

Why NLP?

Introduction to Natural Language Processing
Computer Science 585—Fall 2009
University of Massachusetts Amherst

David Smith

Codes

following

finding

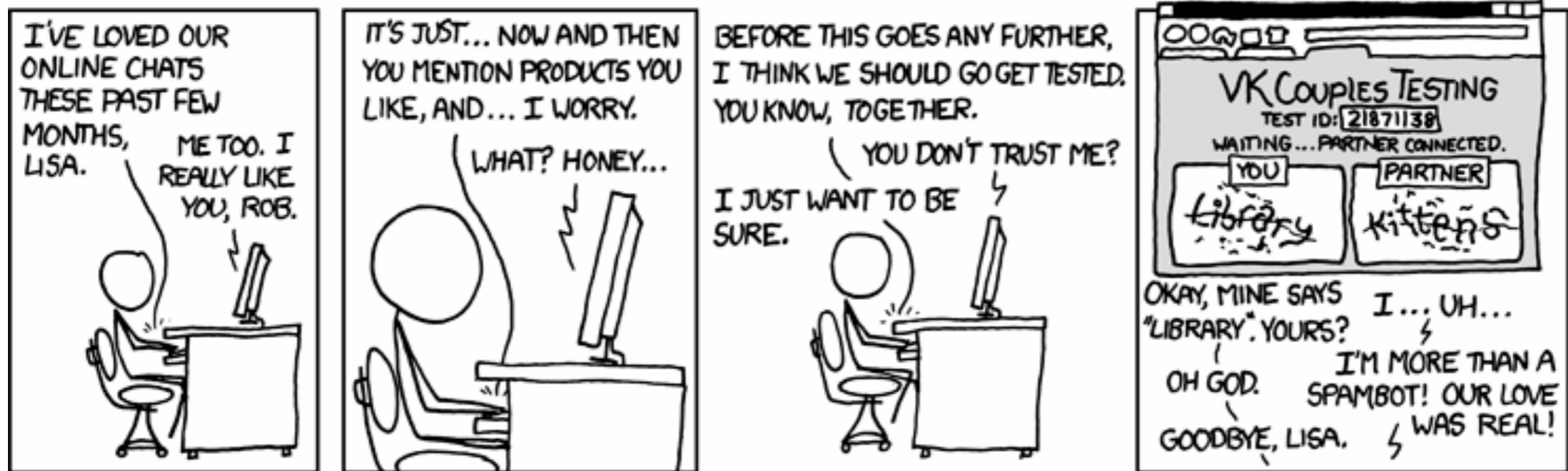
klpsh

3m573

v4p12

CAPTCHA

Completely Automated Public Turing test to tell
Computers and Humans Apart



Fine, walk away. I'm gonna go cry into a pint of Ben&Jerry's Brownie Batter(tm) ice cream [link], then take out my frustration on a variety of great flash games from PopCap Games(r) [link].



TENSE AND MOOD IN INDO-EUROPEAN SYNTAX*

1. THE HISTORICAL PRESENT

The 'historical' or 'dramatic' present tense used in narrating past events, which is common in many Indo-European languages, has always been interpreted in essentially semantic terms. A typical traditional formulation is

it is quite mistaken to transfer it to the earlier stages of Indo-European. In Greek, Old Irish, and Old Norse, for example, the historical present has quite different syntactic and semantic properties, to which the traditional idea, or any of its variants², must utterly fail to do justice.

* This work was supported in part by the Joint Services Electronics Program under Contract DA36-039-AMC-03200(E); in part by the National Science Foundation (Grant GP-2495), the National Institutes of Health (Grant MH-04737-05), the National Aeronautics and Space Administration (Grant NsG-496), and the U.S. Air Force (ESD Contract AF 19 (628)-2487). – I thank Michael Connolly, Eric Hamp, Einar Haugen, George Lakoff, Calvert Watkins, and Roy Wright for offering valuable criticism and/or referring me to some of the examples cited here.



Warren Weaver
to Norbert Wiener
4 March 1947

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: “**This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.**”

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHIEDUNGSPROBLEM

By A. M. TURING.

[Received 28 May, 1936.—Read 12 November, 1936.]

The “computable” numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. Although the subject of this paper is ostensibly the computable *numbers*, it is almost equally easy to define and investigate computable functions of an integral variable or a real or computable variable, computable predicates, and so forth. The fundamental problems involved are, however, the same in each case, and I have chosen the computable numbers for explicit treatment as involving the least cumbersome technique. I hope shortly to give an account of the relations of the computable numbers, functions, and so forth to one another. This will include a development of the theory of functions of a real variable expressed in terms of computable numbers. According to my definition, a number is computable if its decimal can be written down by a machine.

In §§ 9, 10 I give some arguments with the intention of showing that the

with the m -configuration written below the scanned symbol. The successive complete configurations are separated by colons.

:	e	e	0	:	e	e	0	:	e	e	0	:	e	e	0	:	e	e	0	:	0	1	:
b	v	q	q	q	p	p	f	f	v	v	f	v	v	f	v	v	v	f	v	v	v	v	v
e	e	0	0	1	:	e	e	0	0	1	:	e	e	0	0	1	:	e	e	0	0	1	:
p	p	f	f	v	v	f	v	v	f	v	v	v	v	f	v	v	v	f	v	v	v	v	v
e	e	0	0	1	:	e	e	0	0	1	:	e	e	0	0	1	:	e	e	0	0	1	:
e	e	0	0	1	:	e	e	0	0	1	:	e	e	0	0	1	:	e	e	0	0	1	:
e	e	0	0	1	:	e	e	0	0	1	:	e	e	0	0	1	:	e	e	0	0	1	:

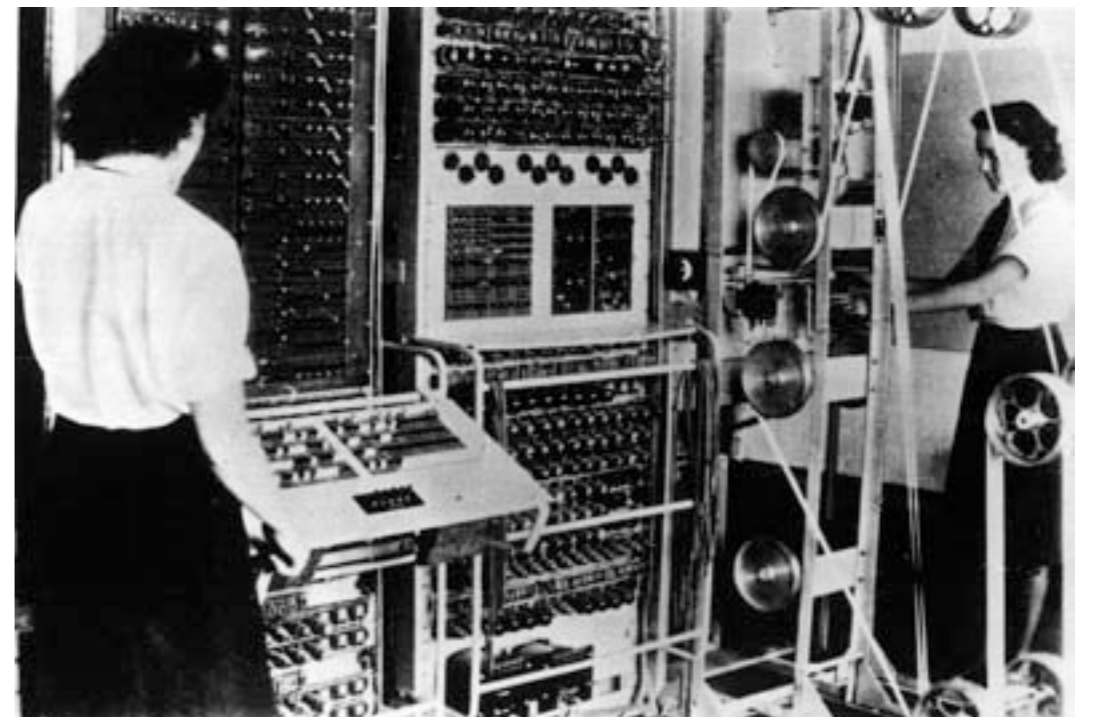
This table could also be written in the form

$b : e e v 0 \quad 0 : e e q 0 \quad 0 : \dots$ (C)





Bruderschwab, BDA 1811796-0220, IIA
Foto: Borchert, Bohn (Bfz) | 1945 Nr. - Juni



The Turing Test



The Turing Test



Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

The Turing Test



Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

The Turing Test



Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

The Turing Test



Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

The Turing Test



Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

The Turing Test



Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

The Turing Test



Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

The Turing Test



Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

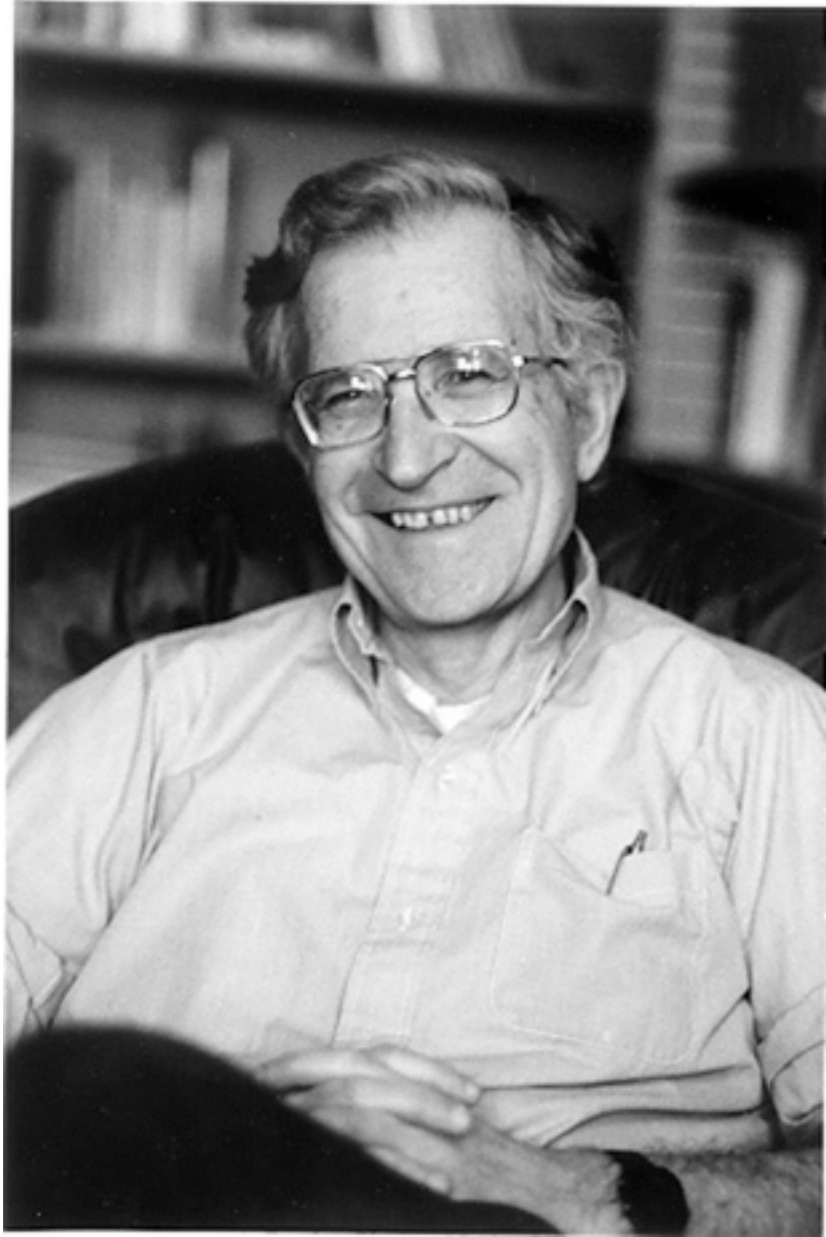
Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.



