre entidades.
- **Deteccción de eventos y clasificación:** Eventos en los cuales participan las entidades.
- **Reconocimiento de expresiones temporales:** En 1980 *united airline*



- **CoNLL-2002 (Conference on Computational Natural Language Learning)** y **CoNLL-2003** (British newswire) para múltiples lenguajes: Español, alemán e inglés: se anotan 4 entidades: Person, Location, Organization, miscellaneous
- **MUC-6 (Message Understanding Conferences)** y **MUC-7** (American newswire) se anotan 7 entidades: Person, Location, Organization, Time, Date, Percent, Money
- **ACE (Automatic Content Extraction)**, se anotan 5 entidades: Location, Organization, Person, GPE (Geo-Political Entity), FAC (Facility entities are limited to buildings)
- **BBN Pronoun Coreference and Entity Type Corpus** (Penn TreeBank), se anotan 22 entidades: Animal, Cardinal, Date, Disease, ...



Anotación del corpus ACE para entidades y relaciones

```
<ne id='e60' gid='E18' fr='w292' to='w295' hfr='w294' hto='w295' t='PER'>
<textspan type='extent'>American saxophonist David Murray</textspan>
<textspan type='head'>David Murray</textspan>
<exattr n='CLASS' v='SPC' />
<exattr n='LDCTYPE' v='NAM' />
</ne>
```

e
60

```
<ne id='e4' gid='E38' fr='w297' to='w298' t='PER'>
<textspan type='extent'>Amidu Berry</textspan>
<textspan type='head'>Amidu Berry</textspan>
<exattr n='CLASS' v='SPC' />
<exattr n='LDCTYPE' v='NAM' />
</ne>
```

e
4

```
<ne id='e5' gid='E1' fr='w300' to='w301' hfr='w301' hto='w301' t='PER'>
<textspan type='extent'>DJ Awadi</textspan>
<textspan type='head'>Awadi</textspan>
<exattr n='CLASS' v='SPC' />
<exattr n='LDCTYPE' v='NAM' />
</ne>
```

e
5

```
<text>
<p>
<s id='s17'>
<w id='w292'>American</w>
<w id='w293'>saxophonist</w>
<w id='w294'>David</w>
<w id='w295'>Murray</w>
<w id='w297'>recruited</w>
<w id='w298'>Berry</w>
<w id='w299'>and</w>
<w id='w300'>DJ</w>
<w id='w301'>Awadi</w>
<w id='w302'>.</w>
</s>
</p>
</text>
```

```
<rels>
<rel id='11-1' gid='11' e1='e61' e2='e62' t='GPE-AFF' st='Citizen-or-Resident' />
<rel id='2-1' gid='2' e1='e60' e2='e4' t='PER-SOC' st='Business' />
<rel id='3-1' gid='3' e1='e60' e2='e5' t='PER-SOC' st='Business' />
</rels>
```

subtipo



Anotación del corpus ConNLL-2002 y CoNLL-2003

CoNLL-2002-español

El DA O
Enty1{Abogado NC B-PER}
Enty1{General AQ I-PER}
Enty1{del SP I-PER}
Enty1{Estado NC I-PER}
, Fc O
Enty2{Daryl VMI B-PER}
Enty2{Williams NC I-PER}
, Fc O

CoNLL-2003- inglés

| | | |
|-----|-------|--------|
| NNP | I-NP | I-ORG |
| VBZ | I-VOP | O |
| JJ | I-NP | I-MISC |
| NN | I-NP | O |
| TO | I-VP | O |
| VB | I-VP | O |
| JJ | I-NP | I-MISC |
| . | O | O |

| Relation Type | Size | Example |
|---------------------|---------|---------------------------------------|
| per:birthPlace | 67 770 | Arnold Schwarzenegger, Thal |
| per:birthDate | 64 237 | Nelson Mandela, 1918-07-18 |
| per:spouse | 7 987 | Bill Clinton, Hillary Rodham Clinton |
| per:residence | 3 576 | François Hollande, Palais de l'Élysée |
| location:country | 72 468 | Wittislingen, Germany |
| location:mayor | 2 237 | Chicago, Rahm Emanuel |
| location:region | 51 484 | Paris, Île-de-France |
| org:adminCenter | 3 029 | UN, New York |
| org:leaderName | 4 080 | Thomson Reuters, David Thomson |
| org:foundedBy | 4 016 | IBM, Thomas J. Watson |
| org:foundingYear | 4 007 | IBM, 1911 |
| org:foundationPlace | 3 085 | Yahoo, Santa Clara |
| Total | 287 976 | |

Anotación de relaciones en dbpedia



Anotación de sentidos y discursiva en PDTB

Explicit

637..644

4,1,1,1,1,1,2,0

Text

because

#####

Features

Arb, PAtt, Null, Null

570..599

4,0;4,1,0;4,1,1,0

Text

Longer maturities are thought

#####

because, Contingency.Cause.Reason

Arg1

570..636

4,0;4,1,0;4,1,1,0;4,1,1,1,0;4,1,1,1,1,0;4,1,1,1,1,1,0;4,1,1,1,1,1,1;4,2

Text

Longer maturities are thought to indicate declining interest rates

#####

Features

Inh, Null, Null, Null

Arg2

645..729

4,1,1,1,1,2,1

Text

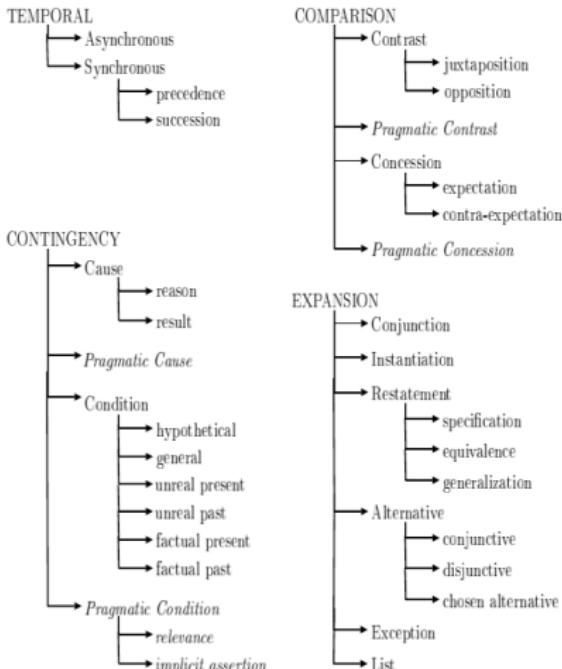
they permit portfolio managers to retain relatively higher rates for a longer period

#####

Features

Inh, Null, Null, Null

Longer maturities are thought to indicate declining interest rates because they permit portfolio managers to retain relatively higher rates for a longer period.



- (2) *Third-quarter sales in Europe were exceptionally strong, boosted by promotional programs and new products – although weaker foreign currencies reduced the company's earnings.*
- (3) Michelle lives in a hotel room, and although she drives a canary-colored Porsche, she hasn't time to clean or repair it.
- (4) *Most oil companies, when they set exploration and production budgets for this year, forecast revenue of \$15 for each barrel of crude produced.*
- (6) But a few funds have taken other defensive steps. *Some have raised their cash positions to record levels.* Implicit = BECAUSE **High cash positions help buffer a fund when the market falls.**
- (9) *Jacobs is an international engineering and construction concern. NoRel Total capital investment at the site could be as much as \$400 million, according to Intel.*

· Arg1 although- Arg2

although-Arg2-Arg1

Arg2 embebido en Arg1

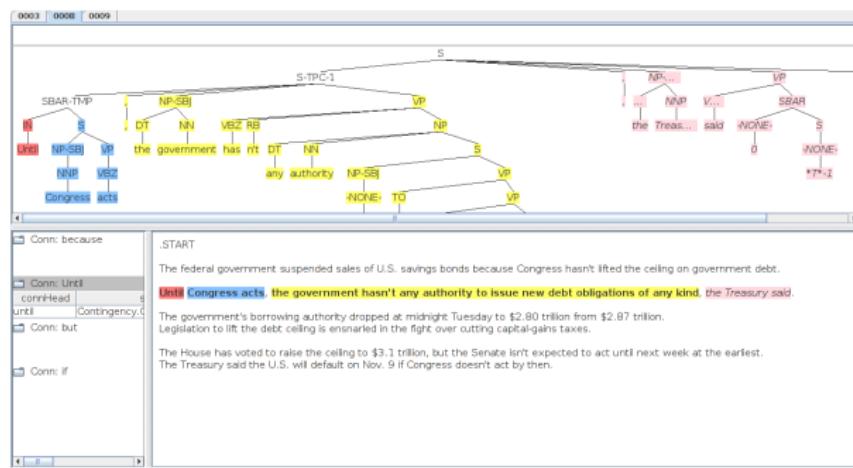
Los Argumentos son objetos abstractos

Dirección inferida: causal
Arg1 - implícito=because- Arg2

No hay relación entre argumentos



- Anotación de las **relaciones del discurso**; relaciones implícitas y explícitas.
- Los **argumentos** de las relaciones también son anotados (Arg1 y Arg2)
- Los **sentidos (semánticos)** de las relaciones, como las características.
- Las **atribuciones** de las relaciones y sus argumentos son anotadas semánticamente.



