

From  
WSD  $\rightarrow$  IR  
to  
WSD  $\leftarrow$  IR

Julio Gonzalo  
UNED

# WSD → IR @ UNED: initial motivation

- 1997: EuroWordNet: let's use it!
- 1998: (manual) indexing with synsets +29%
- 1999: Sanderson pseudo-senses vs. WordNet synsets (EMNLP)
- 1999: WSD versus first sense heuristic (SIGLEX)
- 2000: ITEM conceptual search engine

*WSD strategy*

*Conceptual versus textual indexing*



# Bilingual Search Engine

Query: Celebraciones del milenio en la capital de Italia

Processing finished.

Language of the query	Word Sense Disambiguation	Document Representation	Newspaper
◆ Español	◆ All Senses	◆ Textual	◆ The Washington Post
◆ English	◆ First Sense	◆ Conceptual	◆ El País
◆ Catalá	◆ Conceptual Density		◆ El Periódico

```
#sum( celebration millennium #sum(#phrase(capital of italy) #phrase(italian capital) roma rome ) )
```

Finished. Type a new query.

[Peace Can Be Won Through Respect for Human Rights, Pope Says](#)

- WORLD IN BRIEF
- WORLD
- WORLD In Brief
- Israelis, Palestinians Swap Stereotypes
- Britain Honors M.L. King
- In Poland, a Vast Church to Span the Millennium
- Bard Upstages Contenders For 'Man of the Year'

50 documents found. Double click item.

100%

## Peace Can Be Won Through Respect for Human Rights, Pope Says

Saturday, January 2, 1999; Page A20

VATICAN CITY, Jan. 1—In his last New Year's Day message of the millennium, Pope John Paul II took a retrospective look at the closing century to recall the destruction caused by two world wars and to remind his listeners that only through respect for human rights can peace be achieved.

Before thousands in a Mass in St. Peter's Basilica, the pope said in his homily that the

# ITEM search engine

- Scalable to several languages
- Conceptual query expansion
- Translations via hyperonym relations (e.g governor's race)

but

- Granularity
- Indexing units versus translation units
  - Words are not good for translation
    - » (té cargado/strong tea)
  - Phrases are not good for indexing
    - » “word+sense+disambiguation”/“sense tagging”

→Is *Word* Sense Disambiguation an issue for the semantic web?

Expresa en español o inglés el objeto de su búsqueda:

QUERY

### Términos más relevantes en UNED.es

- ▶ Pinche en [el término](#), para mostrar la lista de documentos que contienen el término seleccionado.
- ▶ Pinche en [Re-consult](#), para reconsultar a Google utilizando como consulta el término seleccionado.

EXPLORE PHRASE

### Primeras 10 páginas de Google en UNED.es

EXPLORE DOCUMENT [as, Resources](#)

[www...com](#) Agency Selection, media Relations, Press Release

#### [Programación II](#)

... utilizado en la asignatura homónima, sirve aquí como referencia para los aspectos del lenguaje Modula-2, utilizado para **prácticas** en **Programación II**.  
[sensei.ieec.uned.es/p2/libros.html](http://sensei.ieec.uned.es/p2/libros.html) 7k

#### [Programación II](#)

... Habrá 2 sesiones de **prácticas**, de presencia obligatoria en los Centros Asociados). ... Problemas de **Programación II**; Guía didáctica de **Programación II**. 1996; ...  
[teelab4 iti.uned.es/Primero/Programacion2.htm](http://teelab4 iti.uned.es/Primero/Programacion2.htm) 8k

#### [Asignatura de Programación II \(98/99\)](#)

... La asignatura de **Programación II** se encuadra en el segundo cuatrimestre ... telemático para la realización de las **prácticas** de la asignatura | tarea para la ...  
[sensei.ieec.uned.es/~miguel/tesis/node48.html](http://sensei.ieec.uned.es/~miguel/tesis/node48.html) 38k

#### [Horario del Curso de Acceso](#)

#### práctica de programación ii

- [Re-consult](#) [práctica de programación ii](#) 4 Docs.
- [Re-consult](#) [incluido la práctica de programación ii](#) 1 Docs.
- [Re-consult](#) [prácticas en programación ii](#) 1 Docs.
- [Re-consult](#) [práctica de programación ii curso 1998](#) 1 Docs.
- [Re-consult](#) [práctica de la asignatura de programación ii](#) 1 Docs.
- [Re-consult](#) [utilizado para prácticas en programación ii](#) 1 Docs.

#### ejercicios sobre programación ii

- [Re-consult](#) [ejercicios sobre programación ii](#) 1 Docs.
- [Re-consult](#) [y ejercicios sobre programación ii](#) 1 Docs.

#### práctica de pr

- [Re-co](#)
- [Re-co](#)
- [Re-consult](#) [prevista la realización de una práctica de programación en el lenguaje de la shell](#) 2 Docs.
- [Re-consult](#) [prácticas de programación sas](#) 2 Docs.
- [Re-consult](#) [prácticas programación orientada ia](#) 2 Docs.

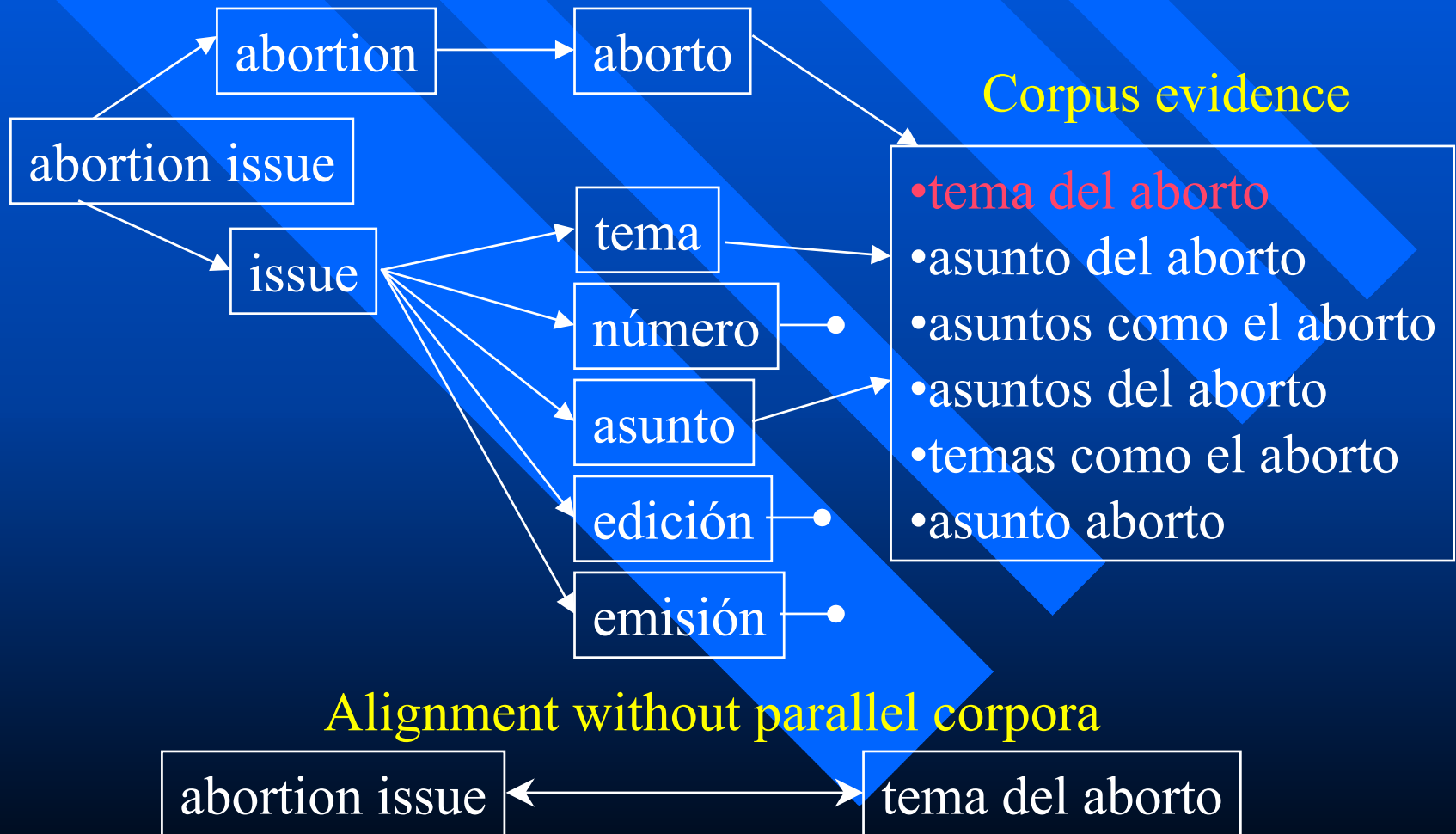
RECONSULT WITH PHRASE

# WTB Evaluation

- 1523 sessions with interaction
- average 5.11 actions per session
- **explore phrase** used in 65.13% sessions