Modularity

Linguistic Modules

- Phonetics and phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse
- *With lots of crossings between levels!*

Phonetics and Phonology

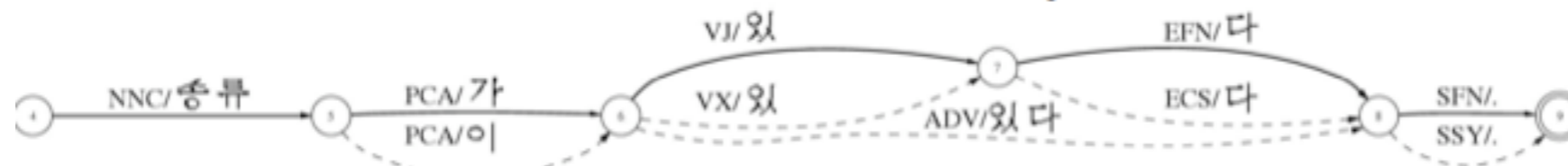
- Phonetics: language sounds & their physiology
- Phonology: systems of discrete sounds in languages
 - E.g.: devoicing of *it is* to *it's*
 - E.g.: syllable structure: *sign*, *signify*

Morphology

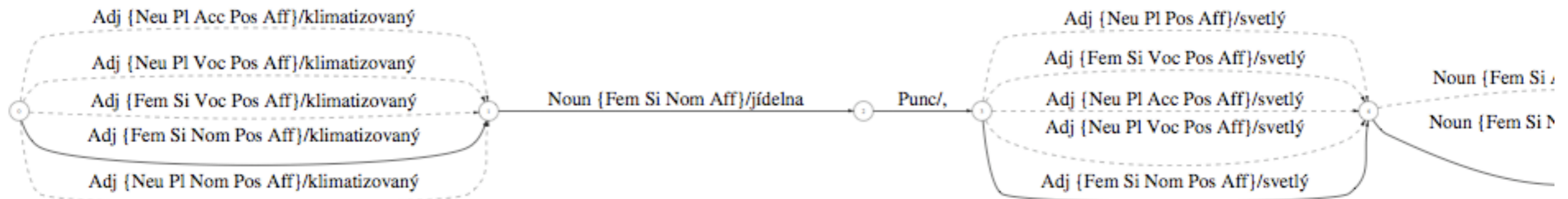
- Inflectional (in some languages):
 - *love → loved*
- Derivational:
 - *tea-cup, un-helpful, with-stand, craisin*
- Turkish: *uygarlastiramadiklarimizdanmissinizcasina*
 - *uygar las tir ama dik lar imiz dan mis siniz casina*
 - *(behaving) as if you are among those whom we could not civilize*

Morphological Tagging

There are many kinds of trench mortars.

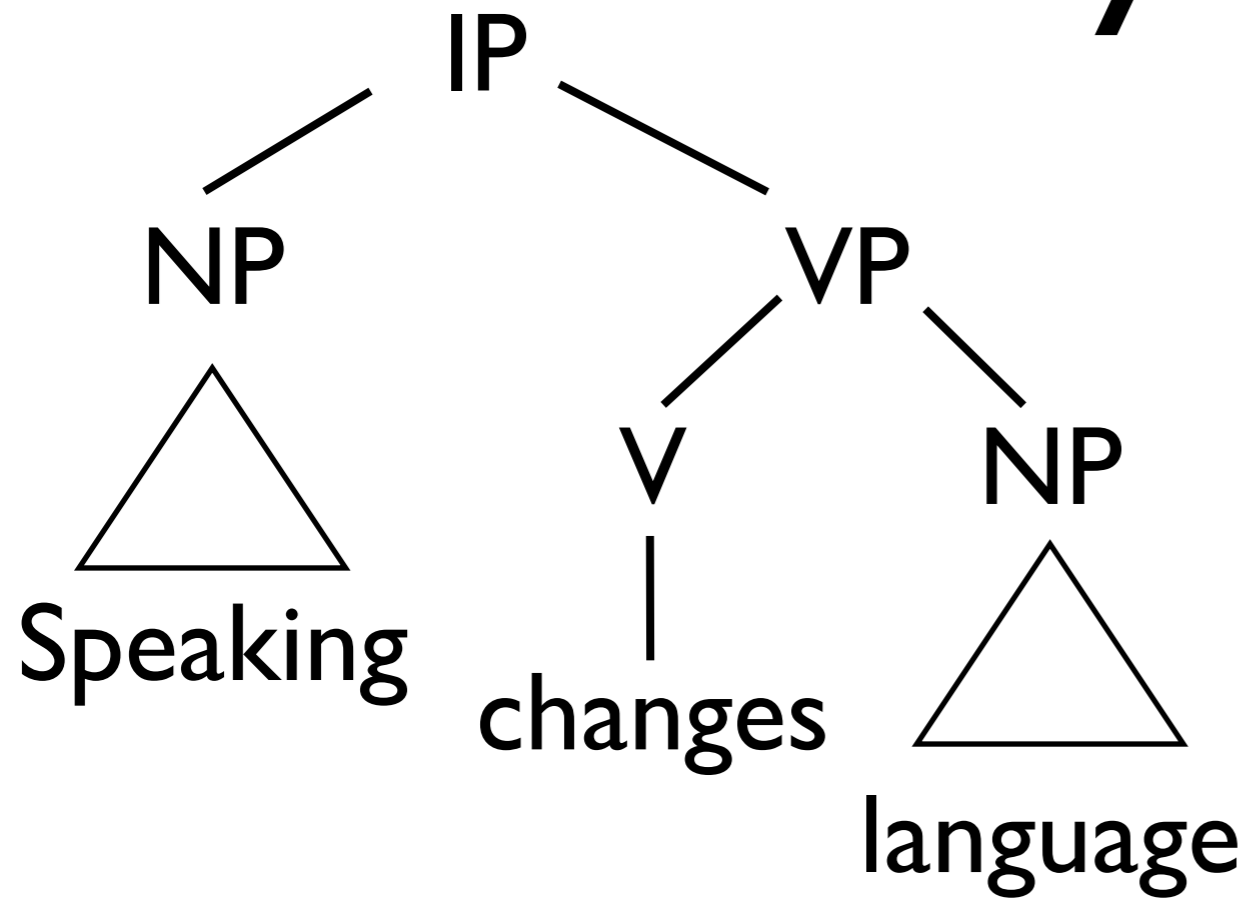


c. Klimatizovaná jídelna, světlá místnost pro snídani.