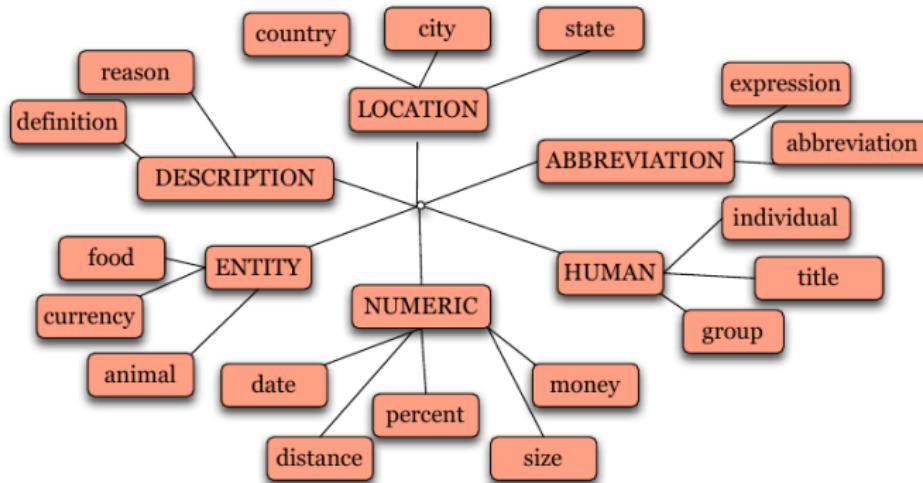
Clasificación de la pregunta o reconocimiento del tipo de respuesta

Se puede conocer el tipo de respuesta de la pregunta y así evitar mirar en toda la sentencia o el sintagma nominal y en lugar enfocarnos en las entidades o clases; personas, lugares, definiciones.

- ¿Qué profesores trabajan en el área de Restricciones? **tipo de respuesta:** (HUMANO-grupo)
- ¿Cuántos artículos sobre cálculo pi se han escrito en la EISC? **tipo de respuesta:** (NUMERICO-conteo).
- ¿Qué porcentaje de estudiantes hacen parte de los grupos de investigación de la EISC? **tipo de respuesta:** (NUMERICO-porcentaje)
- ¿Cuáles son los grupos de investigación de la EISC que están clasificados en COLCIENCIAS como tipo A? **tipo de respuesta:** (ENTIDAD-grupo-investigación)



Xin Li, Dan Roth. 2002. Learning Question Clasifiers. COLLING'02



- 6 Clases generales: ABBREVIATION, DESCRIPTION, ENTITY, HUMAN, LOCATION, NUMERIC
- 50 clases más detalladas:
 - LOCATION: city, country, mountain ...
 - HUMAN: group, individual, title, description ...

Tipos de respuesta, Xin Li, Dan Roth. 2002 (QA Systems)

| Tag | Example |
|------------------|--|
| ABBREVIATION | |
| abb | What's the abbreviation for limited partnership? |
| exp | What does the "c" stand for in the equation E=mc2? |
| DESCRIPTION | |
| definition | What are tannins ? |
| description | What are the words to the Canadian National anthem? |
| manner | How can you get rust stains out of clothing? |
| reason | What caused the Titanic to sink ? |
| ENTITY | |
| animal | What are the names of Odin's ravens? |
| body | What part of your body contains the corpus callosum ? |
| color | What colors make up a rainbow ? |
| creative | In what book can I find the story of Aladdin? |
| currency | What currency is used in China? |
| disease/medicine | What does Salk vaccine prevent ? |
| event | What war involved the battle of Chapultepec? |
| food | What kind of nuts are used in marzipan? |
| instrument | What instrument does Max Roach play? |
| lang | What's the official language of Algeria? |
| letter | What letter appears on the cold-water tap in Spain? |
| other | What is the name of King Arthur's sword? |
| plant | What are some fragrant white climbing roses? |
| product | What is the fastest computer ? |
| religion | What religion has the most members ? |
| sport | What was the name of the ball game played by the Mayans? |
| substance | What fuel do airplanes use? |
| symbol | What is the chemical symbol for nitrogen ? |
| technique | What is the best way to remove wallpaper? |
| term | How do you say " Grandma " in Irish ? |
| vehicle | What was the name of Captain Bligh's ship ? |
| word | What's the singular of dice? |

Ejemplo de una taxonomía anotada manualmente; el tagset de la ontología jerárquica de Li & Roth (2005)

<http://cogcomp.cs.illinois.edu/Data/QA/QC/>



Corpus TASS para análisis de sentimientos

```
<?xml version="1.0" encoding="UTF-8"?>
<tweets>
  <tweet>
    <tweetid>142378325086715906</tweetid>
    <user>jesusmarana</user>
    <content><![CDATA[Portada 'Público', viernes. Fabra al banquillo por 'orden' del Supremo; Wikile</content>
    <date>2011-12-02T00:03:32</date>
    <lang>es</lang>
    <sentiments>
      <polarity><value>N</value></polarity>
    </sentiments>
    <topics>
      <topic>politica</topic>
    </topics>
  </tweet>
</tweets>
```

```
<sentiments>
  <polarity><value>N+</value><type>AGREEMENT</type></polarity>
  <polarity><entity>Sinde</entity><value>N</value><type>AGREEMENT</type></polarity>
  <polarity><entity>SGAE</entity><value>N+</value><type>AGREEMENT</type></polarity>
</sentiments>
```

- El corpus TASS tiene 77.550 tweets en español algunos anotados con polaridad
- Hay múltiples niveles de polaridad con los siguientes valores: **N+** (muy negativo), **N** (negativo), **NEU** (Neutral), **P** (Positivo), **P+** (muy positivo).



Corpus para minería de opinión

```
>>> from nltk.corpus import subjectivity
>>> subjectivity.categories()
['obj', 'subj']
>>> subjectivity.sents()[23]
['television', 'made', 'him', 'famous', ',', 'but', 'his', 'biggest', 'hits',
'happened', 'off', 'screen', '.']
>>> subjectivity.words(categories='subj')
['smart', 'and', 'alert', ',', 'thirteen', ...]
```

```
>>> from nltk.corpus import sentence_polarity
>>> sentence_polarity.sents()
[['simplicistic', ',', 'silly', 'and', 'tedious', '.'], ["it's", 'so', 'laddish',
'and', 'juvenile', ',', 'only', 'teenage', 'boys', 'could', 'possibly', 'find',
'it', 'funny', '.'], ...]
>>> sentence_polarity.categories()
['neg', 'pos']
>>> sentence_polarity.sents()[1]
["it's", 'so', 'laddish', 'and', 'juvenile', ',', 'only', 'teenage', 'boys',
'could', 'possibly', 'find', 'it', 'funny', '.']
```

- The Sentence Polarity dataset contiene 5331 sentencias positivas and 5331 sentencias negativas.
- The Subjectivity Dataset contiene 5000 sentencias subjetivas y 5000 sentencias objetivas.



Herramienta de procesamiento de texto en Español

Esta herramienta permite el procesamiento de texto a nivel morfológico, léxico, sintáctico, semántico , desambiguación, NER, ect.

