

Multilingual Lexical Acquisition by Bootstrapping Cognate Seed Lexicons

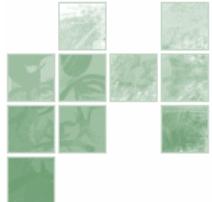
Kornél Markó

Stefan Schulz

Udo Hahn

Medical Informatics,
Freiburg University Hospital
(Germany)

Jena University, Language &
Information Engineering
(Germany)



Cross-Language Text Retrieval



„Korrelation von
Hypertonie und
Läsion der
Weißen Substanz“

Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed: Home Bookmarks mozilla.org Latest Builds Search Print

□ 1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1. Related Articles, Links

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

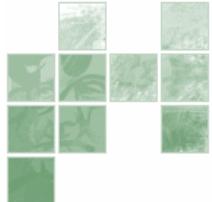
Full text article at
[stroke.ahajournals.org](#)

Interaction between hypertension, apoE, and cerebral white matter lesions.

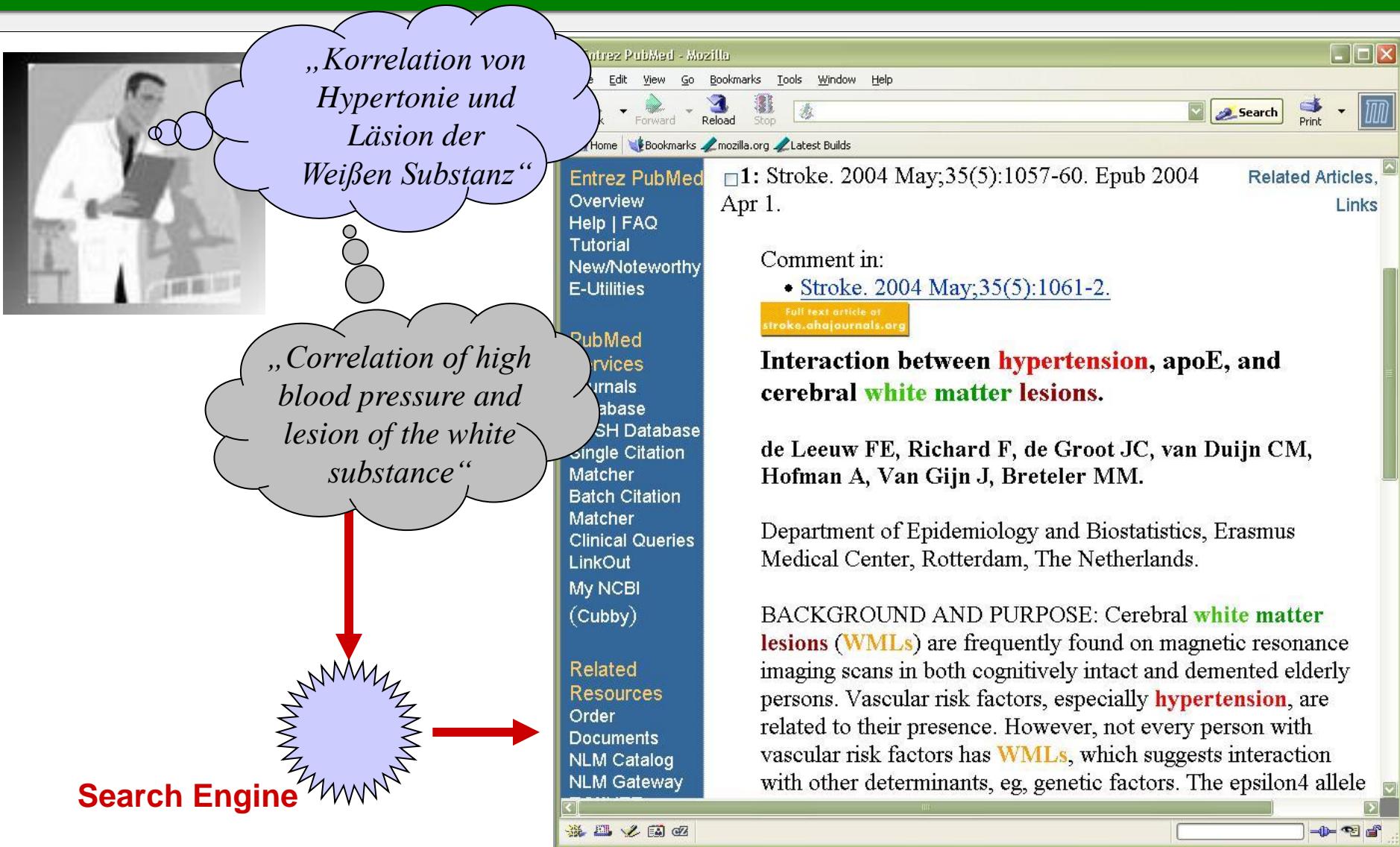
de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele