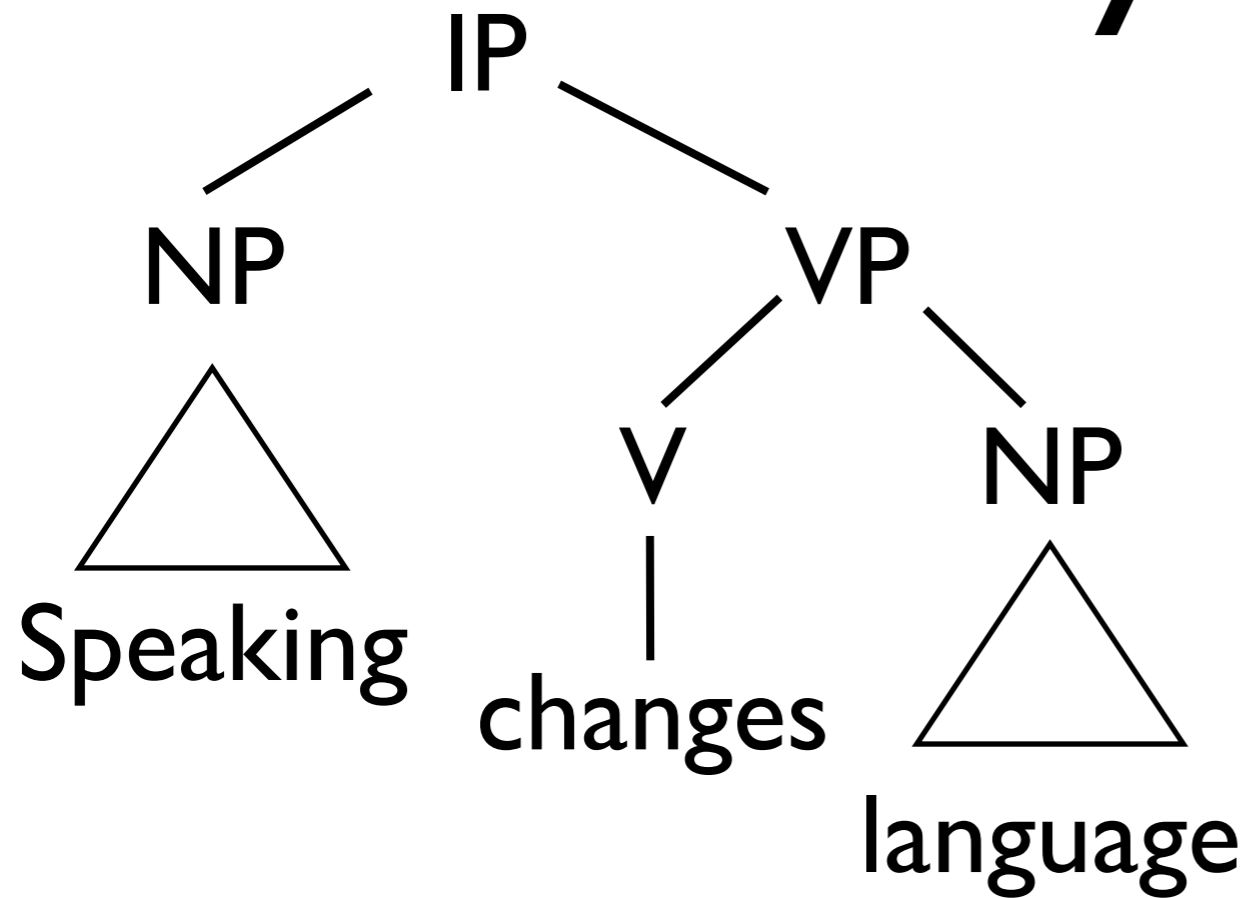
Syntax

Syntax



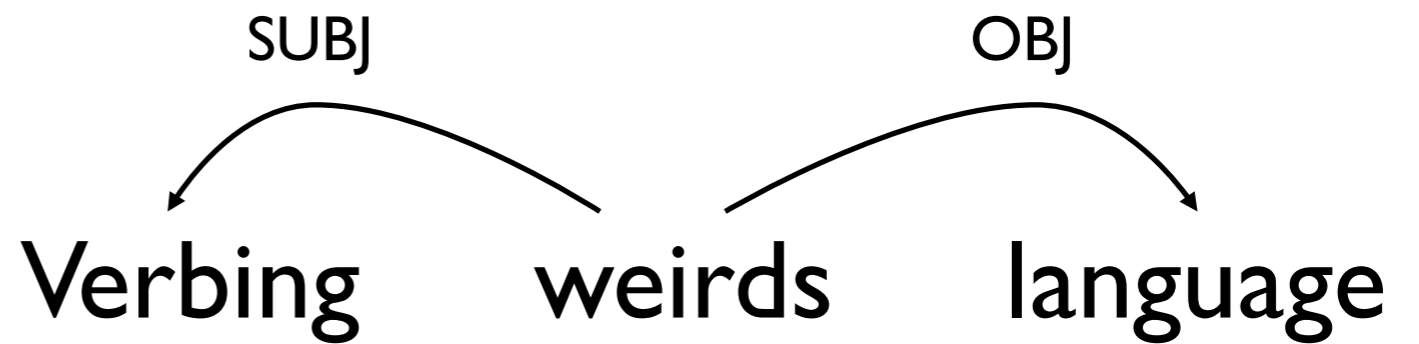
Constituency

Syntax

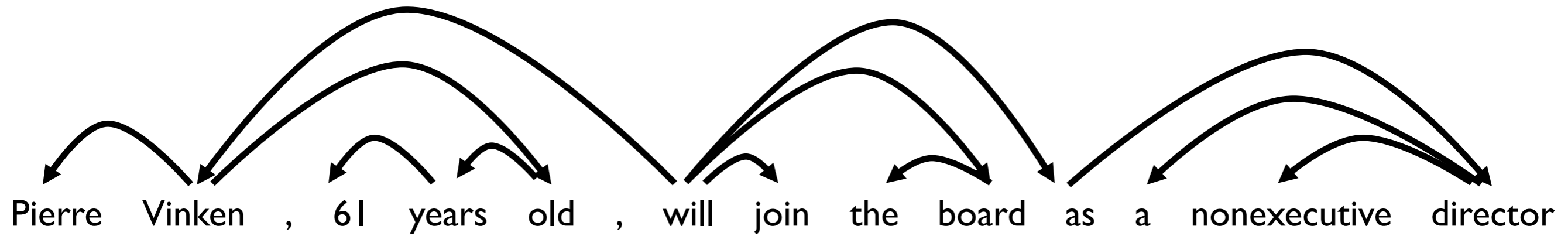


Constituency

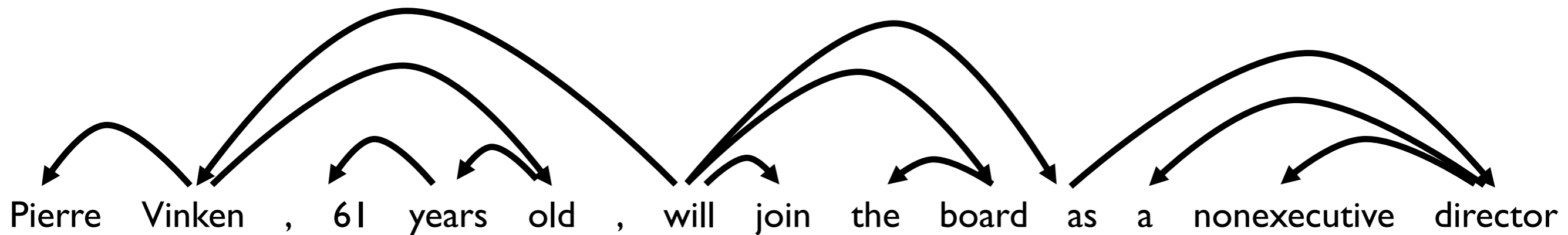
Dependency



Semantics



Semantics



PropBank **join** predicate

ARG0	ARG1	ARG-PRD
Vinken	board	director

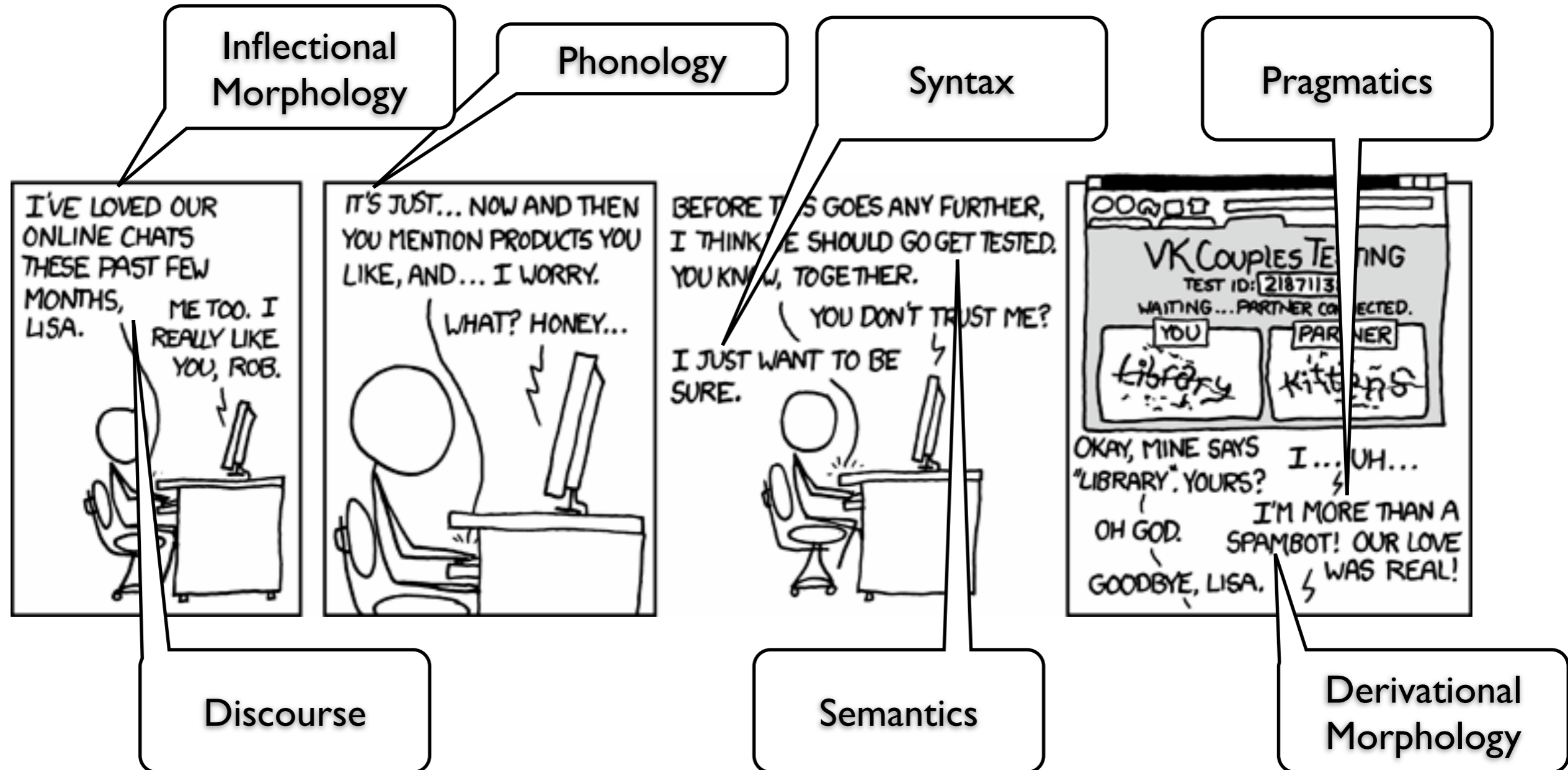
Pragmatics

- Context affects meaning
- Conversational implicature
 - *May I speak to your mother? Yes.*
- Speech acts: “how to do things with words”
 - *I grant you permission to speak.*