- Funciona como librería para lenguajes de programación, stand alone o como cliente servidor
- **analyze -f es.cfg –outf morfo < entrada.txt > salida.txt**
- **analyze -f es.cfg –inpf morfo –outf tagged < salida.txt**



Toda tarea de PLN necesita la normalización del texto:

- ① Segmentación de sentencias en el texto en marcha
- ② Tokenización de palabras sobre el texto segmentado
- ③ Normalización de las palabras y análisis morfológico

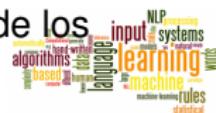


Este es el primer paso en el procesamiento de texto.

- La segmentación de un texto en sentencias es generalmente basada en puntuación.
- Puntos (Periods), signos de interrogación (question marks), signos de exclamación (exclamation points).
- Menos ambiguos: signos de interrogación y exclamación. Más ambiguos el punto (period).
- A veces la ambigüedad del punto hace que la tokenización de sentencias se haga junto con la tokenización de palabras.
- En general, los métodos de tokenización de sentencias trabajan como clasificadores lineales (basados en reglas de secuencias o aprendizaje automático) los cuales deciden si un punto es parte de la palabra o la sentencia (sentence boundary marker).

Ejemplo: *Mr.* o *Inc.*

- Un primer paso para la construcción de segmentadores de sentencias son el uso de expresiones regulares a través de los FSTs (Karttunen-1996 y Beesly-Karttunen-2003)



Problema duro!!!!!!

Text Analysis Result -- NLTK Sentence Segmentation

Original Text

Recognition of Named Entities in Spanish Text. Sofía N. Galicia-Harol, Alexander Gelbunk2,3, and Igor A. Bolshakov2. 1 Faculty of Sciences. UNAM Ciudad Universitaria México City, México. sngb@ciencias.unam.mx. 2 Center for Computing Research National Polytechnic Institute, México City, México {gelbunk,igor}@cic.ipn.mx, www.Gelbunkh.com. 3 Departament of computer Science and Engineering, Chung-Ang University.

#ColombiaNoSeEntrega #ColombiaDiceNo
#unidosPorElNo

HOY mas que NUNCA: #ColombiaNoSeEntrega por
el bien de nuestro país! Digamos NO este 2
de octubre y hoy a esta farsa en Cartagena. ☺☺

Analysis Result

Recognition of Named Entities in Spanish Text.

Sofía N. Galicia-Harol, Alexander Gelbunk2,3, and Igor A. Bolshakov2.

1 Faculty of Sciences.

UNAM Ciudad Universitaria México City, México.

sngb@ciencias.unam.mx.

2 Center for Computing Research National Polytechnic Institute, México City, México {gelbunk,igor}@cic.ipn.mx, www.Gelbunkh.com.

3 Departament of computer Science and Engineering, Chung-Ang University.

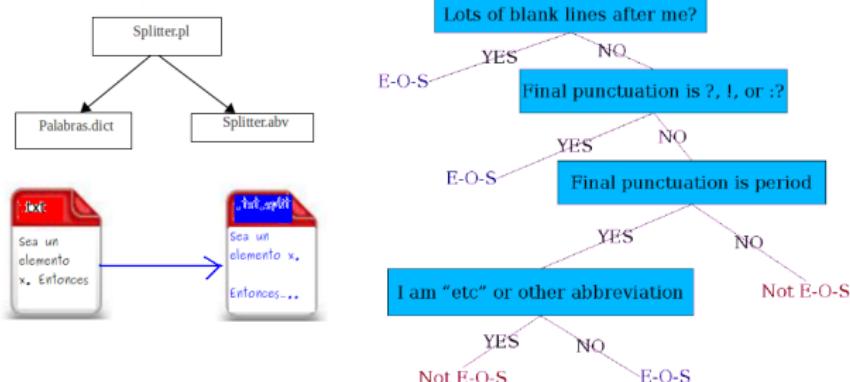
#ElMundoDiceSí, un plan perfecto
para pasar disfrutar la vida con tus
amigos.

Segmentador de nltk:

<http://textanalysisonline.com/nltk-sentence-segmentation>

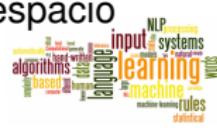


Sentence Splitter [?]



- **Input:** An ASCII text file e.g. `blob.txt`
- **Output:** El archivo de entrada con cada sentencia ya segmentada por línea y todo [.!?] anotado con la categoría [sentence_break].
- **Una simple regla:** Todos los [.!?] son considerados como sentencias terminales si están seguidas de al menos un espacio en blanco y una letra mayúscula.

$$\text{sentence_break} = /[.!?][()"] + [A - Z]$$



Ejemplo de un segmentador de sentencias

Argument

Sea un elemento x que pertenece al conjunto universal U . Entonces el elemento x pertenece al conjunto complemento de A o el elemento x pertenece al conjunto A , porque la unión del conjunto complemento de A con el conjunto A es igual al conjunto universal U . En particular, si el elemento x pertenece al conjunto A , entonces el elemento x pertenece al conjunto B , debido a que el conjunto B contiene al conjunto A . Por lo tanto, se deduce que el elemento x pertenece al conjunto complemento de A o el elemento x pertenece al conjunto B , es decir, el elemento x pertenece a la unión del conjunto complemento de A con el conjunto B .

Sentences splitter

Sea un elemento x que pertenece al conjunto universal U .

Entonces el elemento x pertenece al conjunto complemento de A o el elemento x pertenece al conjunto A , porque la unión del conjunto complemento de A con el conjunto A es igual al conjunto universal U .

En particular, si el elemento x pertenece al conjunto A , entonces el elemento x pertenece al conjunto B , debido a que el conjunto B contiene al conjunto A .

Por lo tanto, se deduce que el elemento x pertenece al conjunto complemento de A o el elemento x pertenece al conjunto B , es decir, el elemento x pertenece a la unión del conjunto complemento de A con el conjunto B .

- Lee de un diccionario y encuentra las palabras comunes.
- Aquí las palabras que comienzan con mayúsculas son simples nombres y uno por línea.



Se quiere mantener lo que internamente ocurre en la palabra

- Ejemplos como: m.p.h., Ph.D.. AT& T, we're
- Caracteres especiales y números que necesitan mantener su precio: (\$45) y fechas (01/02/06)
- Lo mismo pasa URLs (<http://www.univalle.edu.co>), **inicialmente** hashtags de Twitter (#yosoyapoliticoyque), o emails algunacosa@univalle.edu
- Un tokenizador debe manejar contracciones clíticas como **we're** a **we are**, **pa'que** a **para que**
- Se puede pensar la tokenización como un problema de **NER**, para identificar, nombres, lugares, en hashtags...



Tokens

Es la tarea de separar las palabras o tokens de un texto dado.

- Separación de expresiones como *Ph.D.*, *AT&T*, *62.5* y *google.com*
- Con **tr** en unix se puede tokenizar y obtener las frecuencias de cada una de las palabras del texto

```
tr -sc 'A-Za-z' '\n'< sh.txt  
tr -sc 'A-Za-z' '\n'< shakes.t| sort | uniq -c
```

- Búsquedas especiales:
`tr -sc 'A-Za-z' '\n'< shakes.t| grep 'ing$' | sort | uniq -c | sort -nr`
- Morfología en un corpus plano

```
(*v*)ing → ø walking → walk  
                         sing      → sing
```

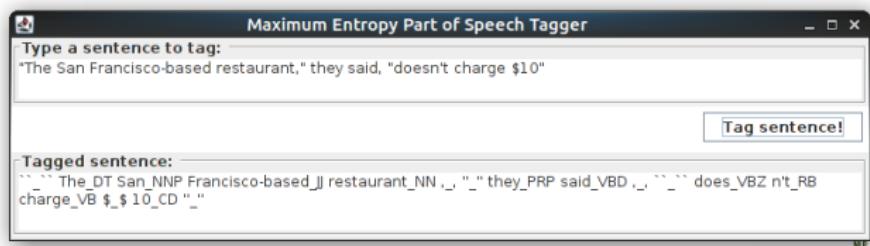


- Este estándar separa clíticos (doesn't por does + n't)
- También mantiene las palabras con guiones juntas y separa toda puntuación
- Los tokens también puede ser normalizados dependiendo del estándar

Input: "The San Francisco-based restaurant," they said, "doesn't charge \$10".