		All queries	1 word queries	>1 word queries
First action	DOC	40.70%	45.49%	37.30%
after QUERY	PHRASE	51.14%	45.65%	55.05%
	RECONSULT	8.141%	8.846%	7.640%
Last action				
Before ending	QUERY	48.74%	53.38%	45.15%
Session with	PHRASE	42.95%	40.85%	44.57%
explore DOC	RECONSULT	8.306%	5.764%	10.27%

# Is WSD easier than MT/CLIR?





# Results on CLEF comparable corpus

Spanish

Size	# Phrases	# Aligned
2	6,577,763	2,004,760
3	7,623,168	252,795

## English

Size	# Phrases	# Aligned
2	3,830,663	1,456,140
3	3,058,698	198,956



# Results on CLEF corpus

2 lemmas

Algorithm

Random Selection

	EN	ES	EN	ES
+ frequent	<b>.83</b>	<b>.80</b>	.02	.02
- frequent	<b>.66</b>	<b>.54</b>	.02	.02

3 lemmas

Algorithm

Random Selection

	EN	ES	EN	ES
+ frequent	<b>.94</b>	<b>.80</b>	.004	.005
- frequent	<b>.81</b>	<b>.62</b>	.004	.004

# Noun Phrase translation

- 1) Select aligned sub-phrase with most frequent translation
- 2) discard overlapping sub-phrases
- 3) iterate.

advances in treatment of a wide variety of diseases

advances in treatment

treatment of a wide

wide variety

variety of diseases

advances in treatment of a wide variety of diseases

advances in treatment

treatment of a wide

wide variety

variety of diseases

tipo de enfermedades

advances in treatment of a wide variety of diseases

advances in treatment

treatment of a wide

wide variety (amplio)

variety of diseases

tipo de enfermedades

advances in treatment of a wide variety of diseases

advances in treatment

treatment of a wide

wide variety (amplio)

variety of diseases

avances en el tratamiento

tipo de enfermedades

advances in treatment of a wide variety of diseases

advances in treatment

treatment of a wide (amplio)

wide variety (amplio)

variety of diseases

avances en el tratamiento

tipo de enfermedades

advances in treatment of a wide variety of diseases

advances in treatment

treatment of a wide

wide variety

variety of diseases

avances en el tratamiento

amplio

tipo de enfermedades



# Is this document relevant?

Query bridge and tunnel construction for the Beijing-Kowloon Railroad

Title coal eastward

Date 99/07/07

西煤东运大通道神华工程生命线 朔黄铁路优质高效加快建设 神肃段明年10月1日前率先开通运煤 990707 西煤东运大通道神华工程生命线 朔黄铁路优质高效加快建设 神肃段明年10月1日前率先开通运煤 本报讯记者严冰报道：我国西煤东运的第二大通道、举世瞩目的跨世纪工程神华工程的生命线——朔黄铁路正在广大建设者手中，优质、高效地建设成为具有我国九十年代末科技水平的现代化铁路。朔黄铁路西起神朔线的神池南站，东至黄骅港的港前站，全长588公里，总概算投资193亿元，由神华集团、铁道部、河北省、山西省共同投资建设，除资本金50.3亿元外，主要资金来源于国家开发银行和日本海外协力基金贷款。目前已开工建设的神（神池南）肃（肃宁北）段正线全长419.6公里，为双线电气化铁路，国家一级干线，概算投资146.2亿元。线路横穿恒山、云中山脉和忻定盆地，沿滹沱河谷穿越太行山进入华北平原。沿线地质情况极为复杂，西段山高谷深，有在我国双线长大隧道中居第2位的长梁山隧道和位居第4和第5位的寺铺尖隧道、水泉湾隧道，线路15次跨越滹沱河；有墩高达61.5米的庄里、红山崖、滴流磴3座悬灌梁大桥，主跨最大达80米。东段河网交错，公路纵横，人口稠密，跨越运输繁忙的京广、京九铁路干线和京深高速公路、107国道等公路干线，1公里以上的特大桥就有5座，跨京深高速公路特大桥（667米）主跨以64米下承钢桁梁跨越公路，为我国铁路之仅有。朔黄铁路的工期目标是2000年5月31日神肃段铺通，10月1日与京九接通运煤，2002年与黄骅港同期配套建成。目前，神肃段已完成路基土石方5000多万立方米，占设计总量的90%多，桥梁工程完成3万多双成桥米，占设计总量的90%多，隧道完成5万多成洞米，占设计总量的70%多，设计的77座隧道已贯通60座；正线铺轨完成200多公里，占设计总长的30%多，站线铺轨完成40多公里，占设计总长的20%多，架梁完成500多孔，占设计总量的20%多。朔黄铁路已开工的各项主体工程均由中标单位承担，不准分包，从制度上保证了工期，质量始终处于可控状态。由于施工单位素质高，能力强，施工管理严格，在朔黄铁路建设中创造了不少奇迹。全线工程质量良好，已完成的工程质量合格率100%，优良率93%。据了解，朔黄铁路建设者提出的质量目标是，路基工程要赶超重载铁路大秦线，桥、隧等结构工程要赶超京九、南昆线，站后工程要面向21世纪。《海外版》（1999年07月07日第2版）人民日报社版权所有，未经授权禁止复制或建立镜像。

info@peopledaily.com.cn 电话：(010)65092993 (010)65091079

Source: Oard 2000

Mozilla: Consulta número 2 con el sistema de 1

## Título de la consulta 2

### Conferencia sobre control de la natalidad

## Descripción de la consulta 2

**¿Cuáles fueron las discusiones y las resoluciones de la Conferencia Mundial de Población sobre el control de la natalidad en El Cairo?**

## Explicaciones complementarias

**Todas las discusiones políticas, propuestas y resoluciones sobre el control de la natalidad en la Conferencia Mundial de Población son de interés. Las posturas de distintos países, organizaciones y grupos son especialmente relevantes.**

### Evalúe el Documento

- No Juzgar Aún
- Relevante
- Algo Relevante
- No Relevante
- ? No hay Inform. suficiente

No obstante, me digo que, decisión de Gianna Beretta podría ser interpretado como heroico. Pero el segundo beatification está disturbando inequívoco. Elisabetta Canori Mora, un romano quién muerta en 1825, seguido siendo en una unión donde su marido abusado le y finalmente abandonado le a cuidar solamente para sus niños. " Elisabetta Canori Mora, " el papa dijo, " demostrado, en el medio de numeroso dificultades conyugales, su fidelidad total a la comisión asumida en sacramento de la unión y de las responsabilidades que derivan de ella." Para permanecer en a unión lousy y destructiva, ella está en el camino a hacer un santo. Ahora hay un buen modelo del papel para las muchachas adolescentes católicas. Considere, también, contexto para la selección de estas dos mujeres: El Vaticano ha estado intentando a influencie la agenda para la conferencia de U.N. de esta caída en El Cairo en la población y desarrollo. El delegado de Vaticano a una reunión preparatoria reciente buscó a suprima la noción de la " maternidad segura " como meta importante de la salud. En a humor refrenado, chastised para pensar eso si usted no está para la caja fuerte maternidad, usted está para la maternidad insegura. Pero estos dos beatifications refuerza el mensaje que el Vaticano no valora la seguridad de las mujeres adentro maternidad y vida de familia.

Mozilla: Consulta número 2 con el sistema de 2

## Título de la consulta 2

Conferencia sobre control de la natalidad

## Descripción de la consulta 2

¿Cuáles fueron las discusiones y las resoluciones de la Conferencia Mundial de Población sobre el control de la natalidad en El Cairo?

## Explicaciones complementarias

Todas las discusiones políticas, propuestas y resoluciones sobre el control de la natalidad en la Conferencia Mundial de Población son de interés. Las posturas de distintos países, organizaciones y grupos son especialmente relevantes.