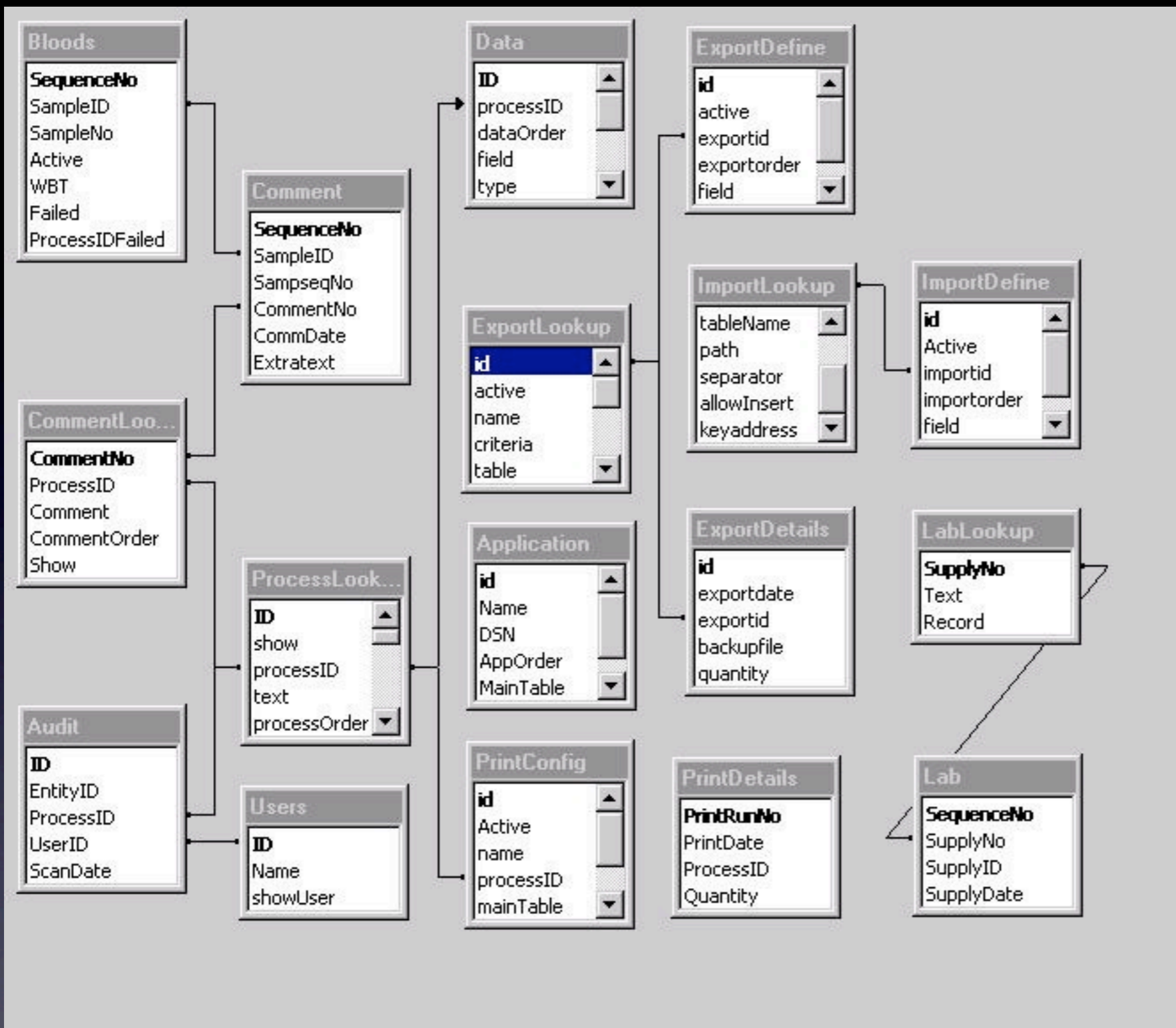
Discourse

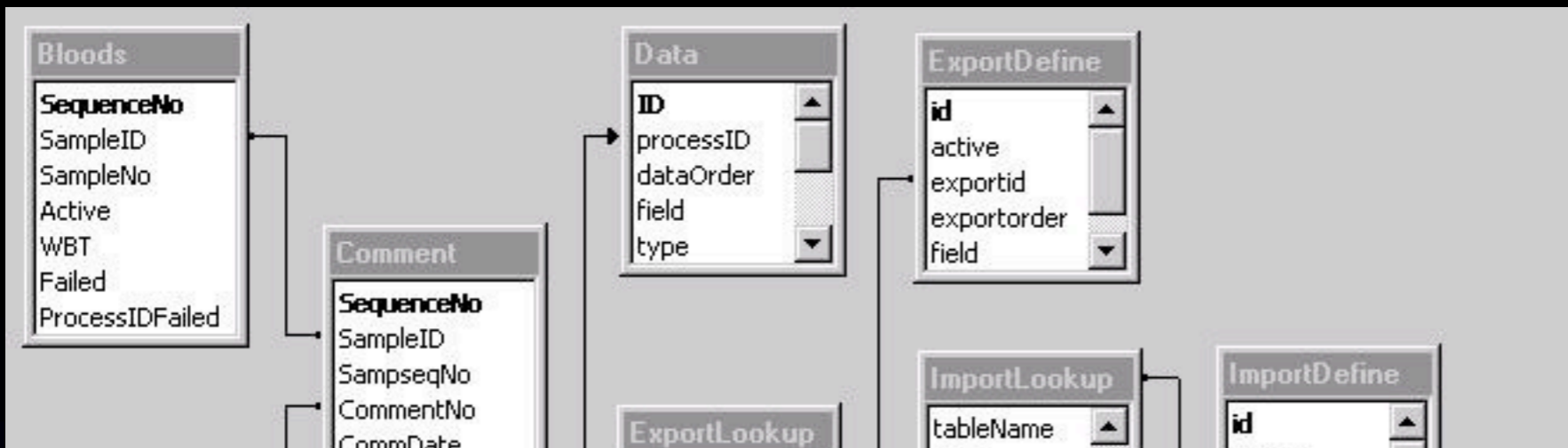
- Study of units larger than a single utterance
 - Turn taking
 - Coreference
 - Organized exposition

It All Hangs Together



Applications

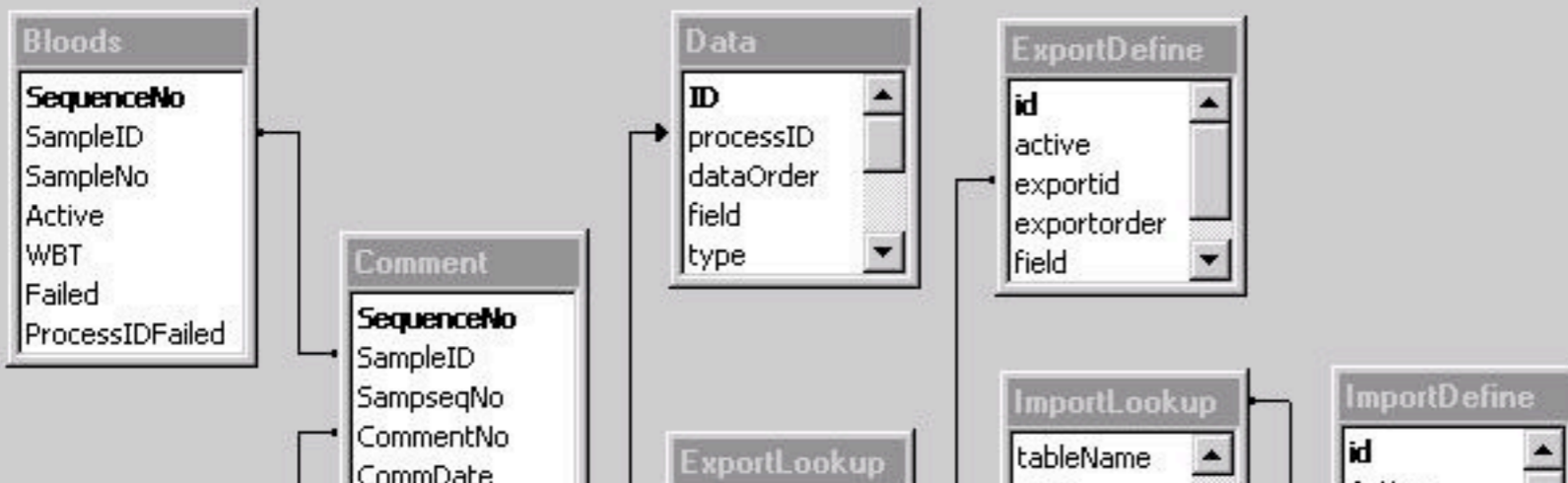




```

<bookstore>
<book category="COOKING">
  <title lang="en">Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>
<book category="CHILDREN">
  <title lang="en">Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>
<book category="WEB">
  <title lang="en">Learning XML</title>
  <author>Erik T. Ray</author>
  <year>2003</year>
  <price>39.95</price>
</book>
</bookstore>

```



```

<bookstore>
<book category="COOKING">
  <title lang="en">Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>
<book category="FICTION">
  <title lang="en">The Hobbit</title>
  <author>J. R. R. Tolkien</author>
  <year>2005</year>
  <price>29.99</price>
</book>
<book category="FICTION">
  <title lang="en">The Lord of the Rings: The Fellowship of the Ring</title>
  <author>Erik T. Johnson</author>
  <year>2003</year>
  <price>39.95</price>
</book>
</bookstore>

```

CommentL

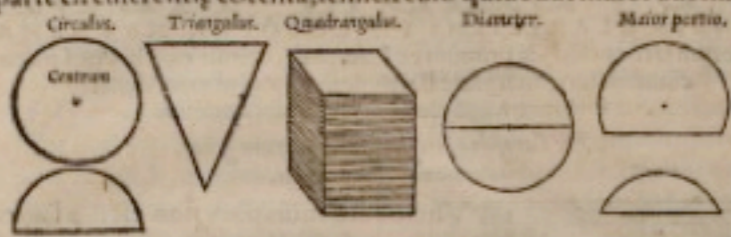
CommentNo
ProcessID
Comment
CommentOr
Show

Audit

ID
EntityID
ProcessID
UserID
ScanDate



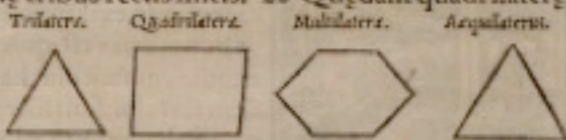
sunt æquales. 16 Et hic quidem punctus, centrum circuli dicitur.
 17 Diameter circuli, est linea recta, que super eius centrū transiens,
 extremitatesq; suas circūferētię applicās, circulū in duo media diuisi-
 dit. 18 Semicirculus, est figura plana diametro circuli, & medietate
 circūferētię cōtēta. 19 Portio circuli, est figura plana, recta linea &
 parte circūferētię cōtēta, semicirculo quidē aut maior aut minor.



Semicirculus.

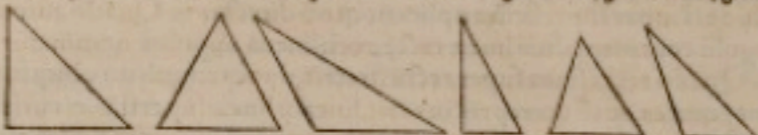
Minor portio.

20 Rectilineę figurę sunt, que rectis lineis continētur. 21 Quarum
 quedā trilaterę, que tribus rectis lineis: 22 Quedam quadrilaterę,
 q̄ quatuor rectis lineis: 23 Quedā
 multilaterę, que
 pluribus q̄ qua-
 tuor rectis lineis

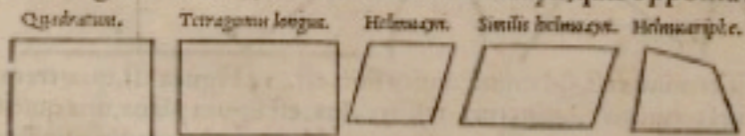


cōtinent. 24 Figurarū trilaterarū, alia est triāgulus, habēs tria latera
 æqualia: 25 Alia triangulus, duo habēs æqualia latera: 26 Alia trian-
 gulus triū inæqualiū laterū. 27 Harū iterū alia est orthogoniū, unū
 scilicet rectū angulū habēs. 28 Alia est amblygoniū, aliquē obtusum
 angulū habens. 29 Alia est oxygoniū, in qua tres anguli sunt acuti.

Dati æqualia latera. Triam inæqualium laterū. Orthogoniū. Oxygoniū. Amblygoniū.



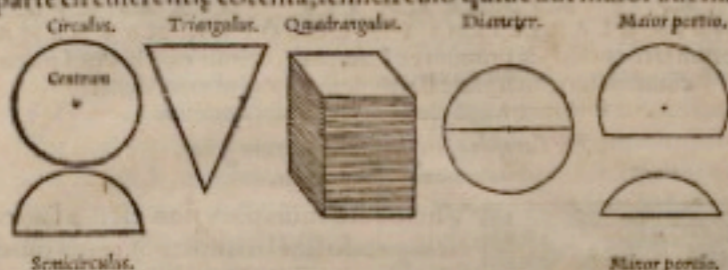
30 Figurarū autē quadrilaterarū, alia est quadratū, quod est æqui-
 laterū rectāguli: 31 Alia est tetragonus lōgus, que est figura rectān-
 gula, sed æquilatera nō est: 32 Alia est helmuayn, que est æquilatera,
 sed rectāgula nō est: 33 Alia est similis helmuayn, que opposita late-



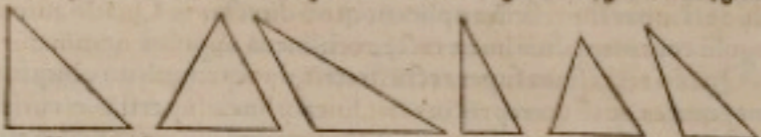
ra habet æqualia atq; oppositos angulos æquales, idē
 tamen nec rectis angulis nec æquis laterib. cōtinetur.