Cross-Language Text Retrieval



„Korrelation von Hypertonie und Läsion der Weißen Substanz“

„Correlation of high blood pressure and lesion of the white substance“

Search Engine

Entrez PubMed - Mozilla Firefox

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop

Home Bookmarks mozilla.org Latest Builds

1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1.

Comment in:

- Stroke. 2004 May;35(5):1061-2.

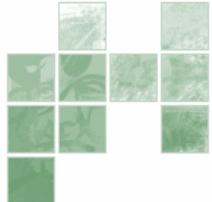
Full text article of stroke.ahajournals.org

Interaction between **hypertension**, apoE, and cerebral **white matter lesions**.

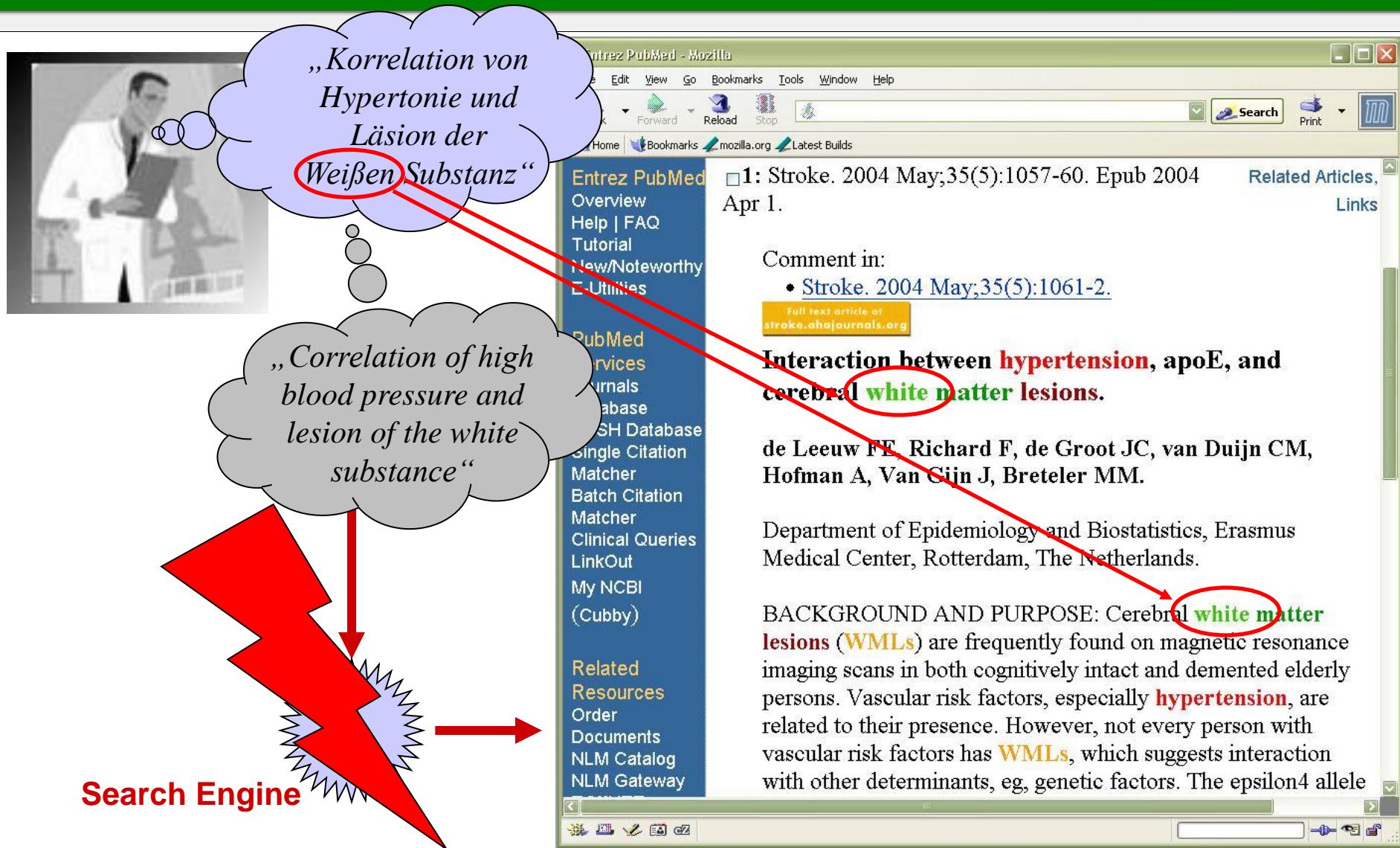
de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