**Evalúe el Documento**

- No Juzgar Aún
- Relevante
- Algo Relevante
- No Relevante
- ? No hay Inform. suficiente

gianna beretta  
segundo beatification  
elisabetta canori mora  
midst numeroso conyugal dificultad  
fidelidad total compromiso  
sacramento del matrimonio responsibility  
piojoso destructor matrimonio  
buen papel modelo  
jóvenes de edades  
contexto selección dos mujer  
**congresos de el cairo población desarrollo**  
reciente preparatorio  
maternidad segura  
salud importante meta  
inseguro maternidad

Best alignment

Phrase with query terms

# Results

System	Precision	Recall
Systran MT	0.48	0.22
UNED NPs	0.47 (-2%)	0.34 (+52%)

cf. U. Maryland experiment: word-by-word translation substantially worse than Systran.

# CLIR Query formulation

## Reference system

- Assisted word-by-word translation.

## UNED system

- Assisted formulation by phrases.
- Automatic translation using alignment.

# UNED query formulation

Mozilla: Consulta número 4 con el sistema 2

## Seleccione las expresiones que desea incluir en la consulta

**La consulta contiene 5 expresiones**

<input type="checkbox"/> atención al paciente	<input type="checkbox"/> consecuencia de la huelga
<input type="checkbox"/> caso de ataque	<input type="checkbox"/> motivo de la huelga
<input type="checkbox"/> huelga de hambre	

**Si lo desea puede añadir nuevos términos individuales a la consulta**

### Leyenda de Símbolos

- No Juzgado
- Relevante
- No Relevante
- ? Dudoso

1	<input checked="" type="radio"/>	<a href="#">NOTICIAS DE LA COMUNIDAD: CENTRAL: PARQUE DEL ECO; PLAN EXCESIVO EFECTUADO DEL ALZA DEL AUTOBÚS DE LA HUELGA DE HAMBRE.</a>
2	<input checked="" type="radio"/>	<a href="#">MUJER PARA TERMINAR HUELGA DE HAMBRE DE GUATEMALA; DERECHOS HUMANOS: EL ABOGADO DE ESTADOS UNIDOS BUSCARÁ CARGAS DEL CRIMINAL CONTRA OFICIALES DEL EJÉRCITO. ELLA SE HABÍA MUERTO DE HAMBRE POR 32 DÍAS ADENTRO ESFUERZO DE APRENDER DEL SINO DE SU MARIDO.</a>
3	<input checked="" type="radio"/>	<a href="#">EL ACUERDO TERMINA HUELGA DE HAMBRE DE SANTA DE LOS ESTUDIANTES UC DE BARBARA.</a>
4	<input checked="" type="radio"/>	<a href="#">UCLA ESCOGE LA CABEZA DEL INTERINO PARA LOS ESTUDIOS DE CHICANO; EDUCACIÓN: ALGUNOS ACTIVISTAS CRITIQUE LA CITA DEL WHO DEL ADMINISTRADOR AYUDADA PARA NEGOCIAR EL EXTREMO PARA TENER HAMBRE HUELGA.</a>
5	<input checked="" type="radio"/>	<a href="#">ABOGADOS EGIPCIOS EN LOS COLEGAS EXCESIVOS DE LA HUELGA DE HAMBRE QUE ENCARCELAN</a>



# UNED relevance feedback

Mozilla: Consulta número 4 con el sistema 2

## Seleccione las expresiones que desea incluir en la consulta

**La consulta contiene 5 expresiones**

<input type="checkbox"/> atención al paciente	<input type="checkbox"/> consecuencia de la huelga
<input type="checkbox"/> caso de ataque	<input type="checkbox"/> motivo de la huelga
<input checked="" type="checkbox"/> huelga de hambre	

**Si lo desea puede añadir nuevos términos individuales a la consulta**

### Evalúe el Documento

Relevante  
 No Relevante  
 ? Dudoso

**huelga de hambre en guatemala** ←

**derechos humanos**

**criminal acusaciones contra oficiales del ejército**

**jornada de trabajo**

---

**abogados estadounidenses**



# UNED relevance feedback

Mozilla: Consulta número 4 con el sistema 2

## Seleccione las expresiones que desea incluir en la consulta

**La consulta contiene 6 expresiones**

<input type="checkbox"/> atención al paciente	<input type="checkbox"/> consecuencia de la huelga
<input type="checkbox"/> caso de ataque	<input type="checkbox"/> motivo de la huelga
<input type="checkbox"/> huelga de hambre	<input type="checkbox"/> huelga de hambre en guatemala

**Si lo desea puede añadir nuevos términos individuales a la consulta**

### Leyenda de Símbolos

- No Juzgado
- Relevante
- No Relevante
- ? Dudoso

- [MUJER PARA TERMINAR HUELGA DE HAMBRE DE GUATEMALA; DERECHOS HUMANOS: EL ABOGADO DE ESTADOS UNIDOS BUSCARÁ CARGAS DEL CRIMINAL CONTRA OFICIALES DEL EJÉRCITO. ELLA SE HABÍA MUERTO DE HAMBRE POR 32 DÍAS ADENTRO ESFUERZO DE APRENDER DEL SINO DE SU MARIDO.](#)
- [ABOGADO DE ESTADOS UNIDOS EN LA HUELGA DE HAMBRE PARA EL ESPOSO PERDIDO DEL GUERRILLA; GUATEMALA: JENNIFER HARBURY DICE QUE ELLA SE MORIRÁ DE HAMBRE A LA MUERTE A MENOS QUE EL EJÉRCITO POSEA HASTA LO QUE ELLA VE COMO UN EXPEDIENTE 2 1/2-YEAR DEL ENGAN@O. ELLA INSISTE QUE ÉL ES UN PRISIONERO DE GUERRA Y SE HA TORTURADO SIN PIEDAD.](#)
- [NOTICIAS DE LA COMUNIDAD: CENTRAL; PARQUE DEL ECO; PLAN EXCESIVO EFECTUADO DEL ALZA DEL AUTOBÚS DE LA HUELGA DE HAMBRE](#)
- [ALBANIA FACILITA EL SOPORTE EN CARACTERÍSTICA CONFISCADA](#)
- [EL ACUERDO TERMINA HUELGA DE HAMBRE DE SANTA DE LOS](#)

# Results

<b>System</b>	<b><math>F_{\alpha=0.8}</math></b>
<b>Reference</b>	<b>.23</b>
<b>UNED</b>	<b>.37 (+65%)</b>

Statistical significance:

$p < 0.05$  linear mixed-effects model + ANOVA.

# Initial Query formulation

System	Average time
Reference	286 s.
UNED	44 s.

# Initial query formulation

System	P @ 20
Reference	.19
UNED	.29

# And what about WSD?

- Supervised systems have little to be supervised with...
  - Research on unsupervised systems (Senseval 2)
  - Solve the acquisition bottleneck of supervised systems: obtain training instances automatically.
- Better understanding of the problem: sense inventories, test suites, polysemy.

# WSD is harder than the applications

- IR → WSD: automatic assignment of web directories to word senses (Computational Linguistics, to appear)
- MT → WSD: use aligned phrases for partial disambiguation (no need for parallel corpora!) (work in progress)
- WSD: go to the basics: study sense inventories, and polysemy distinctions for clustering (SIGLEX 00, 02)




Google Search

[Directory Help](#)

Search only in Tigers  Search the Web

## Tigers

[Society](#) > [Issues](#) > [Environment](#) > [Conservation and Endangered Species](#) > [Species Loss](#) > [Mammals](#) > [Cats](#) > [Tigers](#)

[Go to Directory Home](#)

Related Category:

[Science](#) > [Biology](#) > [Flora and Fauna](#) > [Animalia](#) > [Chordata](#) > [Mammalia](#) > [Carnivora](#) > [Felidae](#) > [Panthera](#) > [Tiger](#) (33)