Output:

| | | | | | | | | | |
|------|-----|-----|-----------------|------------|--------|----|------|---|---|
| " | The | San | Francisco-based | restaurant | , | " | they | | |
| said | , | " | does | n't | charge | \$ | 10 | " | . |



TweetTokenizer

Esta es una herramienta de nltk muy poderosa que puede ayudar en el preprocesamiento de tweets

- Desde python, luego import nltk
- from nltk.tokenize import TweetTokenizer
- Uso de **reduce_len** reemplaza secuencias de 3 o más por secuencias de longitud 3

```
>>> from nltk.tokenize import TweetTokenizer
>>> tknzs = TweetTokenizer()
>>> s0 = "This is a coooool #dummymiley: ::) :-P <3 and some arrows < > -> <- -"
>>> tknzs.tokenize(s0)
['This', 'is', 'a', 'coooool', '#dummymiley', ':', '::)', '-P', '<3', 'and', 'some', 'arrows', '<', '>', '->', '<- -']
```

Examples using *strip_handles* and *reduce_len* parameters:

```
>>> tknzs = TweetTokenizer(strip_handles=True, reduce_len=True)
>>> s1 = '@remy: This is waaaayyy too much for you!!!!!!'
>>> tknzs.tokenize(s1)
[':', 'This', 'is', 'waaaayyy', 'too', 'much', 'for', 'you', '!', '!', '!']
```

Algoritmo MaxMatch

Inicia desde el primer carácter y encuentra la palabra más grande en un diccionario o lexicón que coincida con la entrada en la posición actual. El puntero avanza hacia el final de la sentencia. Si ninguna palabra coincide, el puntero avanza carácter por carácter. El algoritmo se aplica iterativamente a partir de la nueva posición del puntero.

```
#bigbangtheory -> big bang theory  
#chickensoup -> chicken soup  
#running -> running  
#30times -> 30 times  
#neverstop -> never stop
```

```
function MAXMATCH(sentence, dictionary D) returns word sequence W  
    if sentence is empty  
        return empty list  
    for i ← length(sentence) downto 1  
        firstword = first i chars of sentence  
        remainder = rest of sentence  
        if InDictionary(firstword, D)  
            return list(firstword, MaxMatch(remainder,dictionary) )  
  
    # no word was found, so make a one-character word  
    firstword = first char of sentence  
    remainder = rest of sentence  
    return list(firstword, MaxMatch(remainder,dictionary) )
```

- **Stemming** (information retrieval and web search (IR)), se mapea foxes desde fox sin necesidad de conocer que foxes is plural.
Despojarse de los finales de las palabras.
- **Lematización:** Es la tarea de determinar que dos palabras tienen la misma raíz, por ejemplo; sing es el lema de song y sang. También am, are e is comparten el lema be
- <http://nlp.lsi.upc.edu/freeling/demo/demo.php>

| El | hombre | bajo | toca | el | bajo | bajo | la | escalera |
|-------------------|-------------------------------|--------------------------------|--------------------------------|-------------------|--------------------------------|--------------------------------|------------------------------|--------------------------|
| el DA0MS0 1 | hombre NCMS000 0.961347 | bajo SPS00 0.879562 | toca NCF5000 0.764439 | el DA0MS0 1 | bajo SPS00 0.879562 | bajo SPS00 0.879562 | el DA0FS0 0.972269 | escalera NCFS000 1 |
| | hombre I 0.0386534 | bajo AQ0MS0 0.0766423 | tocar VMIP3S0 0.233251 | | bajo AQ0MS0 0.0766423 | bajo AQ0MS0 0.0766423 | lo PP3FSA00 0.0277039 | |
| | | bajo NCMS000 0.040146 | tocar VMM02S0 0.00231023 | | bajo NCMS000 0.040146 | bajo NCMS000 0.040146 | la NCMS000 2.74025e-05 | |
| | | bajar VMIP1S0 0.00364964 | | | bajar VMIP1S0 0.00364964 | bajar VMIP1S0 0.00364964 | | |

- Las palabras pueden clasificarse en **clases/categorías** (Part of Speech PoS) con base a su función dentro de la frase o sentencia.
- Clases de palabras:
 - **Clases cerradas:** Número fijo de palabras (no se pueden añadir más). Ej: proposiciones, pronombres, conjunciones....
 - **Clases abiertas:** Permiten añadir nuevas palabras (mediante flexión, derivación, composición..) Ej: nombres, adjetivos, verbos, adverbios.....
- Dentro de una clase pueden haber **subclases**: la clase nombre tiene dos subclases: los nombres propios y comunes. Las subclase comunes: contables e incontables.
- Hay pronombres; personales, posesivos, interrogativos y relativos.



Lexemas y morfemas (Tomado de EDU365.CAT)

http://www.edu365.cat/eso/faqs/castella/lex_morf.htm)

Lexemas

Unidad con significado léxico, es decir, aporta a la palabra una idea comprensible para los hablantes. Normalmente es la parte que se repite. Ej: panadero: **pan**, destornillador: **tornill**

Libr- Lexema, puede anexar morfemas como:

- o-s Morfemas dependientes flexivos de masculino y plural
- eta Morfema dependiente derivativo sufijo.
- ería Morfema dependiente derivativo sufijo.
- ito Morfema dependiente derivativo sufijo.
- eto Morfema dependiente derivativo sufijo.

El lexema es **hac**: deshacer, hacer, deshacía, haces.

Morfemas

Unidad con significado gramatical, es decir, complementa al lexema en género, número, aumentativo, diminutivo y otras terminaciones

Ej: casita, **destaar**, **inaguantable**.

- **Morfemas gramaticales:**

- **Morfema de género:** Para indicar si la palabra está en masculino o femenino, Ej: león, leon-**a**.
- **Morfema de número:** Indica si la palabra está en plural. Ej: león, leon-**es**
- **Desinencias:** Son morfemas que se añaden al lexema de los verbos para indicar la persona, el número, el tiempo y el modo. Ten-**emos**, compró, despertará. Son importantes porque indican los modos del verbo: *Indicativo, subjuntivo, imperativo*.

- **Morfemas derivativos:**

- **Prefijos:** Van antes del lexema. **Extra-muros**; **Pre-historia**
- **Sufijos:** Van después del lexema. Metró-**polis**; hidro-**terapia**



Definición lingüística

Es una unidad autónoma constituyente del léxico con unidad significativa que se considera una entrada de diccionario o enciclopedia

- Los lemas se calculan para verbos, sustantivos, adjetivos, etc.
- El lema sobre verbos deja la palabra en su parte infinitiva
- Grampal: Etiquetador morfológico y lematizador

[http://cartago.111f.uam.es/grampal/grampal.cgi?
m=analiza&e=naturales](http://cartago.111f.uam.es/grampal/grampal.cgi?m=analiza&e=naturales)

- Anotador de AnCora es más preciso
- los perros bandidos corren apresurados por los montes

| los | perros | bandidos | corren | apresurados | por | los | montes |
|--------------|------------------|--------------------|-------------------|----------------------|-----------|--------------|------------------|
| el DAOMPO | perro NCMP000 | bandido AQOMP00 | correr VMIP3PO | apresurar VMP00PM | por SP | el DAOMPO | monte NCMP000 |

Normalización de palabras

- Hay dos enfoques de normalización de palabras: **Stemming** y la **Lematización**
- La **lematización (cabeza en el diccionario)** reemplaza el sufijo de una palabra con una diferente o remueve completamente el sufijo de una palabra y genera **la palabra básica (lema) o canónica**.
- El **stemming** no necesariamente produce la forma básica de la palabra solamente una aproximación de ella llamada **stem** o **forma normalizada**. Su uso es propio de la **recuperación de información**.
- Ej: las palabras *calcular*, *calcula*, *calculado* o *calculando* serán lematizadas usando stemming como *calcul* pero su forma normal (lematización) es el infinitivo de la palabra *calcular*.