Præter

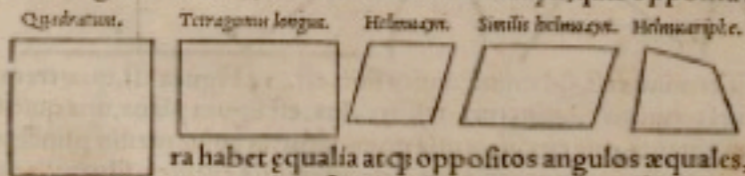
16 Et hic quidem punctus, centrum circuli dicitur.
 17 Diameter circuli, est linea recta, que super eius centrū transiens,
 extremitatesq; suas circūferētię applicās, circulū in duo media diuis
 dit. 18 Semicirculus, est figura plana diametro circuli, & medietate
 circūferētię cōtēta. 19 Portio circuli, est figura plana, recta linea &
 parte circūferētię cōtēta, semicirculo quidē aut maior aut minor.



20 Rectilineę figurę sunt, que rectis lineis continētur. 21 Quarum
 quedā trilaterę, que tribus rectis lineis: 22 Quedam quadrilaterę,
 q̄ quatuor rectis lineis: 23 Quedā
 multilaterę, quę
 pluribus q̄ qua-
 tuor rectis lineis
 cōtinent. 24 Figurarū trilaterarū, alia est triāgulus, habēs tria latera
 æqualia: 25 Alia triangulus, duo habēs æqualia latera: 26 Alia trian-
 gulus triū inæqualiū laterū. 27 Harū iterū alia est orthogoniū, unū
 scilicet rectū angulū habēs. 28 Alia est amblygoniū, aliquē obtusum
 angulū habens. 29 Alia est oxygoniū, in qua tres anguli sunt acuti.
 Dū æqualia latera. Triū inæqualia latera. Orthogoniū. Oxygoniū. Amblygoniū.



30 Figurarū autē quadrilaterarū, alia est quadratū, quod est æqui-
 laterū rectāgulū: 31 Alia est tetragonus lōgus, que est figura rectān-
 gula, sed æquilatera nō est: 32 Alia est helmuayn, que est æquilatera,
 sed rectāgula nō est: 33 Alia est similis helmuayn, quę opposita late-
 ra habet æqualia atq; oppositos angulos æquales, idē
 tamen nec rectis angulis nec æquis laterib. cōtinetur.



ra habet æqualia atq; oppositos angulos æquales, idē
 tamen nec rectis angulis nec æquis laterib. cōtinetur.

paper.pdf (9 pages)

7 Page Back/Forward Zoom In Zoom Out

Baselines	Dependency accuracy [%]			
	German		Spanish	
Modify prev.	18.2		28.5	
Modify next	27.5		21.4	
Training sentences	1k	10k	1k	10k
EM	30.2	30.8	25.6	24.9
Hard proj.	66.2	64.8	59.1	60.1
Hard proj. w/EM	58.6	59.8	53.0	52.8
QG w/EM	68.5	66.9	64.8	64.8

Table 2: Unlabeled dependency accuracy for German and Spanish with different training conditions and training set sizes.

to find sentences where enough links were projected to completely determine a target language tree. Of course, we needed to filter more than 1000 sentences of bitext to output 1000 training sentences in this way. With this subset, we can simply perform supervised training. As discussed in §2, these links are still quite noisy. Performance in fact suffers when we add more of this noisy training data. Still, this method is a substantial improvement over the baselines and unsupervised EM.

Instead of finding fully projected trees, we can simply take the one-to-one projected links are given, impute expected counts for the remaining structures with EM, and update our models. This approach ("hard projection with EM"), however, performed worse than using only the fully projected trees. In fact, only the first iteration of EM with this method made any improvement. Afterwards, EM degraded accuracy further from the numbers in table 2.

5.2 Unsupervised Learning

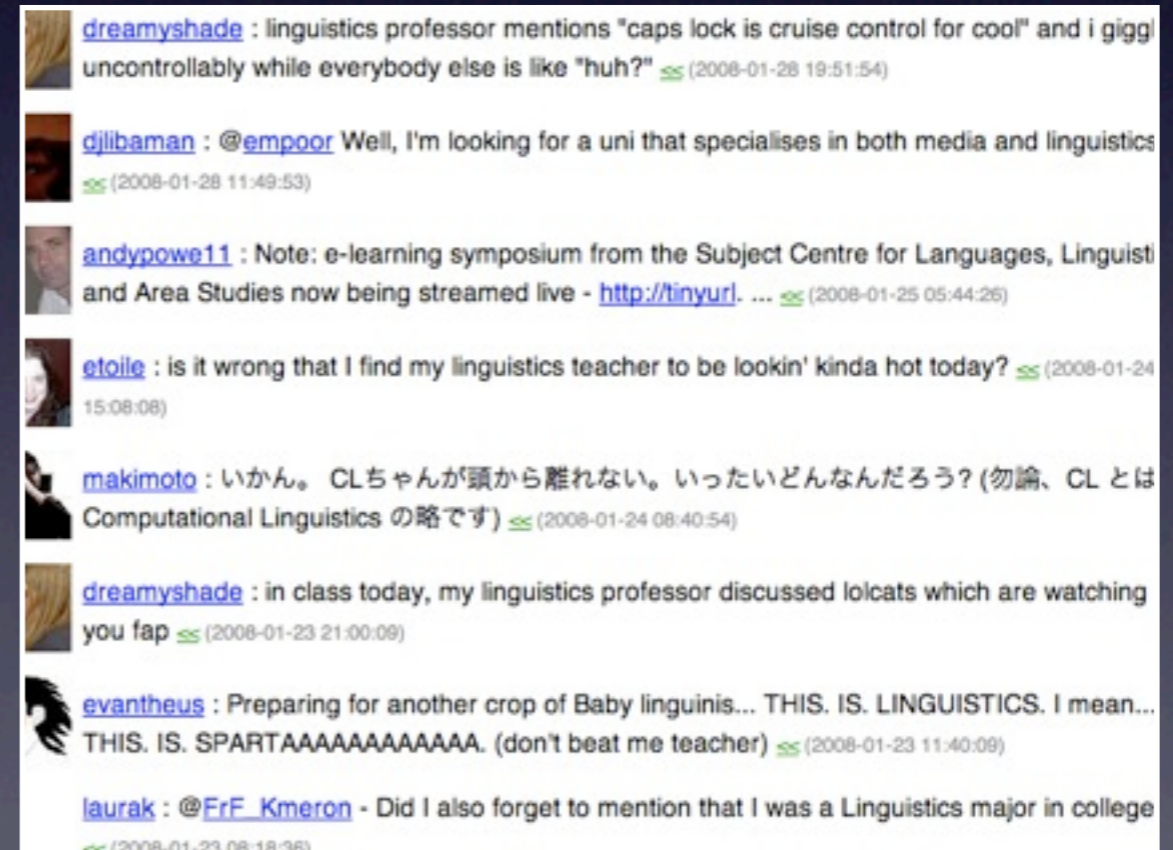
5.4 QG Projection

[Delete all spam messages now](#) (messages that have been in Spam more than 30 days will be automatically deleted)

Roland R. Duncan	Show your love to your special people! - EMAIL ID: ReJjr
=?UNKNOWN?B??=]	[JHSPAM-ALERT-IP] Piranesi 4 =?UNKNOWN?B?s8y3c6qpLW03Tw==?= - zT381OoDMFP3 Adobe Pt
Doctor Jeanne Holt	[JHSPAM-ALERT-IP] Something interesting for you - XMAILOE %XMIMEOE You can make your girl-frie
peggy.ruby	[JHSPAM-ALERT-IP]
Janie Adkins	[JHSPAM-ALERT-IP] Ficken wie ein Weltmeister ? - Meinung von unserem Kunden: Ich glaube, ich hab
Cary Phelps	No Hassle Business Loans - If you have your own business and need IMMEDIATE money to spend ANY \
Elias Moran	[JHSPAM-ALERT-IP] Loans - Small Business - If you have your own business and require IMMEDIATE €
Dr John Avery	[JHSPAM-ALERT-IP] Something interesting for you - XMAILOE %XMIMEOE X-Antivirus: avast! (VPS 08
Roland Reid	Business Loans - If you have your own business and need IMMEDIATE ready money to spend ANY way y
Doctor Max Stokes	[JHSPAM-ALERT-IP] It's important for you - XMAILOE %XMIMEOE Make your lady-love contented! You
¶W±]r«¬ μ∅±d∞Ñ°∂∞»!X∞@».	[JHSPAM-ALERT-IP] =?UNKNOWN?B?qr6mV6q6qL6scrNuxektUnFU?= - oOUQs3Xk Adobe After Effe



The screenshot shows the Wikipedia article for "Parsing". At the top left is the Wikipedia logo and navigation menu. The article title "Parsing" is prominently displayed. Below the title, it says "From Wikipedia, the free encyclopedia". The main text begins with a note: "For the computer programming language, see *Parser (CGI language)*." The article then defines parsing in computer science and linguistics, explaining that a parser is a component of a compiler that transforms input text into a data structure (usually a tree) and creates tokens from a sequence of input characters. It also mentions that parsing is an earlier term for the diagramming of sentences of natural languages or Latin. The article notes that parser generators are tools that can automatically generate a parser (in some cases, a compiler) and that parsers can be constructed as executable specifications of grammars in functional programming languages, often using higher-order functions called parser combinators.



The screenshot shows a forum thread with several posts. The first post is by user "dreamyshade" and discusses a linguistics professor's comment about "caps lock is cruise control for cool" and how the user giggled uncontrollably. The second post is by "djlibaman" looking for a university specializing in media and linguistics. The third post is by "andypowe11" about an e-learning symposium. The fourth post is by "etoile" asking if it's wrong to find their linguistics teacher hot. The fifth post is by "makimoto" in Japanese, asking about the difficulty of Computational Linguistics. The sixth post is by "dreamyshade" about a linguistics professor discussing lolcats. The seventh post is by "evantheus" expressing their love for linguistics. The eighth post is by "laurak" mentioning they were a linguistics major in college.

Would you like to...



[Add to calendar](#)

Feature space Maximu...

Fri Feb 8 12pm – Fri F...

Would you like to...



Dar al hayat
ENGLISH

PDF



PDF



الحياة

الطبعة السعودية

دار الحياة

F

F

المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق



الادعاء يطلب أقصى عقوبة لثلاثة متهمين بالتخطيط لاغتيال بوش

عمان الحياة - //07/02/08

طالب مدعي عام محكمة أمن الدولة الأردنية بتوقيع أقصى عقوبة ممكنة على ثلاثة متهمين بالتخطيط لاغتيال الرئيس الأميركي جورج بوش أثناء زيارته عمان في تشرين الثاني (نوفمبر) 2006، بعدما وجه إليهم اتهامات بينها «القيام بأعمال إرهابية، واستخدام مواد ملتهبة، وحمل وحيازة أسلحة اتوماتيكية ومفرقات من دون ترخيص بقصد استعمالها على وجه غير مشروع».