### Web Pages

Viewing in Google PageRank order

View in alphabetical order

- [Tiger Information Center](http://www.5tigers.org/) - <http://www.5tigers.org/>  
 Dedicated to providing information to help preserve the remaining five subspecies of tigers.
- [Tiger Foundation](http://www.tigerfdn.com/) - <http://www.tigerfdn.com/>  
 A Canadian conservation organization exclusively dedicated to the preservation of wild tigers.
- [Save China's Tigers](http://www.savechinastigers.org/) - <http://www.savechinastigers.org/>  
 Dedicated to in-situ conservation of native cat species in China, particularly the South China Tiger.
- [Tigers in Crisis](http://www.tigersincrisis.com/) - <http://www.tigersincrisis.com/>  
 Information about endangered tigers, conservation status and solutions.
- [Tiger Missing Link Foundation](http://www.tigerlink.org/) - <http://www.tigerlink.org/>  
 "Saving tigers one by one" -- The only non-government organization working in tiger genetics with the private sector tiger owners to preserve tiger biodiversity.
- [TigerAid Foundation](http://www.tigeraid.org/) - <http://www.tigeraid.org/>  
 Through public-private partnerships in alliance with Internet users, our organization is creating the largest global online community acting on protecting tigers and their habitats.
- [Forever Tigers](http://www.forevertigers.com/) - <http://www.forevertigers.com/>  
 Information about threats to wild tiger populations.
- [Awesome Tigers](http://www.everwonder.com/david/tigers/) - <http://www.everwonder.com/david/tigers/>  
 Information about living and extinct tigers.
- [Tiger Havens Of India](http://www.geocities.com/pateluday/tiger1.html) - <http://www.geocities.com/pateluday/tiger1.html>  
 A site by an amateur naturalist dedicated to tigers and their conservation
- [Tigers in the Wild](http://panda.org/resources/publications/species/tiger99/) - <http://panda.org/resources/publications/species/tiger99/>  
 Describes the conservation efforts of the World Wildlife Fund (WWF) to save tigers from extinction due primarily to habitat loss and illegal killing.
- [Endangered Species on Ecobeetle - Tiger](http://www.ecobeetle.com/tiger.htm) - <http://www.ecobeetle.com/tiger.htm>  
 Ecobeetle is your environmental webpage with a special focus on endangered tigers.






Google Search

[Directory Help](#)
 Search only in Woods, Tiger
  Search the Web

## Woods, Tiger

[Sports](#) > [Golf](#) > [Golfers](#) > Woods, Tiger

[Go to Directory Home](#)

### Web Pages

Viewing in Google PageRank order

View in alphabetical order

- [First Church of Tiger Woods](http://www.TigerWoodsisGod.com) - <http://www.TigerWoodsisGod.com>  
Dedicated to exploring the possibility that Tiger Woods may actually be God.
- [Tiger Woods Timeline](http://www.infoplease.com/spot/tigertime1.html) - <http://www.infoplease.com/spot/tigertime1.html>  
A potted life history presented as a year-by-year timeline.
- [AskMen.com - Tiger Woods](http://www.askmen.com/men/sports/28_tiger_woods.html) - [http://www.askmen.com/men/sports/28\\_tiger\\_woods.html](http://www.askmen.com/men/sports/28_tiger_woods.html)  
Commentary on his star qualities, plus a biography and related links.
- [PGA Tour.com - Tiger Woods](http://www.pgatour.com/players/intro/8793.html) - <http://www.pgatour.com/players/intro/8793.html>  
Current statistics or tour history.
- [Rick's Tiger Woods Page](http://world.std.com/~rgu/tigerwoods/index.html) - <http://world.std.com/~rgu/tigerwoods/index.html>  
FAQ, anecdotes and quotes, mailing list, and links.
- [Golf Stars Online: Tiger Woods](http://www.golfstarsonline.com/WW/Tiger_Woods/) - [http://www.golfstarsonline.com/WW/Tiger\\_Woods/](http://www.golfstarsonline.com/WW/Tiger_Woods/)  
Reviewed links to websites, articles, interviews, endorsements, photos and other online resources, plus fan comments.
- [Billion Dollar Athlete](http://www.askmen.com/sports/business/15_sports_business.html) - [http://www.askmen.com/sports/business/15\\_sports\\_business.html](http://www.askmen.com/sports/business/15_sports_business.html)  
Article on Tiger Woods as marketing phenomenon.
- [All Star Tiger Woods](http://www.allstarz.org/tigerwoods/enter.htm) - <http://www.allstarz.org/tigerwoods/enter.htm>  
Includes facts, pictures, wallpaper, statistics, swing analysis, and related links.
- [Tiger's Den](http://www.geocities.com/Colosseum/Track/6331/index.html) - <http://www.geocities.com/Colosseum/Track/6331/index.html>  
Biography, pictures, chat room, screen saver, and audio and video clips.
- [Will's Tiger Woods Page](http://home.swipnet.se/~w-61837/tiger.htm) - <http://home.swipnet.se/~w-61837/tiger.htm>  
Photos, audio, video and personal information.
- [Tiger Woods Site in Thailand](http://members.tripod.com/~t_woods/) - [http://members.tripod.com/~t\\_woods/](http://members.tripod.com/~t_woods/)  
Tiger Woods' biography, with postcards, news, and pictures in Thailand. Photo gallery, updates, articles, facts, chatroom, and links.
- [Tiger Fan-Attics](http://www.tigerfans.com/) - <http://www.tigerfans.com/>  
Tiger Woods fan club with information, links, photos, crossword puzzles, and information about books and other publications.
- [Tiger Watch](http://members.aol.com/Pgatour01/tiger.htm) - <http://members.aol.com/Pgatour01/tiger.htm>  
Track the Tiger throughout his career. Watch his wins, major championships and earnings on and off the course.
- [Tiger's swing](http://www.msnbc.com/news/546343.asp) - <http://www.msnbc.com/news/546343.asp>

# WordNet senses ↔ web dirs

## ■ Circuit 1 (electrical circuit)

business/industries/electronics and electrical/contract manufacturers (.98)

manufacturers/printed circuit boards/fabrication (.88)

computers/cad/electronic design automation... (.78)

### Sense specializations:

business/industries/electronics/components/integrated circuits (.98)

## ■ Circuit 2 (tour, journey around a particular area)

Sports/cycling/travel/travelogues/europe/france (.58)

Regional/asia/nepal/travel and tourism/travel guides (.66)

## ■ Circuit 5 (racing circuit)

Sports/motorsports/auto racing/stock cars/drivers and teams (.78)

Sports/motorsports/auto racing/tracks (.82)

Sports/motorsports/auto racing/driving schools (.78)

# Applications

- Automatic acquisition of training corpora.
- Sense clustering.
- Increase lexical coverage
  - Specialized senses (*integrated circuit*)
  - New senses (*oasis, jaguar, tiger*)

Cleaner than the full web as corpus

More stable: ODP is downloadable!

Less redistribution problems

# Algorithm

1. Retrieve ODP directories for every word sense with wordnet-based queries.
2. Directory: Extract lemmas in every directory full path.
3. Word sense: extract lemmas in related synsets, including hyperonym chain.
4. Compare both representations for cooccurrence.
5. Apply heuristic filters.

# Evaluation: Senseval-2 nouns

# senses	# dirs	# labeled senses	# sense extensions	Coverage	precision
147	148	43	28	.73/.88	.86

Irrelevant	Mildly relevant	Highly relevant
.15	.17	.67

# Processing of wordnet nouns

Candidate nouns	51,168
Candidate senses	73,612
Associations	29,291
Characterized nouns	24,558
Characterized senses	27,383
Sense specializations	1,800

# Automatic acquisition of WSD training corpora

- **Circuit 1 (electrical circuit)**

Electromechanical products for brand name firms; offers printed circuit boards (..) Offers surface mount, thru-hole, and flex circuit assembly, in circuit and functional

- **Circuit 2 (tour, journey around a particular area)**

The Tour du Mont-Blanc is a circuit of 322 km based in the northern French Alps  
A virtual tour of the circuit by Raimon Bach

- **Circuit 5 (racing circuit)**

The circuit is a smooth 536 yards of racing for Hot Rod and Stock Car's at the east  
History of the circuit and its banked track and news of Formula 1



# Results of supervised WSD

Word senses	# train instances Senseval	# train instances Directories	# test instances	Recall Senseval	Recall Directories
Bar 1,10	127,11	1,1	62,6	.91	.50
Child 1,2	39,78	3,80	35,27	..57	.44
Circuit 1,2,5	67,6,7	229,2,5	23,2,8	.70	.70
Facility 1,4	26,61	4,18	15,28	.79	.67
Grip 2,7	6,1	17,6	4,0	1	1
Holiday 1,2	4,57	5,17	26,2	.96	.96
Material 1,4	65,7	63,10	30,9	.79	.79
Post 2,3,4,7,8	1,64,20,11,7	2,7,1,9,3	2,25,13,12,4	.45	.25
Restrain 1,4,6	17,32,11	2,2,2	8,14,4	.65	.50
Stress 1,2	3,45	8,50	1,19	.95	.95
<b>TOTAL</b>	<b>773</b>	<b>547</b>	<b>379</b>	<b>.73</b>	<b>.58</b>