Try the Spanish stemming algorithm:

calculator the calculated calculating the calculation that she calculate

calcul lo calcul calcul el calcul que ella calcul

Ejemplo de lematización y stemming de la palabra calcula

El análisis morfológico, stems y lexemas

El análisis morfológico se basa en las características lingüísticas de composición de una palabra y usa los lexemas como stems o raíces léxicas (canónicas)

- Palabra a lematizar: Calcula

Lema: calcular

Categoría: verbo transitivo

Flexión: 2^a per. sing. imperat. y 3^a per. sing.
pres. ind.

Clasificación semántica: De significación
inmaterial,

Tecnicismos de ciencias y artes Matemáticas

- Stemming: calcul



- Hay algoritmos simples para el análisis morfológico sin FSTs que son empleados en IR (Information Retrieval)
- Estos algoritmos sin grandes lexicones sirven para tareas de búsquedas en la web, en la cual una pregunta (query) es una combinación Booleana de relevantes palabras claves (**keywords**) o frases.
- La información morfológica en IR es solamente usada para determinar que dos palabras tienen el mismo **stem**.
- Uno de los más ampliamante usados tales como los algoritmos **stemming** es el de Porter, simple y eficiente, el cual está basado en una serie de simples reglas de reescritura.
- El algoritmo de Porter Stemmer se encuentra en:
<http://tartarus.org/~martin/PorterStemmer/>
- Stemming and Lemmatization with Python NLTK
<http://text-processing.com/demo/stem/>



Porter Stemmer

Step 1a

| | | | |
|------|------|----------|----------|
| sses | → ss | caresses | → caress |
| ies | → i | ponies | → poni |
| ss | → ss | caress | → caress |
| s | → Ø | cats | → cat |

Step 1b

| | | | |
|----------|-----|-----------|-----------|
| (*v*)ing | → Ø | walking | → walk |
| | | sing | → sing |
| (*v*)ed | → Ø | plastered | → plaster |

Step 2 (for long stems)

| | | | |
|---------|-------|------------|------------|
| ational | → ate | relational | → relate |
| izer | → ize | digitizer | → digitize |
| ator | → ate | operator | → operate |
| ... | | | |

Step 3 (for longer stems)

| | | | |
|------|-----|------------|----------|
| al | → Ø | revival | → reviv |
| able | → Ø | adjustable | → adjust |
| ate | → Ø | activate | → activ |

Stem Text

Choose stemmer

Porter

Enter text

walking and ponies and revival
adjustable and caresses, activate
digitizer

Stemmed Text

walk and poni and reviv adjust and caress , activ digit

Snowball stemmer

Pasos1. Eliminación del sufijo estándar

Buscar el sufijo más largo entre los siguientes, y llevar acabo la acción

anza anzas ico ica icos icas ismo ismos able ables ible ibles ista **istas** oso osa osos
amiento amientos imiento imientos

delete if in R2



Try the **Spanish** stemming algorithm:

admirables los allanamientos rosas sudorosas, listas generalistas **admir los allan ros sudor, list general**

Demo

Try the **Spanish** stemming algorithm:

José calcula los porcentajes del año **jos calcu los porcentaj del año**

| | | | | | | |
|------------------------|----------------------------|---------------------|------------------------------|---------------------------|-----------------------|-----------------------|
| José | calcula | los | porcentajes | del | año | Lematizador de AnCora |
| josé <i>NP00000</i> | calcular <i>VMIP350</i> | el <i>DAOMPO</i> | porcentaje <i>NCMP000</i> | de+el <i>SP+DAOMSO</i> | año <i>NCMS000</i> | |

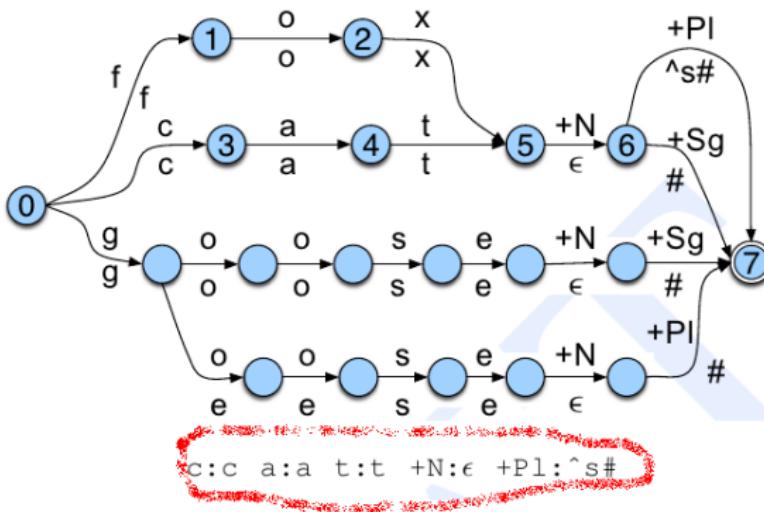
- Página del demo <http://snowballstem.org/demo.html>
- Página del Stemmer <http://snowball.tartarus.org/algorithms/spanish/stemmer.html>



Definición de A.M

Es el problema de reconocer que una palabra (*foxes*) se descompone en componentes de morfemas (*fox*, *es*), construyendo una representación estructural de este hecho.

cats → cat + N + PL



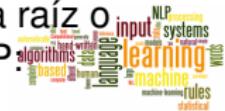
Análisis morfológico

El objetivo es tomar una entrada (palabras) y producir como salida la información morfológica de la entrada.

| English | | Spanish | | |
|---------|-------------------------------|---------|-------------------------------|---------------|
| Input | Morphologically Parsed Output | Input | Morphologically Parsed Output | Gloss |
| cats | cat +N +PL features | pavos | pavo +N +Masc +Pl | 'ducks' |
| cat | cat +N +SG | pavo | pavo +N +Masc +Sg | 'duck' |
| cities | city +N +Pl | bebo | beber +V +PInd +1P +Sg | 'I drink' |
| geese | goose +N +Pl | canto | cantar +V +PInd +1P +Sg | 'I sing' |
| goose | goose +N +Sg | canto | canto +N +Masc +Sg | 'song' |
| goose | goose +V | puse | poner +V +Perf +1P +Sg | 'I was able' |
| gooses | goose +V +1P +Sg | vino | venir +V +Perf +3P +Sg | 'he/she came' |
| merging | merge +V +PresPart | vino | vino +N +Masc +Sg | 'wine' |
| caught | catch +V +PastPart | lugar | lugar +N +Masc +Sg | 'place' |
| caught | catch +V +Past | | stems | |

Figure 3.2 Output of a morphological parse for some English and Spanish words. Spanish output modified from the Xerox XRCE finite-state language tools.

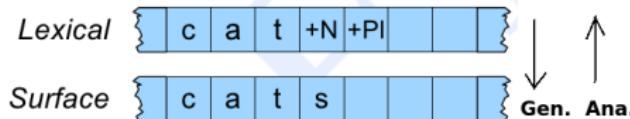
Las **features** especifican la información adicional acerca de la raíz o stem. Por ejemplo, +N :sustantivo; +Sg: singular, Pl: plural, 1P: primera persona, V: verbo, etc.



Tarea del análisis morfológico

For example, given a surface form of a word, it will give us the lexical form (analysis); or given the lexical form, it will give us the surface form (generation).