وتلا رئيس المحكمة تقريراً طبياً صادراً عن لجنة قيمت الوضع العقلي للمتهم صطام الزواهره الذي طلب محاميه إحالته على لجنة لمعرفة مدى إدراكه لطبيعة أفعاله وإمكان مثوله أمام المحكمة. وأثبت التقرير أن «المتهم مدرك لأفعاله وأقواله». وطلب وكيل الدفاع المحامي عبدالكريم الشريفة من المحكمة إمهاله لتقديم بلاغته في جلسة 13 من الشهر الحالي.

Would you like to...



Dar al hayat
ENGLISH

PDF



PDF

الحياة

الطبعة السعودية

دار الحياة

F
F

المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق



الادعاء يطلب أقصى عقوبة لثلاثة متهمين بالتخطيط لاغتيال بوش

عمان الحياة - 07/02/08

ابحث

ENGLISH

دار الحياة

على ثلاثة متهمين بالتخطيط لاغتيال
نمبر) 2006، بعدما وجه إليهم
وحيازة أسلحة اتوماتيكية ومفرقات

للمتهم صطام الزواهره الذي طلب
« أمام المحكمة. وأثبت التقرير أن
بم الشريفة من المحكمة إمهاله
تسعين الثلاثة من الزواهره (28)

Prosecution requests the maximum penalty for the three accused of planning to assassinate Bush

Oman life - 07/02/08 / /

Asked prosecutor in the state security court of Jordan signed the maximum possible sentence to three accused of planning to assassinate President George Bush during his visit to Amman in November (November) 2006, after it drew them was charged with «terrorist acts, the use of flammable materials, and carrying and possessing automatic weapons and explosives without a licence in order to use illegal ».

The President of the Court read out a medical report issued by the Commission assessed the mental status of the accused Azwahrh Stam, who asked his lawyer referred to a committee to determine the extent aware of the nature of his acts and the possibility of his appearance before the court. The report proved that «the accused was aware of his actions and sayings». Asked assistant defense lawyer `court delayed by the court to

Would you like to...

Dar al hayat ENGLISH PDF الحياة الطبعة السعودية دار الحياة

المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق المشرق

الادعاء يطلب أقصى عقوبة لثلاثة متهمين بالتخطيط لاغتيال بوش

عمان الحياة - //07/02/08

ابحث

على ثلاثة متهمين بالتخطيط لاغتيال (نمذ 2006، بعدما وجه الهم

دار الحياة



machine translation

Search

[Advanced Search](#)
[Preferences](#)

Web

Results 1 - 10 of about 2,790,000 for machine translation [definition]. (0.15 seconds)

[Machine translation - Wikipedia, the free encyclopedia](#)

Machine translation, sometimes referred to by the acronym **MT**, is a sub-field of computational linguistics that investigates the use of computer software to ...
en.wikipedia.org/wiki/Machine_translation - 56k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Free Online Translator](#)

Free Online Language **Translation**. Translates from English, Chinese Simplified, Chinese Traditional, Dutch, French, German, Greek, Italian, Japanese, Korean, ...
www.worldlingo.com/products_services/worldlingo_translator.html - 12k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Machine Translation Engine](#)

TranslateNow! Access 20 on-line **machine translation** systems from a single screen. Have various online **machine translation** systems translate your texts ...
www.foreignword.com/Tools/transnow.htm - 3k - [Cached](#) - [Similar pages](#) - [Note this](#)

[machine translation](#)

www.springerlink.com/openurl.asp?genre=journal&issn=0922-6567 - [Similar pages](#) - [Note this](#)

Sponsored Links

[Machine Translation](#)

Google is looking for Engineering experts to join our team. Apply!
www.google.com/jobs

[Language Weaver](#)

High quality, statistically-based **translation** software.
www.languageweaver.com

[Translation Software](#)

Spanish, French, German, Japanese, Chinese, Korean, Polish, Russian.
www.allvirtualware.com

[Free Translation Software](#)

Download your free trial of SYSTRAN

Translation

Translation

Er wird in den Strassen wandern

Translation

Er wird in den Strassen wandern

He will in the streets walk

Translation

Er wird in den Strassen wandern

He will in the streets walk

He will walk in the streets



Translation

Er wird in den Strassen wandern

He will in the streets walk

He will walk in the streets



Er wird in den **kleinen** Strassen wandern

Translation

Er wird in den Strassen wandern

He will in the streets walk

He will walk in the streets



Er wird in den **kleinen** Strassen wandern

He will in the small streets walk

Translation

Er wird in den Strassen wandern

He will in the streets walk

He will walk in the streets



Er wird in den **kleinen** Strassen wandern

He will in the small streets walk

He is in the small streets hike



Translation



Er wird in den Strassen wandern

He will in the streets walk

He will walk in the streets



Er wird in den **kleinen** Strassen wandern

He will in the small streets walk

He is in the small streets hike



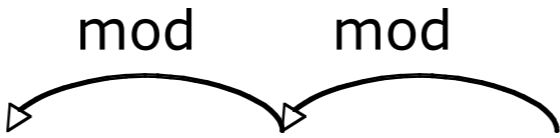
Question Answering

Who is the leader of France ?

The diagram illustrates the syntactic structure of the question. It shows two dependency arcs, both labeled 'mod', indicating a modification relationship. The first arc connects the word 'leader' to the word 'of'. The second arc connects the word 'of' to the word 'France'. The words 'leader' and 'France' are highlighted in green in the original image.

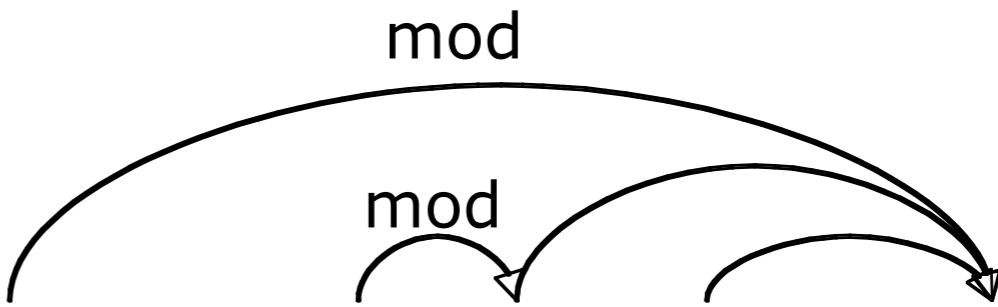
Question Answering

Who is the leader of France ?



The diagram shows two curved arrows labeled 'mod' above the words 'leader' and 'France'. One arrow points from 'leader' to 'France', and the other points from 'France' to 'leader'.

Henri Hadjenberb , who is the leader of France 's Jewish community



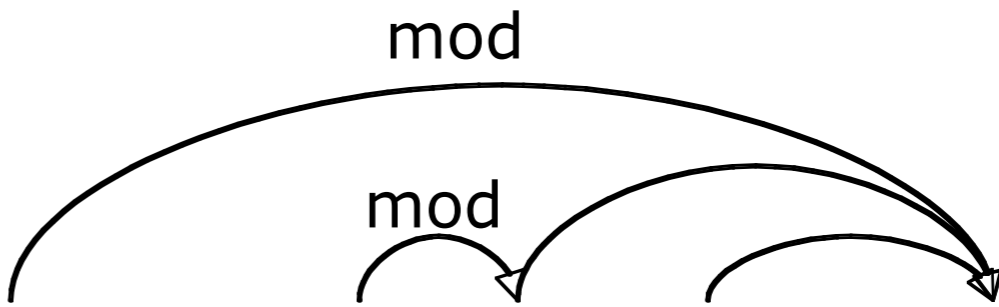
The diagram shows three curved arrows labeled 'mod' above the words 'leader', 'France', and 'Jewish community'. One arrow points from 'leader' to 'France', another from 'France' to 'Jewish community', and a third from 'leader' to 'Jewish community'.

Question Answering

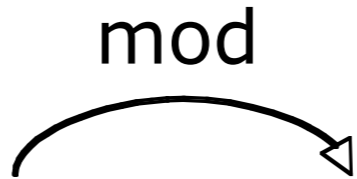
Who is the leader of France ?



Henri Hadjenberb , who is the leader of France 's Jewish community



Bush met with French president Jacques Chirac



Multilingual “Topics” European Parliament Corpus

DA børn familie udnyttelse børns børnene seksuel
DE kinder kindern familie ausbeutung familien eltern
EL παιδιά παιδιών οικογένεια οικογένειας γονείς παιδικής
EN **children family child sexual families exploitation**
ES niños familia hijos sexual infantil menores
FI lasten lapsia lapset perheen lapsen lapsiin
FR enfants famille enfant parents exploitation familles
IT bambini famiglia figli minori sessuale sfruttamento
NL kinderen kind gezin seksuele ouders familie
PT crianças família filhos sexual criança infantil
SV barn barnen familjen sexuellt familj utnyttjande

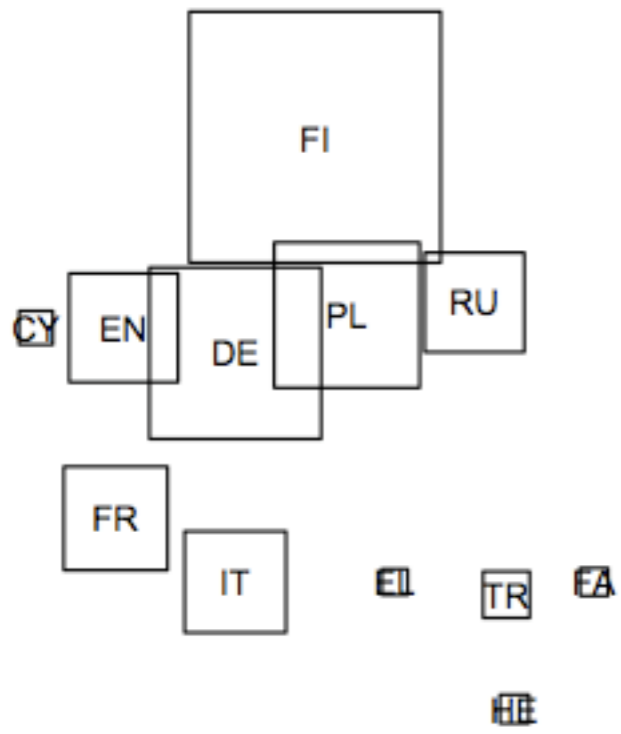
DA mål nå målsætninger målet målsætning opnå
DE ziel ziele erreichen zielen erreicht zielsetzungen
EL στόχους στόχο στόχος στόχων στόχοι επίτευξη
EN **objective objectives achieve aim ambitious set**
ES objetivo objetivos alcanzar conseguir lograr estos
FI tavoite tavoitteet tavoitteena tavoitteiden tavoitteita tavoitteen
FR objectif objectifs atteindre but cet ambitieux
IT obiettivo obiettivi raggiungere degli scopo quello
NL doelstellingen doel doelstelling bereiken bereikt doelen
PT objetivo objetivos alcançar atingir ambicioso conseguir
SV mål målet uppnå målen målsättning målsättning