# Comparable training material

Word senses	# train instances SENSEVAL	# train instances DIRS	# test instances	Recall SENSEVAL	Recall DIRS
Bar 1,10	127,11	1,1	62,6	.91	.50
Child 1,2	39,78	3,80	35,27	..57	.44
Circuit 1,2,5	67,6,7	229,2,5	23,2,8	.70	.70
Facility 1,4	26,61	4,18	15,28	.79	.67
Grip 2.7	6,1	17,6	4,0	1	1
Holiday 1,2	4,57	5,17	26,2	.96	.96
Material 1,4	65,7	63,10	30,9	.79	.79
Post 2,3,4,7,8	1,64,20,11,7	2,7,1,9,3	2,25,13,12,4	.45	.25
Restrain 1,4,6	17,32,11	2,2,2	8,14,4	.65	.50
Stress 1,2	3,45	8,50	1,19	.95	.95
<b>TOTAL</b>	<b>773</b>	<b>547</b>	<b>379</b>	<b>.73</b>	<b>.58</b>

# Incorrect directories

Word senses	# train instances Senseval	# train instances Directories	# test instances	Recall Senseval	Recall Directories
Bar 1,10	127,11	1,1	62,6	.91	.50
Child 1,2	39,78	3,80	35,27	..57	.44
Circuit 1,2,5	67,6,7	229,2,5	23,2,8	.70	.70
Facility 1,4	26,61	4,18	15,28	.79	.67
Grip 2.*7	6,1	17,6	4,0	1	1
Holiday 1,*2	4,57	5,17	26,2	.96	.96
Material *1,*4	65,7	63,10	30,9	.79	.79
Post 2,*3,*4,7,*8	1,64,20,11,7	2,7,1,9,3	2,25,13,12,4	.45	.25
Restrain 1,4,6	17,32,11	2,2,2	8,14,4	.65	.50
Stress 1,2	3,45	8,50	1,19	.95	.95
<b>TOTAL</b>	<b>773</b>	<b>547</b>	<b>379</b>	<b>.73</b>	<b>.58</b>

# Characterization of sense inventories for WSD

- Given two senses of a word,
  - How are they related? (polysemy relations)
  - How closely? (sense proximity)
  - In what applications should be distinguished?
- Given an individual sense of a word
  - Should it be split into subsenses? (sense stability)

# Semantic distance in WordNet

- Wordnet conceptual relations (Resnik, Agirre & Rigau, etc.) → topic relatedness?
- Cross-Linguistic evidence: Resnik & Yarowsky 99.

# Cross-Linguistic evidence

## Fine 40129

Mountains on the other side of the valley rose from the mist like islands, and here and there flecks of cloud as pale and `<tag>fine</tag>` as sea-spray, trailed across their sombre, wooded slopes.

**TRANSLATION:** \* \*

# Sense proximity

$$\mathbf{P}_L(\text{same lexicalization} | w_i, w_j) \equiv \frac{1}{|w_i| |w_j|} \sum_{\substack{x \in \{w_i \text{ examples}\} \\ y \in \{w_j \text{ examples}\}}} \text{tr}_L(x) = \text{tr}_L(y)$$

$$\text{Proximity}(w_i, w_j) \equiv \frac{1}{|\text{languages}|} \sum_{L \in \text{languages}} \mathbf{P}_L(\text{same lexicalization} | w_i, w_j)$$

# Sense Stability

$$\text{Stability}(w_i) \equiv \frac{1}{|\text{languages}|} \sum_{L \in \text{languages}} \frac{1}{|\{x, y \in \{w_i \text{ examples}\}, x \neq y\}|} \sum_{\substack{x, y \in \{w_i \text{ examples}\}, \\ x \neq y}} t_L(x) = t_L(y)$$



# Experiment Design

## MAIN SET

44 words (nouns and adjectives)