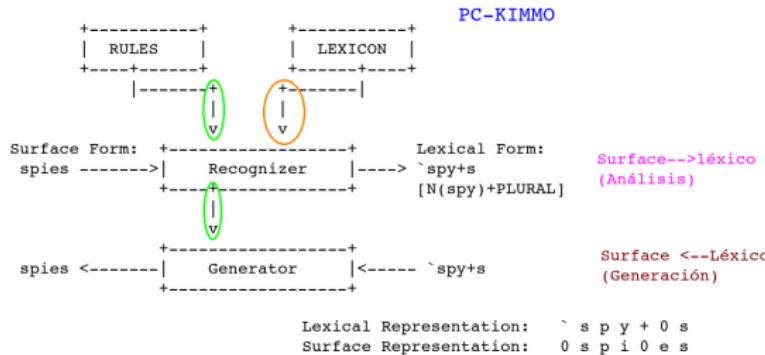
- Entonces representamos una palabra con su respectivo **nivel léxico (lexical level)**, el cual representa una concatenación de morfemas y el **nivel de superficie (surface level)** que representa la palabra procesada.
- Un FST tiene dos cintas; la de la parte superior o **cinta léxica**, y la parte inferior es la **cinta de superficial (surface tape)** o la **cinta intermedia**



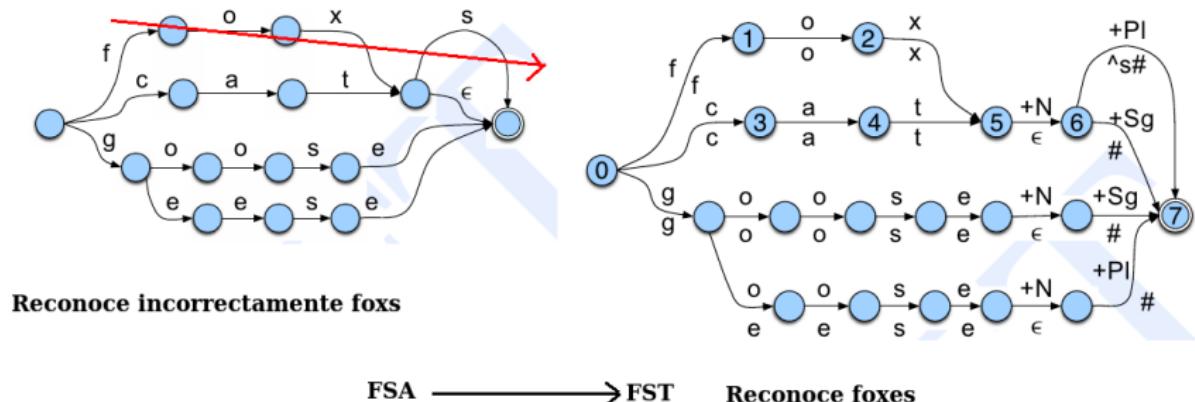
Partes de un analizador morfológico

En general, un analizador morfológico de dos niveles tiene las siguientes partes:

- ① Un **LEXICÓN** con las palabras, stems o afijos junto con información básica (si un stem es SUS o VERB) (glosas)
- ② **REGLAS**
 - Un conjunto de **reglas morfológicas** para entender que tipos de morfemas son precedentes o siguientes
 - Un conjunto de **reglas ortográficas** para la adición de sufijos y afijos: Se escribe con c: Las terminaciones -cito, -cita, -cillo, -cilla, -cecillo, -cecilla se escriben con c. Ejemplos: pedacito, nochecita, calzoncillo, manecilla, pececillo, lucecilla



Importancia de los transductores



- Se generan niveles: léxico, surface (input form) e intermedio
- Solamente se ven el léxico e intermedio que hace referencia a las reglas morfológicas de adición e inserción.

- Un transductor de estado finito FST es otra máquina de estado finito que produce una salida grabando la estructura de entrada.
- Un FST es esencialmente un FSA que trabaja sobre dos o más cintas.



$a : b$ en el arco significa que en esta transición el FST lee a desde la primera cinta y escribe b sobre la segunda cinta.

Funcionalidad de los transductores

El FST es una función más general que los FSAs, un FSA define un lenguaje formal a través de un conjunto de cadenas, y un FST define una **relación** entre conjuntos de cadenas.

Un transductor de estado finito FST es una 6-tupla

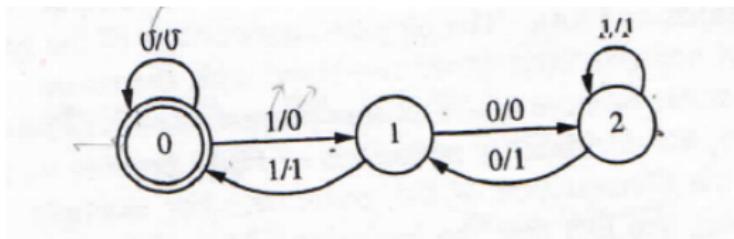
$T = (\Sigma_1, \Sigma_2, Q, q_0, F, E)$ tal que:

- Q es un conjunto finito de estados
- Σ_1 es el conjunto correspondiente al alfabeto de entrada.
- Σ_2 es el conjunto correspondiente al alfabeto de salida (los caracteres "b", "a" y los símbolos multi carácter $\langle n \rangle$, $\langle Sg \rangle$, $\langle PI \rangle$.)
- $q_0 \in Q$ es el estado inicial
- $F \subseteq Q$ conjunto de estados finales
- $E \subseteq Q \times \Sigma_1^* \times \Sigma_2^* \times Q$ es el conjunto de lados
- Sea δ una función de $Q \times \Sigma_1^*$ a 2^Q (Posibles subconjuntos de Q). Dado un estado $q \in Q$ y una cadena $w \in \Sigma_1^*$, entonces $\delta(q, w)$ retorna un conjunto de nuevos estados $Q' \in 2^Q$. Dado un estado $q \in Q$ y una cadena $w \in \Sigma_1^*$, entonces $\sigma(q, w)$ retorna un conjunto de cadenas de salida, donde cada cadena $\beta \in \Sigma_2^*$. σ es la función de salida de $Q \times \Sigma^*$ a $2^{\Sigma_2^*}$.



Ejemplo de un FST que representa la división binaria por 3

Sea la tupla $T = (\{0, 1\}, \{0, 1\}, \{0, 1, 2\}, \{0\}, \{0\})$ donde,
 $E = \{(0, 0, 0, 0), (0, 1, 0, 1), (1, 0, 0, 2), (1, 1, 1, 0), (2, 1, 1, 2), (2, 0, 1, 1)\}$,
 $\Sigma_1 = \Sigma_2 = \{0, 1\}$



Sea $\delta(0, 11) = \delta(\delta(0, 1), 1) = \delta(1, 1) = 0$ donde $0 \in Q$

Sea $\sigma(0, 11) = \sigma(\underbrace{\sigma(0, 1)}_0, 1) = \underbrace{\sigma(1, 1)}_1$

Reconoce cadenas tales como

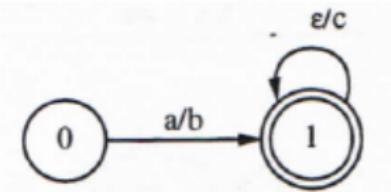
11(3)/01(1), 110(6)/010(2), 1001(9)/0011(3), 1100(12)/0100(4), etc.
101 no es reconocida.



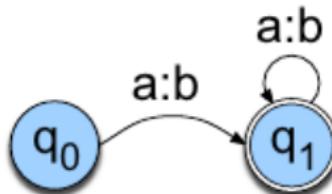
Ejemplo de FST

En a) el FST mapea la entrada a a un infinito conjunto de salidas correspondientes a la expresión regular bc^* (Una entrada es asociada con infinitas salidas)

a)

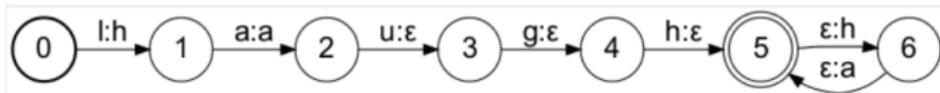


b)



En b) el FST mapea la entrada de a o a^* a un conjunto infinito de salidas correspondientes b^+ , tal que a^+ / b^+





La cadena de entrada **laugh** es mapeada al conjunto infinito
 $\{(ha)^n : n \geq 1\}$

(laugh,ha)
(laugh,haha)
(laugh,hahaha)



Existen diferentes modos de comportarse un FST:

- **FST generador:** Una máquina tiene como salida pares de cadenas de un lenguaje. Así la salida sea un sí o un no, y un par de cadenas de salida.
- **FST reconocedor:** El FST toma un par de palabras como entrada y salida; acepta si el par de cadenas está en el lenguaje de pares de cadenas, de lo contrario el FST rechaza.
- **FST Traductor:** Una máquina que lee una cadena y tiene como salida otra cadena.
- **FST Relacionador:** una máquina que computa relaciones entre conjuntos.

Máquinas de doble cinta

Otra manera de ver un FST es que es una máquina que lee una cadena y genera otra.