DA andre anden side ene andet øvrige
DE anderen andere einen wie andererseits anderer
EL άλλες άλλα άλλη άλλων άλλους όπως
EN **other one hand others another there**
ES otros otras otro otra parte demás
FI muiden toisaalta muita muut muihin muun
FR autres autre part côté ailleurs même
IT altri altre altro altra dall parte
NL andere anderzijds anderen ander als kant
PT outros outras outro lado outra noutros
SV andra sidan å annat ena annan

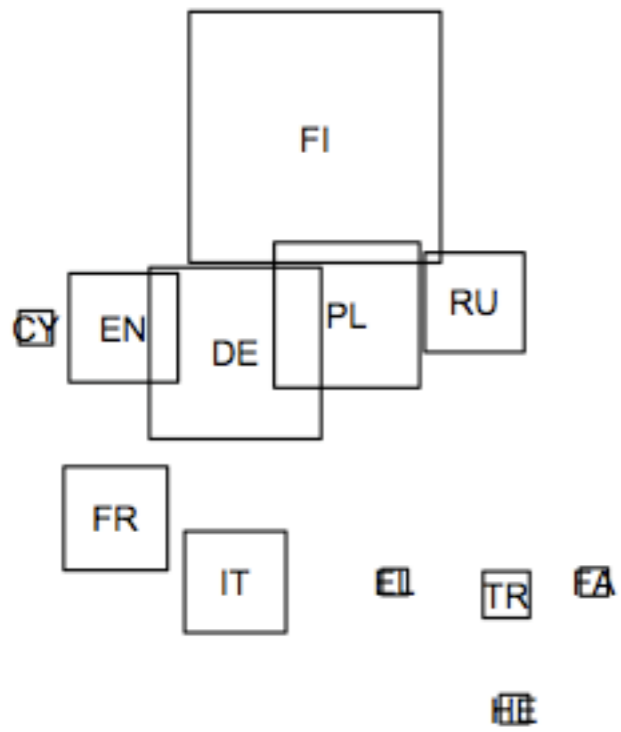
Multilingual “Topics” Wikipedia comparable articles

CY sbaen madrid el la josé sbaeneg
DE de spanischer spanischen spanien madrid la
EL ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN **de spanish spain la madrid y**
FA اسپانيا اسپانيايي کوبا مادريد
FI espanja de espanjan madrid la real
FR espagnol espagne madrid espagnole juan y
HE ספרד ספרדית דה מדריד הספרדית קובה
IT de spagna spagnolo spagnola madrid el
PL de hiszpański hiszpanii la juan y
RU де мадрид испании испания испанский de
TR ispanya ispanyol madrid la küba real

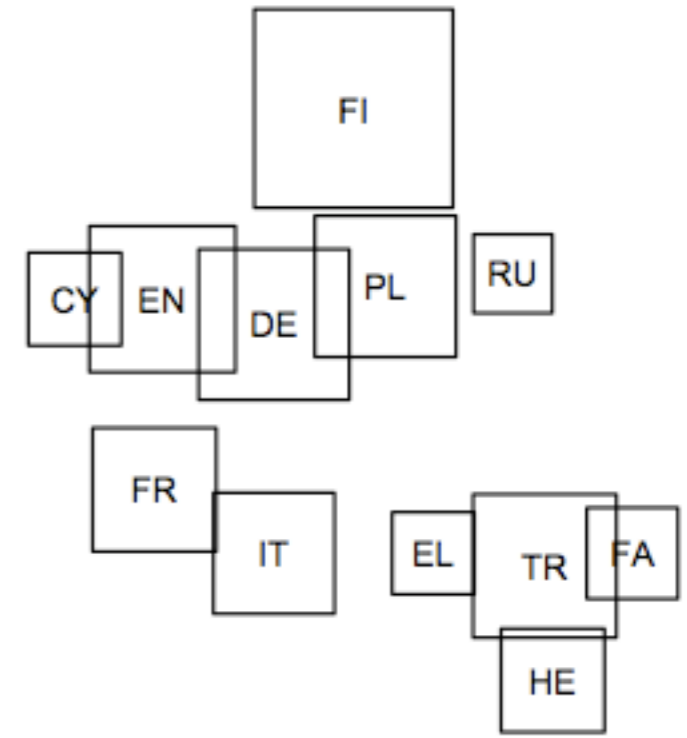
CY bardd gerddi iaith beirdd fardd gymraeg
DE dichter schriftsteller literatur gedichte gedicht werk
EL ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN **poet poetry literature literary poems poem**
FA شاعر شعر ادبيات فارسی ادبی آثار
FI runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR poète écrivain littérature poésie littéraire ses
HE משורר ספרות שירה סופר שירים המשורר
IT poeta letteratura poesia opere versi poema
PL poeta literatury poezji pisarz in jego
RU поэт его писатель литературы поэзии драматург
TR şair edebiyat şiir yazar edebiyatı adlı