182 senses

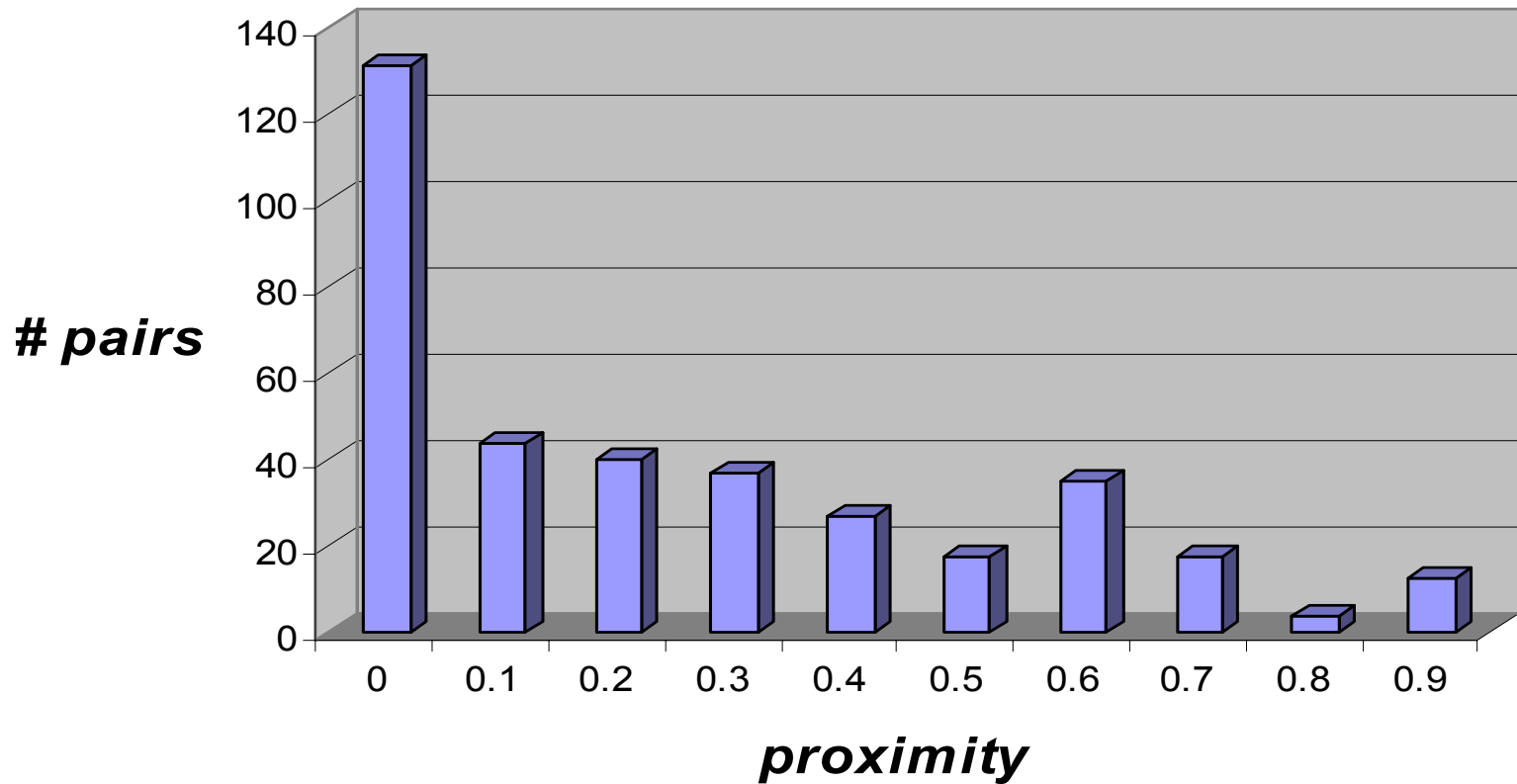
508 examples

11 native/bilingual speakers of 4 languages

*Bulgarian*  
*Russian*  
*Spanish*  
*Urdu*

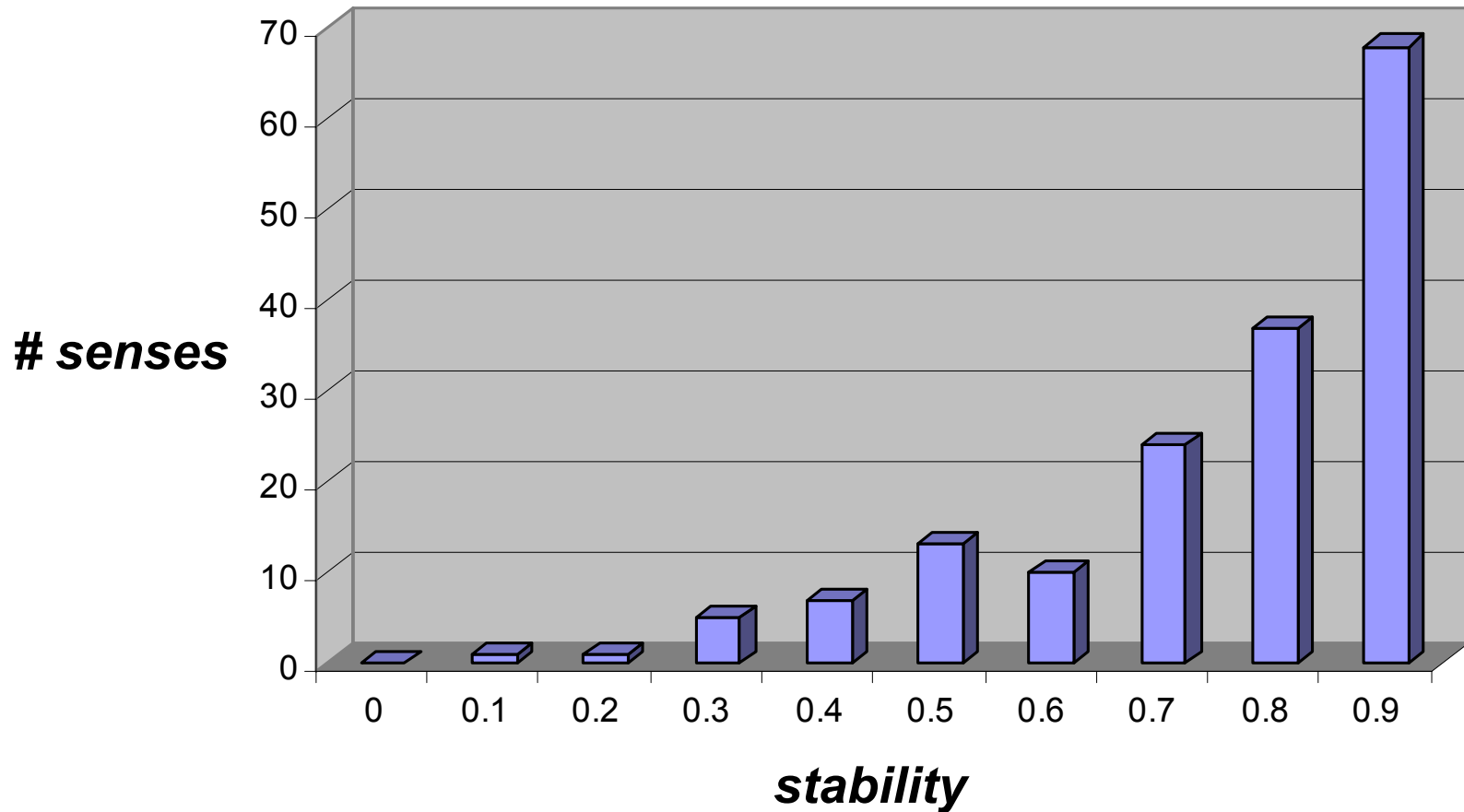
# RESULTS:

## distribution of proximity indexes



Average proximity = 0.29

# Results: distribution of stability indexes



Average stability = 0.80

# Re-scoring Senseval systems

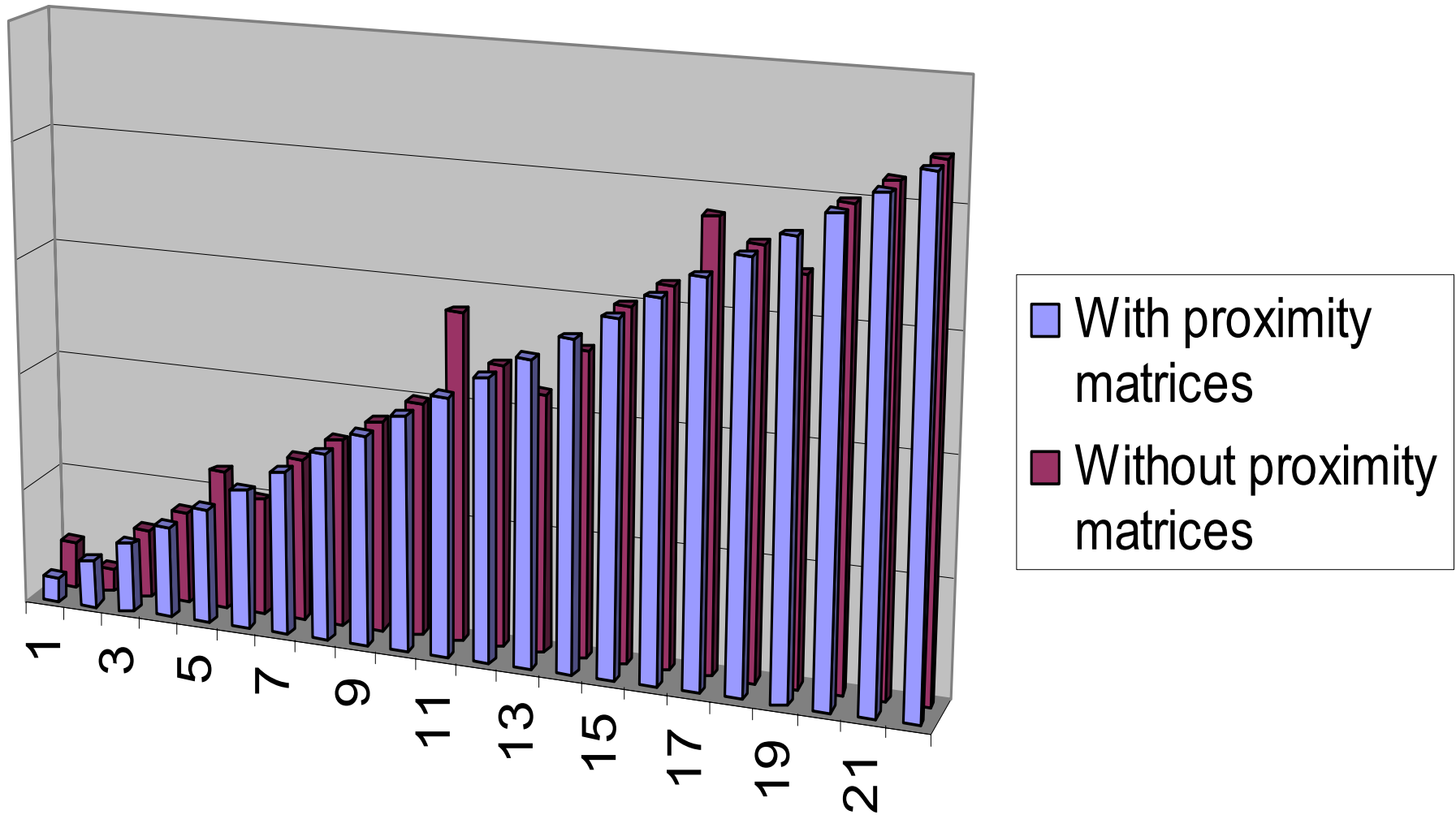
SENSEVAL-2

score =  $P(\text{correct sense})$

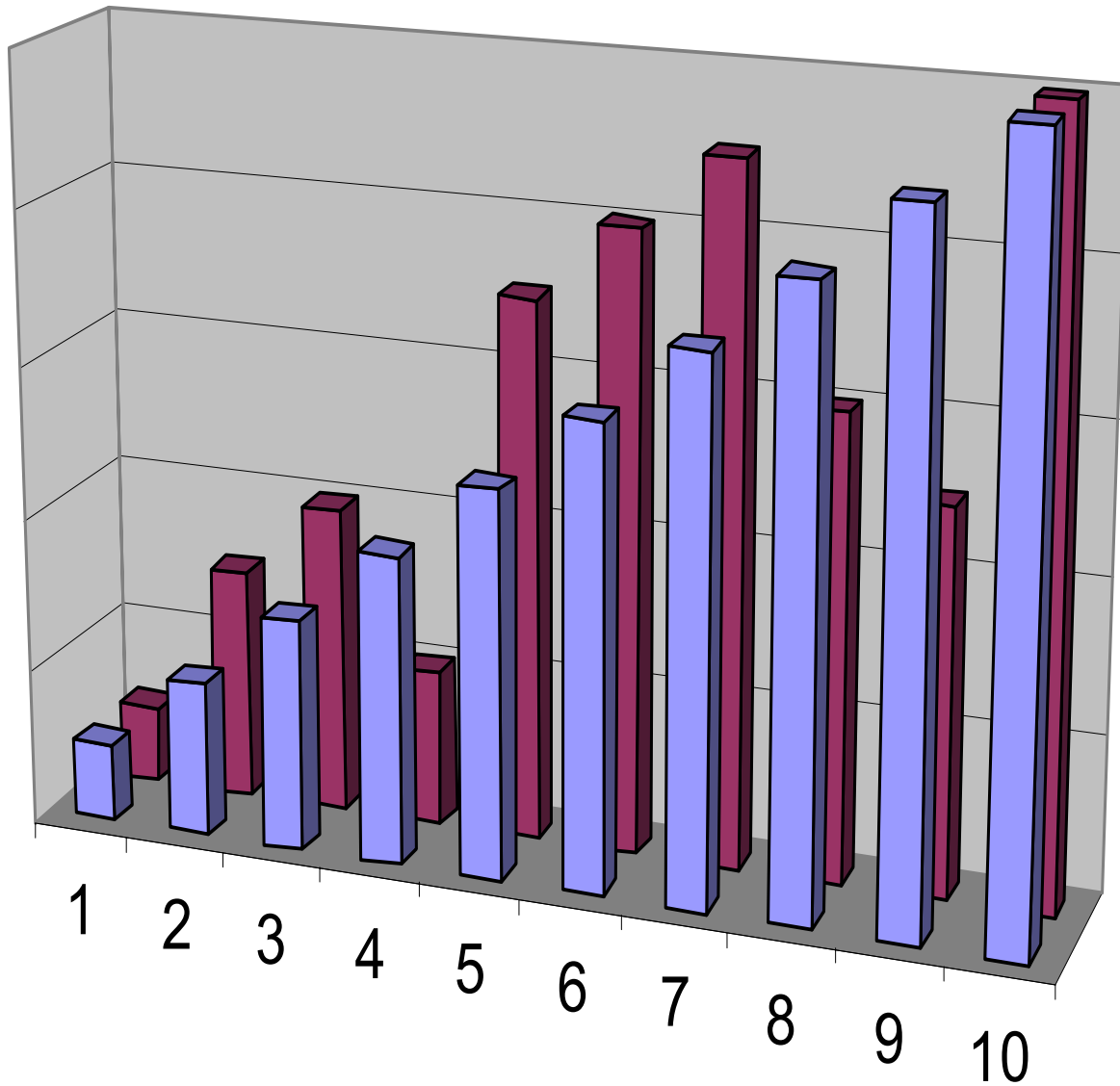
NEW

score =  $\sum \text{proximity}(i, \text{correct}) P(i)$

# SUPERVISED SYSTEMS (25)

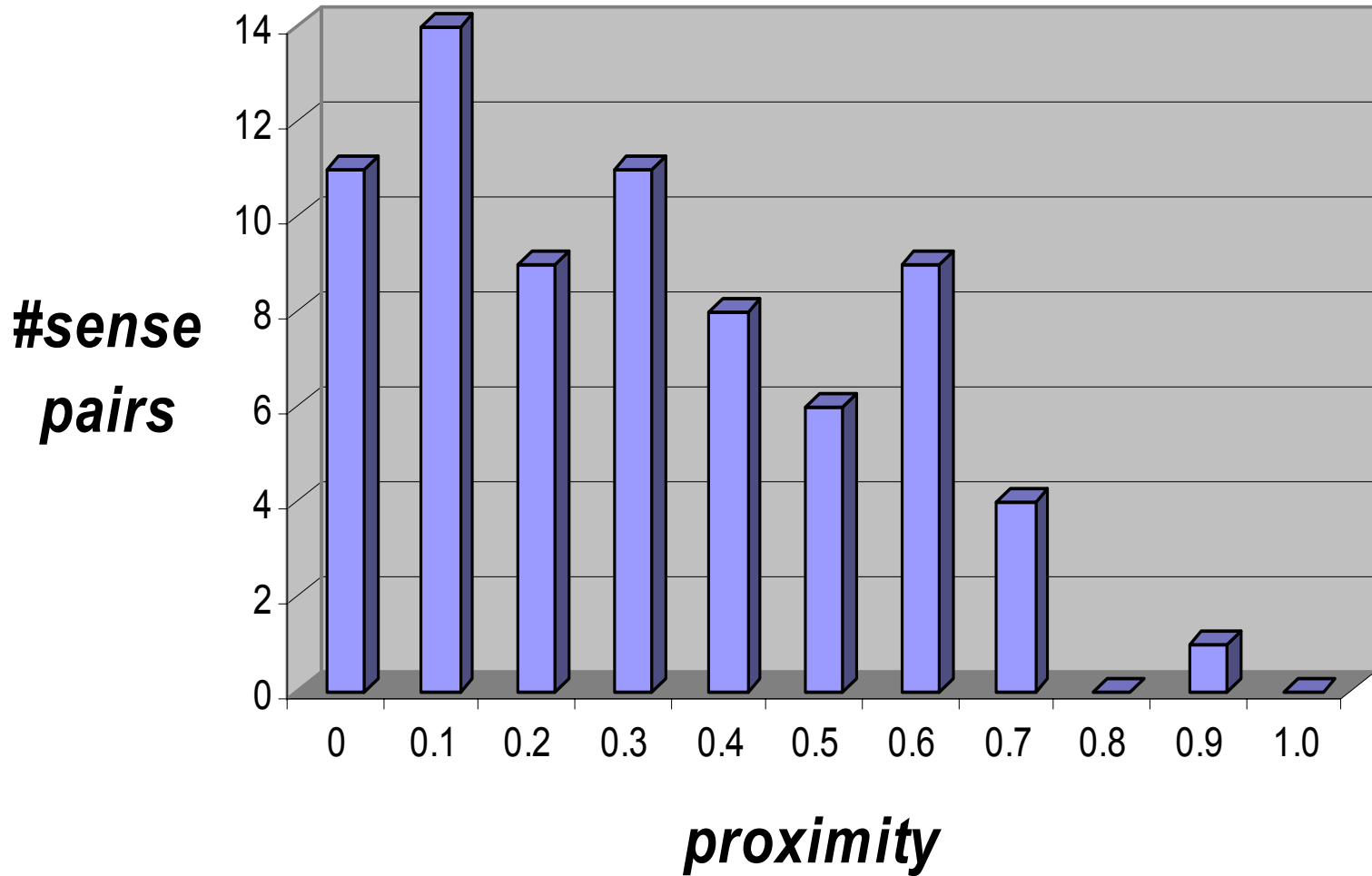


# UNSUPERVISED SYSTEMS (10)



- with proximity matrices
- without proximity matrices

# distribution of metaphors



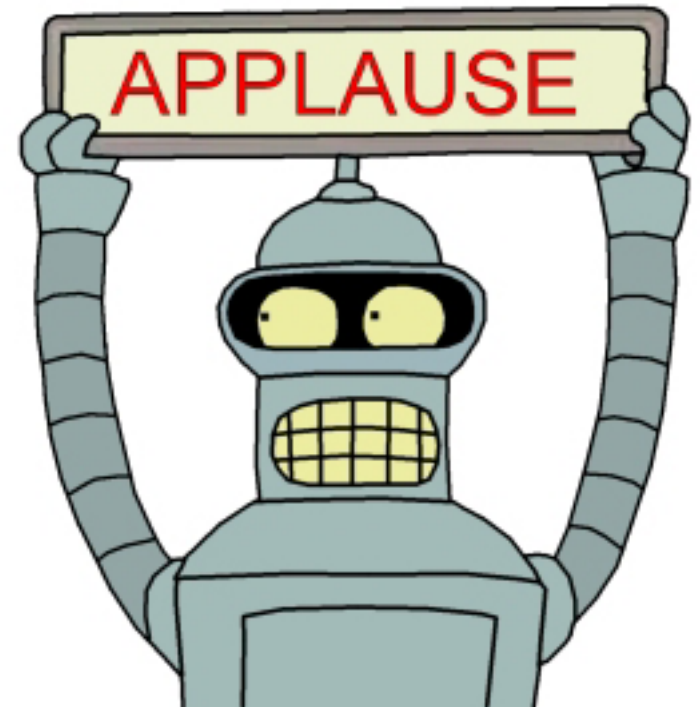
# Suggestions for Meaning

- IR → WSD: Use web directories
- MT → WSD: Use comparable corpora
- Lexical representation: enrich WordNet with polysemy relations
- Evaluation: Focus on interactive applications. Beware of the semantic web!