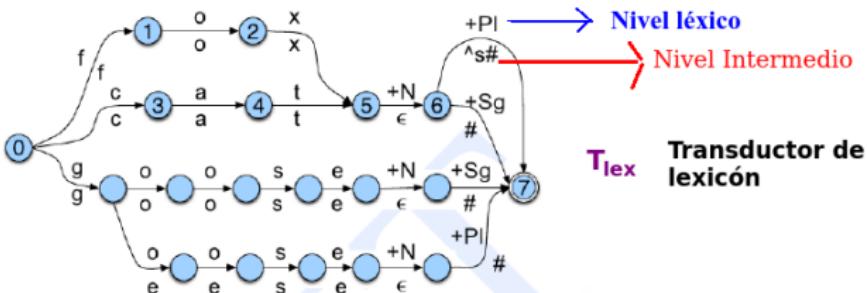
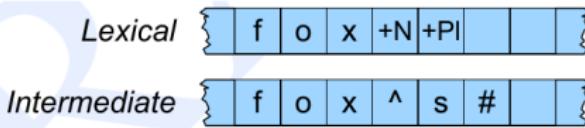
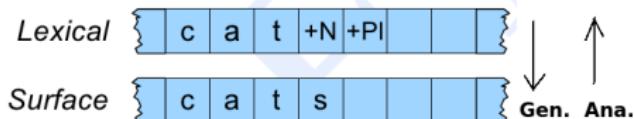
- Entonces combinamos los símbolos de los alfabetos Σ_1 y Σ_2 para crear un nuevo alfabeto Σ' , el cual es un alfabeto finito de símbolos complejos.
- Cada símbolo complejo está compuesto de un par entrada-salida $i : o$; donde $i \in \Sigma_1$ y $o \in \Sigma_2$. Así $\Sigma' \subseteq \Sigma_1 \times \Sigma_2$.
- Σ_1 y Σ_2 pueden incluir ϵ
- Entonces un FSA acepta un lenguaje sobre $\Sigma = \{b, a, !\}$ y un FST acepta una relación entre cadenas.

$$\Sigma' = \{a : a, b : b, ! : !, a : !, a : \epsilon, \epsilon : !\}$$

- Los pares de símbolos en Σ' son llamados pares viables (**Feasible pairs**). Así cada par viable $a : b$ expresa que el símbolo a a partir de una cinta se asigna al símbolo b en la otra cinta.



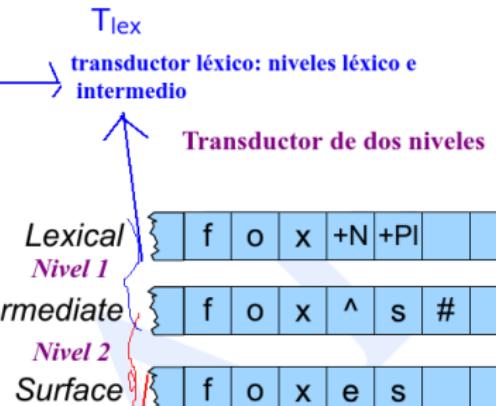
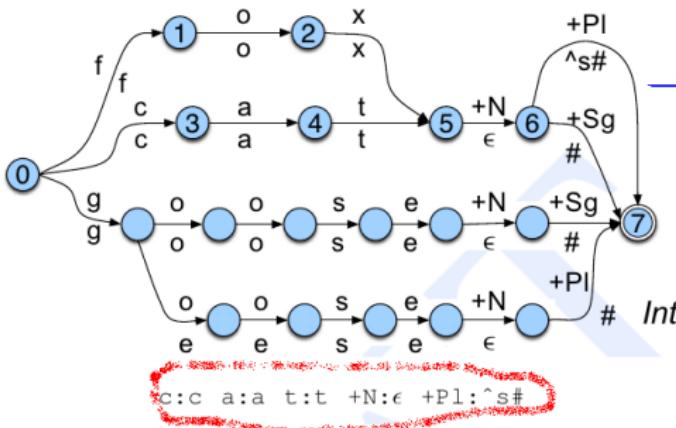
Generación y análisis en FSTs



c:c a:a t:t +N:ε +Pl:^s# → **cat +N +Pl /cats**

c:c a:a t:t +N:ε +Sg:# → **cat +N +Sg / cat**

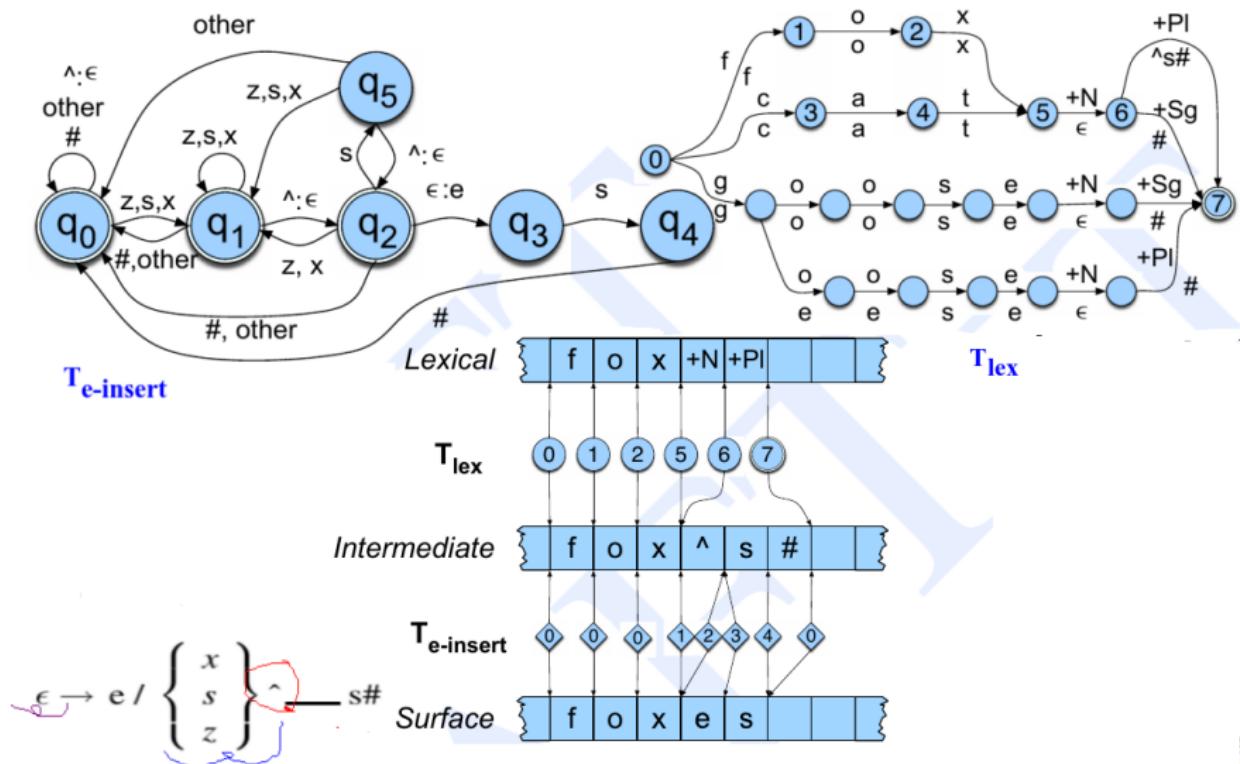
Transductores y reglas ortográficas



| Name | Description of Rule | Example |
|--------------------|--|----------------|
| Consonant doubling | 1-letter consonant doubled before -ing/-ed | beg/begging |
| E deletion | Silent e dropped before -ing and -ed | make/making |
| E insertion | e added after -s,-z,-x,-ch,-sh before -s | watch/watches |
| Y replacement | -y changes to -ie before -s, -i before -ed | try/tries |
| K insertion | verbs ending with vowel + -c add -k | panic/panicked |