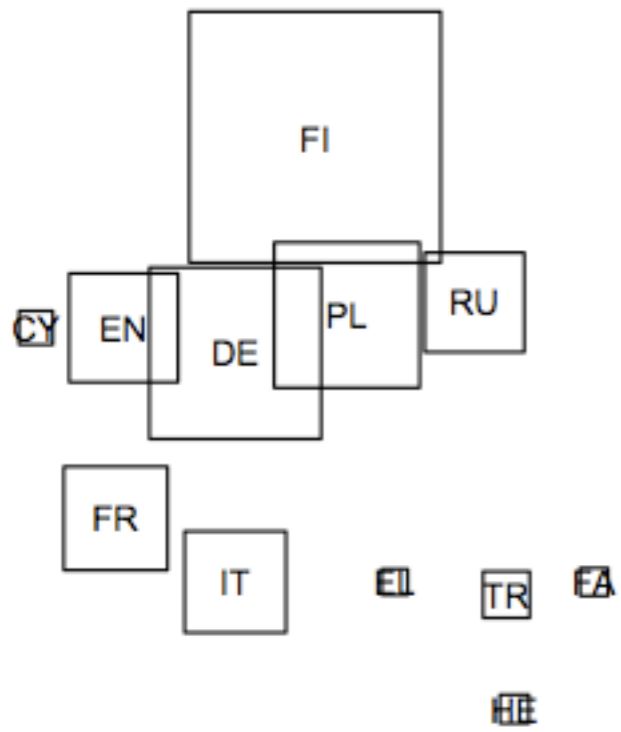
world ski km won



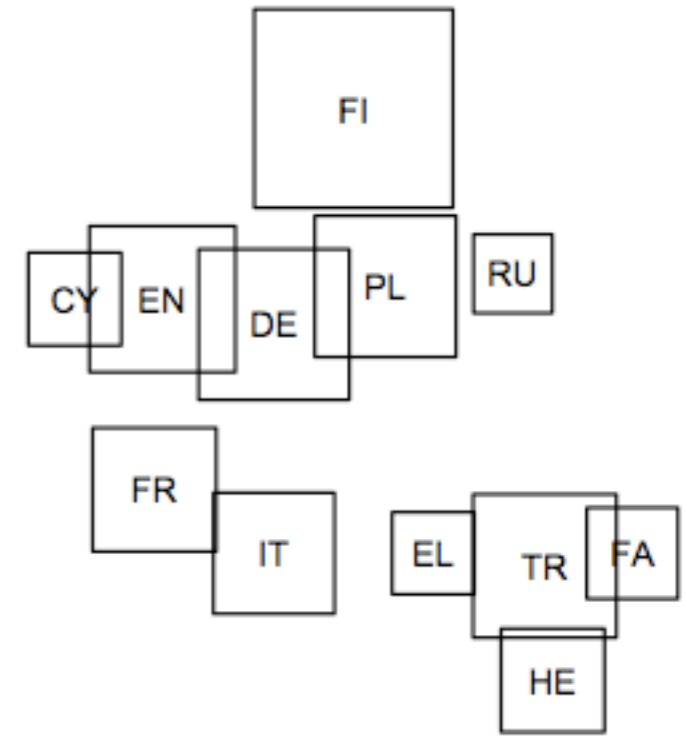
world ski km won



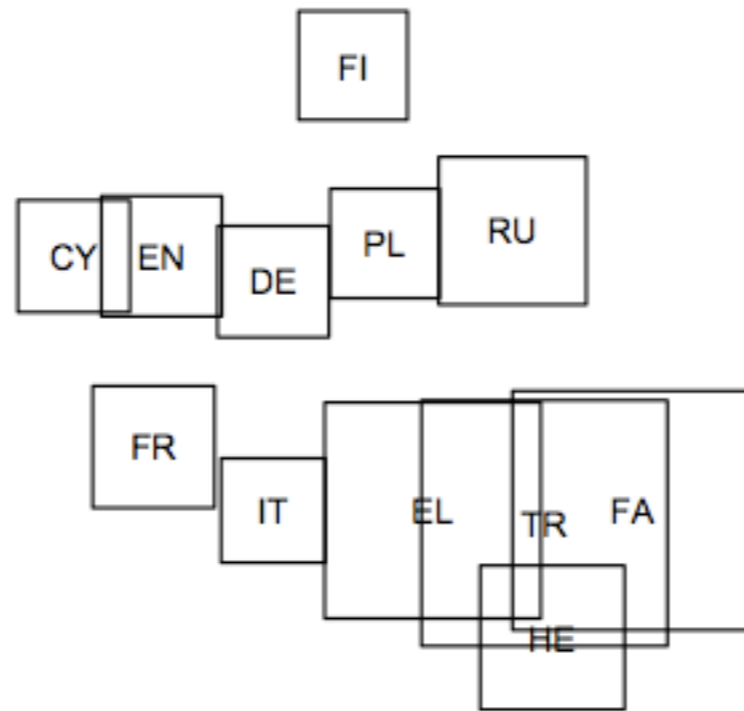
actor role television actress



world ski km won



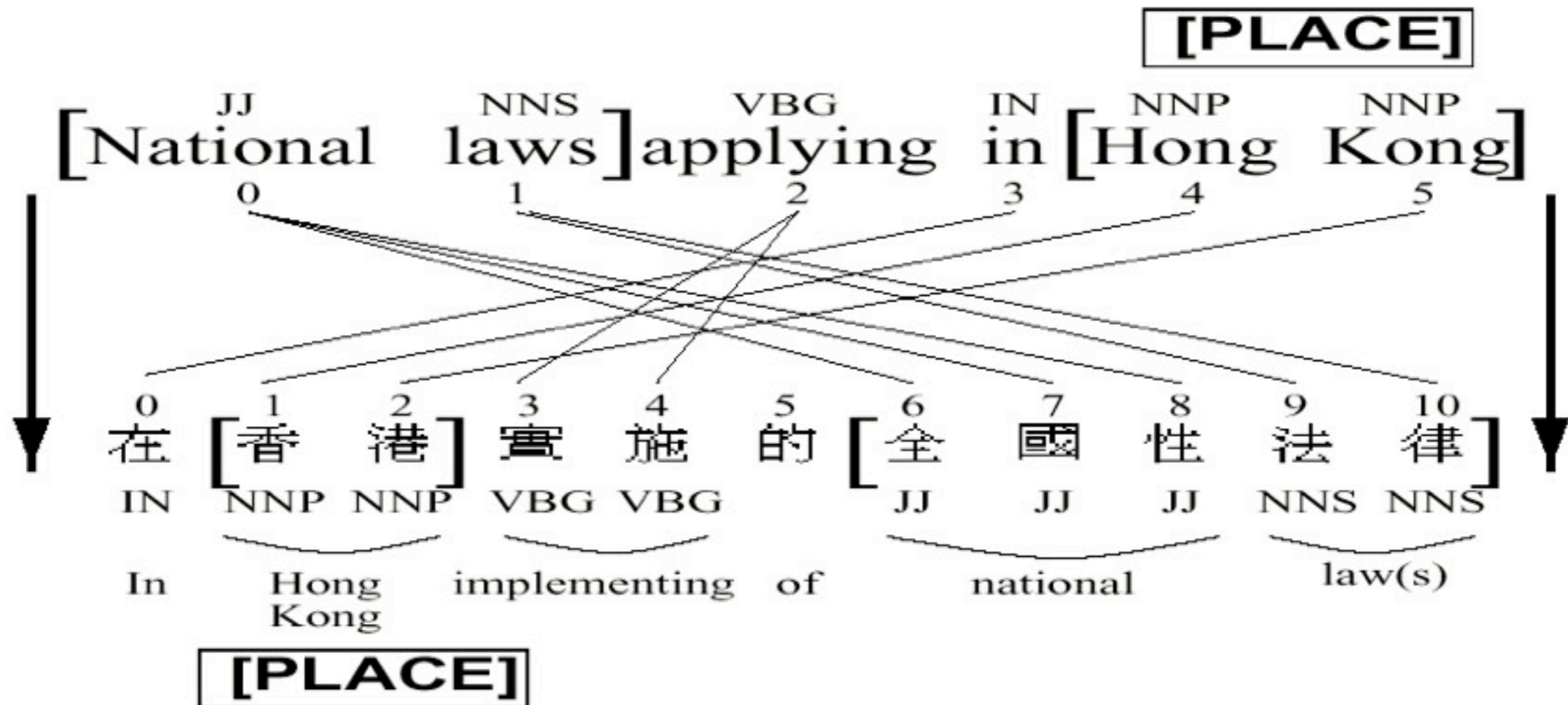
actor role television actress



ottoman empire khan byzantine

Projecting Hidden Structure

Annotations From Existing English Tools



Induced Annotations for Chinese

NLP Tasks

- Analog to digital
 - OCR, Speech Recognition
- Individual language modules
 - Morphology, Syntax, Semantics, and Discourse
- Language to data
 - Information extraction and retrieval, translation

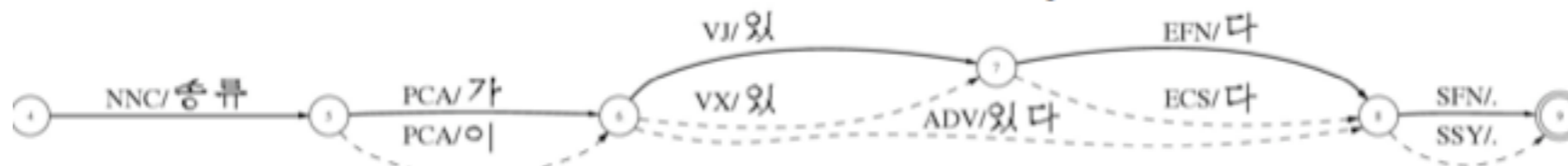
Monolingual & Multilingual

- Analysis technologies for languages
 - Morphology, syntax, semantics
- Translation technologies
 - Dictionaries, cross-lingual IR, MT
- Multilingual exploratory data analysis
 - Clustering, classification → model building

Problems

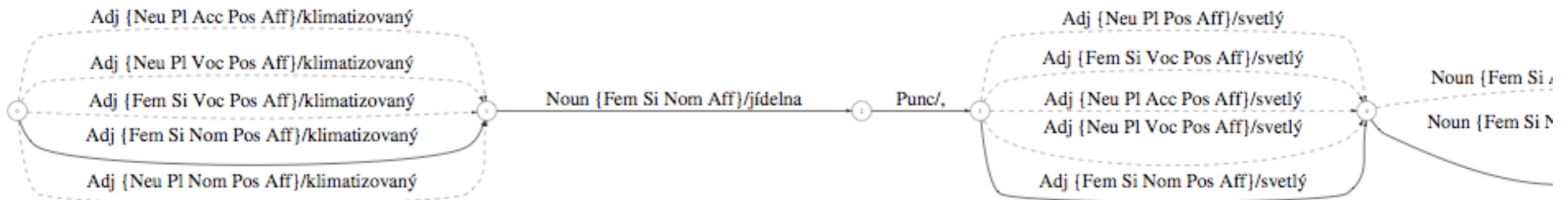
Morphological Ambiguity

There are many kinds of trench mortars.

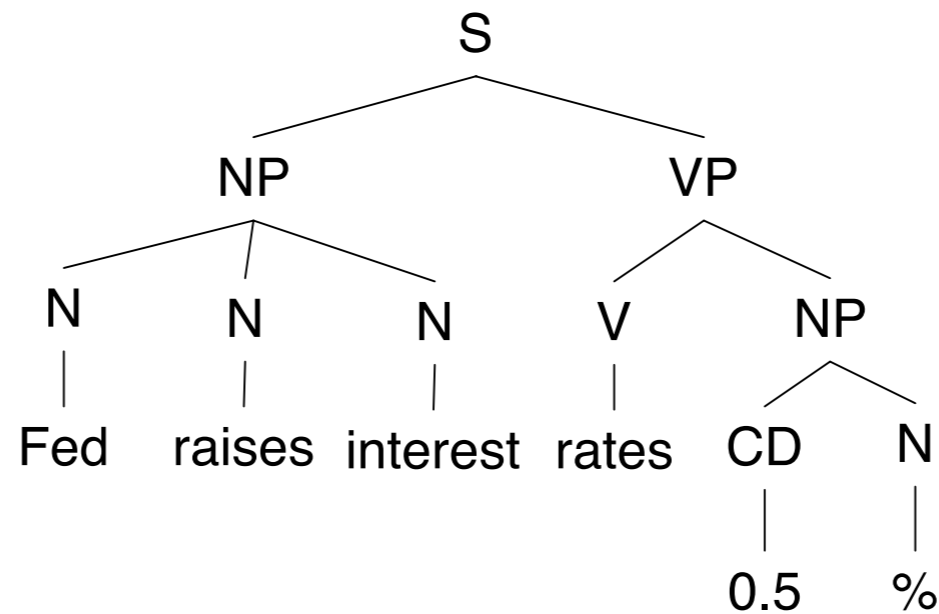
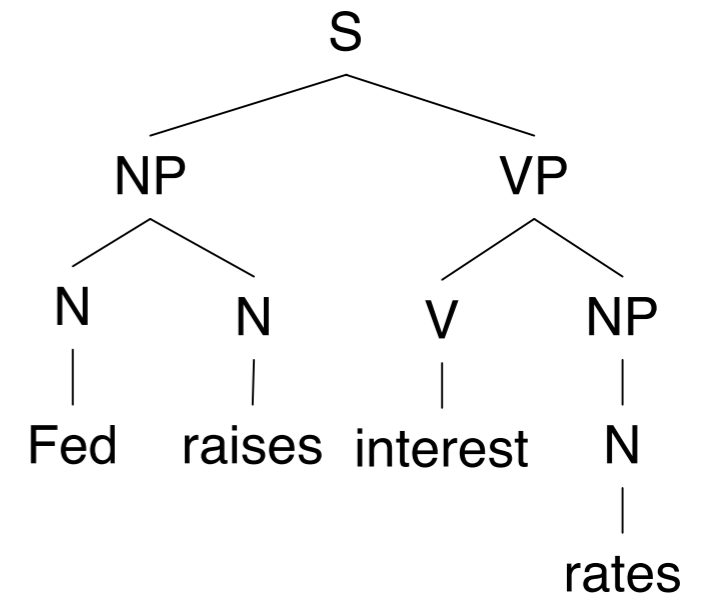
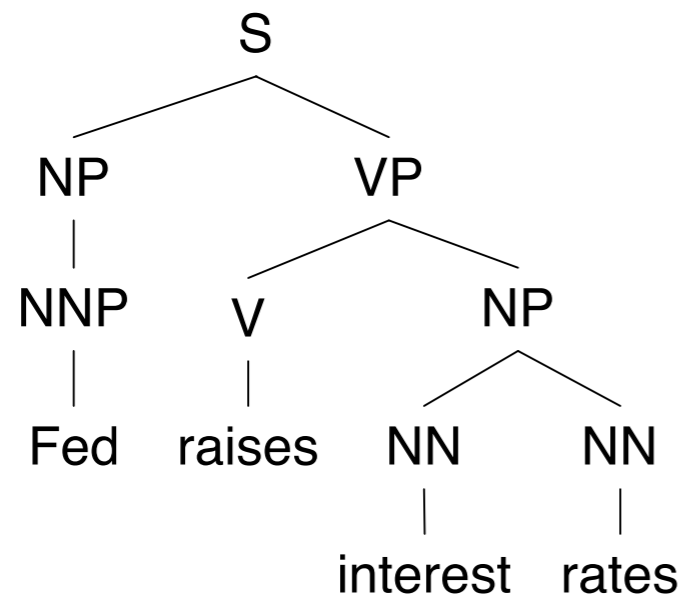


c. Klimatizovaná jídelna, světlá místnost pro snídani.

Air-conditioned dining room,



Syntactic Ambiguity



More Ambiguity

- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Ban on Nude Dancing on Governor's Desk

Why is NLP Hard?

- The rules are ambiguous
- We don't know the rules
- We need to combine lots of weak evidence
- It's *AI complete*
- Language is nearly co-extensive with humanity
- To the rescue: probability, machine learning

Why is NLP in CS?

- How about...
 - Linguistics
 - Statistics
 - Psychology and Cognitive Science
 - The Lang/Lit Humanities
- All of the above!
 - Focus on algorithms, data analysis, engineering

What You'll Learn in NLP

- Looking at data
 - Phenomena and problems
- Modeling data
 - Linguistic and statistical tools
- Algorithms and implementation
 - Efficient computation, practical systems

No Really, What'll I Learn?

- Models of language
 - n-grams, grammars, generative, discriminative
- Algorithms to tame complexity
 - Finite-state models and regular expressions
 - Context-free grammars and parsers
- Problem solving: language ID, spam filters, translation

Who – Where – When

- Professor: David Smith
 - dasmith@cs.umass.edu
 - CS 358, Th after class till 6, W 1:30-3:30
- TA: Jason Naradowsky
 - narad@cs.umass.edu
 - CS 266 after class till 6:15, W 1-2
- Time: T/Th 4:00-5:15
 - www.cs.umass.edu/~dasmith/in1p2009
 - Five medium programming homeworks, midterm, final

Thanks