More info

<http://nlp.uned.es>



# Typology of Polysemic Relations

## **METONYMY**

*(post-letters vs. post-delivery)*

- **METAPHOR**

*(window-house vs. window-computer)*

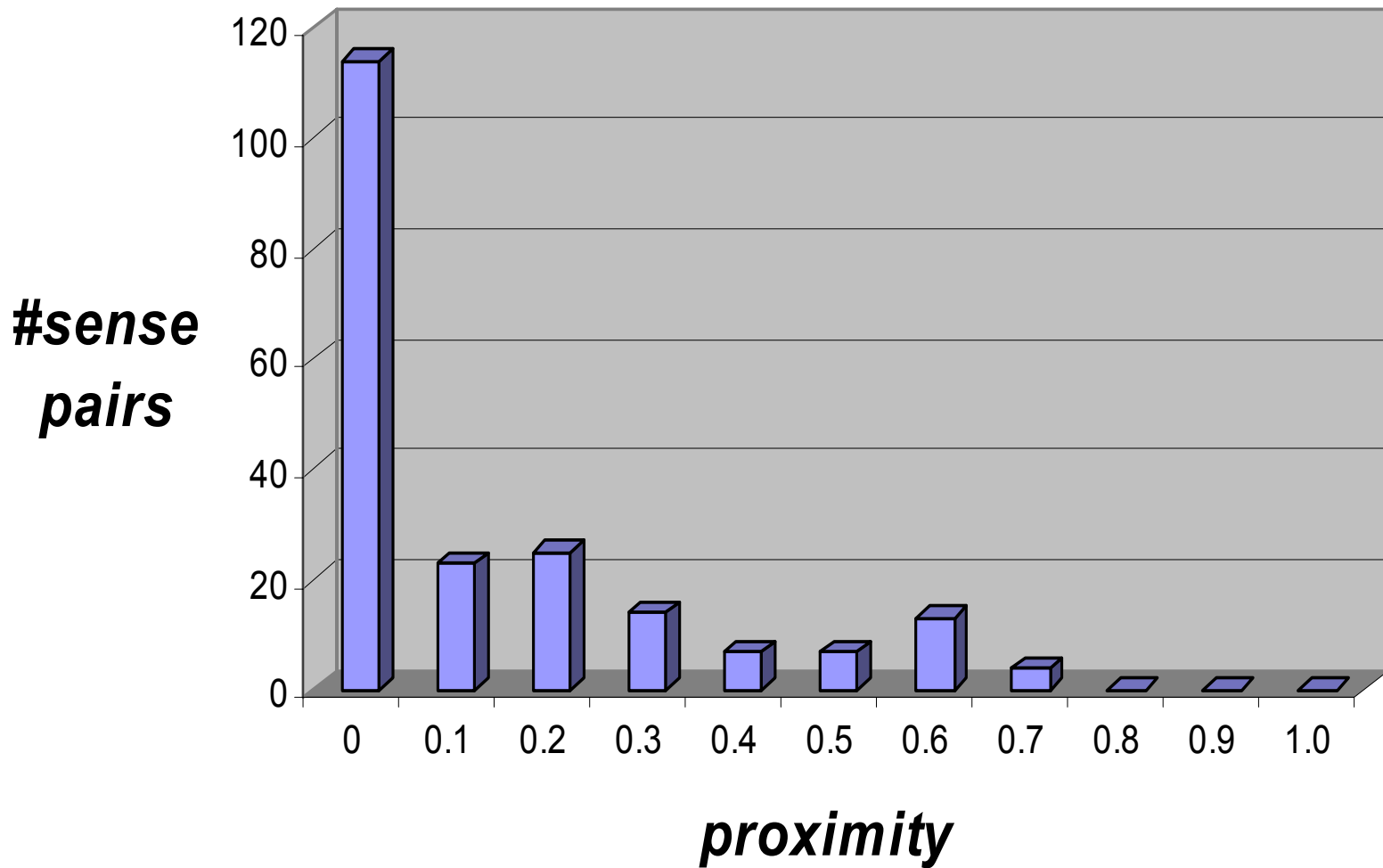
- **SPECIALIZATION / GENERALIZATION**

*(fine-ok vs. fine-greeting)*

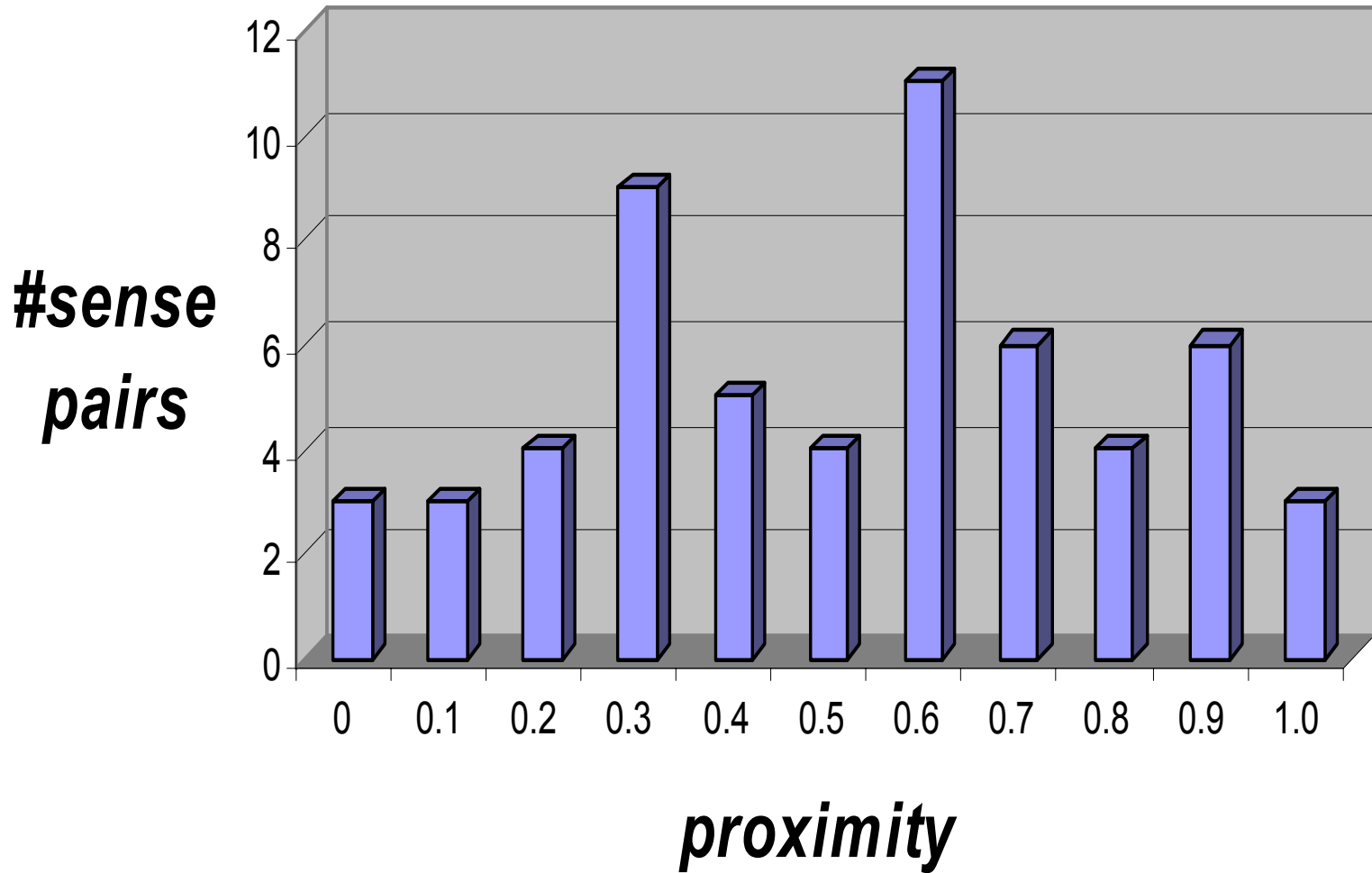
- **HOMONYMY** → no relation

*(bar-law vs. bar-unit\_of\_pressure)*

# distribution of homonyms



# distribution of metonymy



# distribution of specialization/generalization