Reglas de e-inserción: niveles intermedio y surface

Configuración que acepta el mapeo `fox +N +PL` a `foxes`



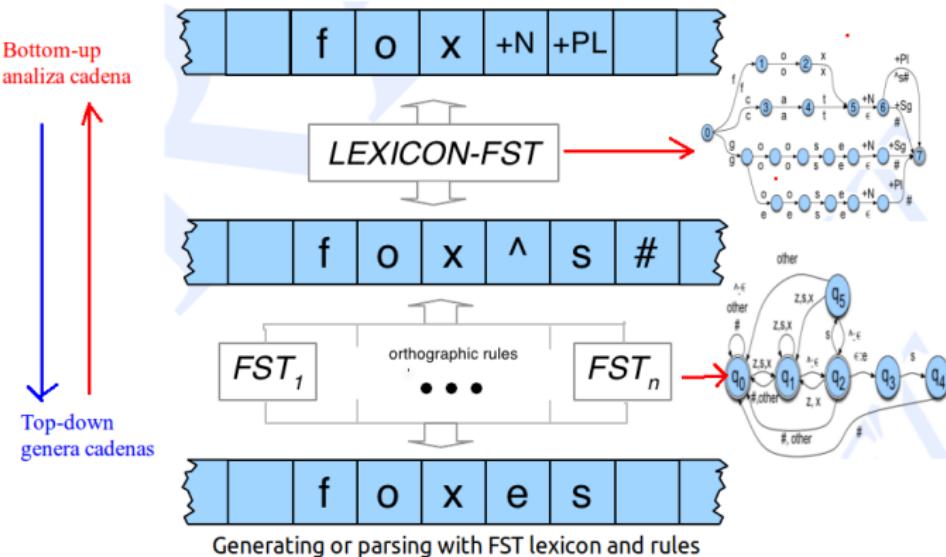
La regla de inserción-e inserta

Se inserta una **e** sobre la **cinta superficie** cuando la **cinta intermedia** tiene un morfema delimitador $\hat{}$ seguido por el morfema $-s$ y cuando la **cinta léxica** tiene un morfema finalizador x o (z, s, ch, sh) .



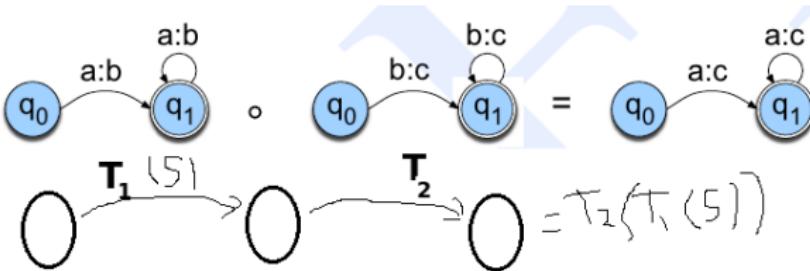
- 1 Esta regla de Chomsky-Halle (1968) es de la forma $a \rightarrow b|c\dots d$ significa que a se reescribe como b cuando este ocurre entre c y d
- 2 Entonces ϵ se reescribe por **e** (inserta una **e**) después del morfema x, z, s y antes del morfema s



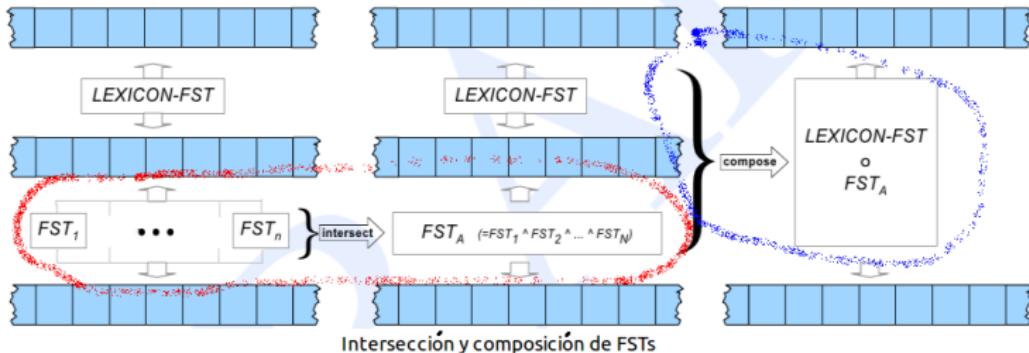


- Dos FSTs en cascada significa en **serie**, la salida de uno alimenta la entrada del otro. (**two-level cascade of transducers**)
- El sistema cascada tiene dos FSTs en serie: el FST que mapea los niveles léxico e intermedio y la colección de FSTs en **paralelo** que mapean los niveles intermedio y surface.





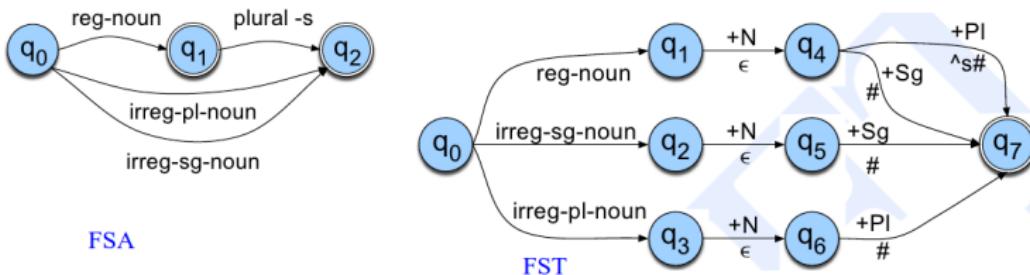
- La composición de $[a : b]_{T_1}^+$ con $[b : c]_{T_2}^+$ produce $[a : c]^+$.
- La **composición** es útil porque permite tomar dos transductores en serie y reemplazarlos con otro FST más complejo, es decir, $T_2 \circ T_1(S) = T_2(T_1(S))$, donde S es una secuencia de entrada.
- Sea $S = a^+$, entonces $T_1(S) = b^+$ y $T_2(b^+) = c^+$ ahora $T_2(T_1(S)) = c^+$ por lo tanto, $[a : c]^+$ es el resultado de aplicar $T_2(T_1(S))$ cuando $S = a^+$
- La **proyección** de un FST es el FSA que es producido extrayendo sólo un aldo de la relación.



- Los FSTs en paralelo pueden ser combinados por **intersección** de FSTs.
- La intersección de FSTs es el producto cartesiano de estados, para cada estado q_i en la máquina 1 y el estado q_j en la máquina 2, se crea un nuevo estado q_{ij} . Entonces para cualquier entrada a , si la máquina 1 hace transición al estado q_n y la máquina 2 hace transición a q_m la transición final es al estado q_{nm} .

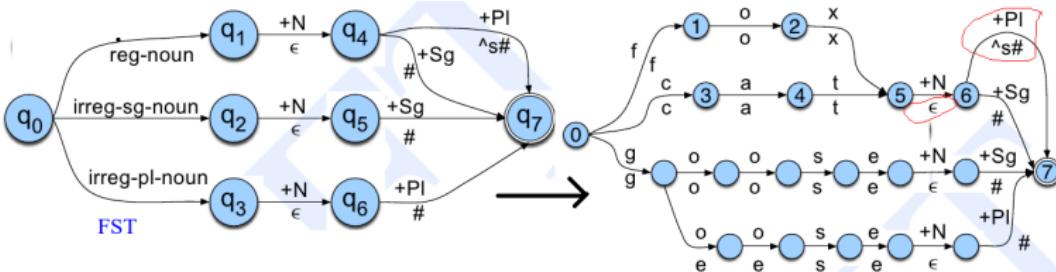
Ejemplo de cintas intermedias en un diccionario

- En las morfologías de dos niveles son definidos los pares predeterminados (**default pairs**) tales como $a : a$ los cuales consisten de una sola letra a .



| reg-noun | irreg-pl-noun | irreg-sg-noun |
|----------|-----------------|---------------|
| fox | g o:e o:e s e | goose |
| cat | sheep | sheep |
| aardvark | m o:i u:e s:c e | mouse |

- El FST es un FSA aumentado con características morfológicas (+Sg y +Pl) que corresponden a cada morfema. El símbolo \wedge indica un morfema límite (**morpheme boundary**), mientras que $\#$ indica una palabra límite (**word boundary**)



- Es importante ver el aumento del lexicón de FSTs de dos niveles: **léxico e intermedia**
- Cuando en la cinta intermedia no hay límites de morfema (morpheme boundary) es por la flexibilización de la palabra, por ejemplo; *fox*, *foxes* y *cat*, *cats* (en este caso si hay marcador) a diferencia de *geese* y *goose* (en este caso no hay marcador)
- Independiente de la arquitectura las palabras superficie mapean la palabra léxica