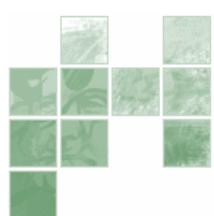
Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral **white matter lesions** (**WMLs**) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially **hypertension**, are related to their presence. However, not every person with vascular risk factors has **WMLs**, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele

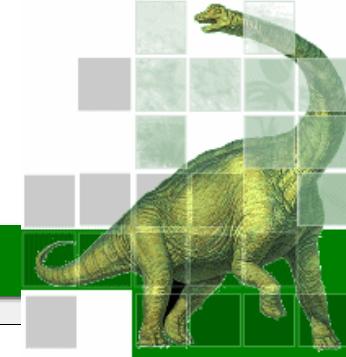


Cross-Language Text Retrieval





MorphoSaurus* semantic indexing system



- Subword oriented:
 - Subwords are atomic conceptual or linguistic units
 - Multilingual subword lexicon:
 - Good coverage for German, English, Portuguese (manually), French, Spanish, Swedish under construction
 - Subword thesaurus
 - Groups synonyms and translations are in equivalence classes.
- #female** = { *woman*, *women*, *female*, *frau-*, *weib-*, *mulher-* }
- ↑
MID (MorphoSaurus Identifier)

stomach, gastr-, diophys-, anti-, bi-, hyper- , -itis -ary, -ion, -it, -is, -o-, -s-

*<http://www.morphosaurus.net>

Example of MorphoSaurus Indexing

High TSH values suggest the diagnosis of primary hypothyroidism ... 

Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose ... 

Original

Orthographic
Normalization
*Orthografic
Rules*

Interlingua of MIDs

#up tsh #value #suggest
#diagnost #primar #small
#thyre

#up tsh #value #permit
#diagnost #primar #small
#thyre



Semantic
Normalization
*Subword
Thesaurus*

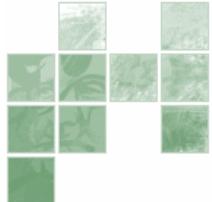
high tsh values suggest the diagnosis of primary hypothyroidism ... 

erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose ... 

Segmenter
Subword Lexicon

high tsh value s suggest the diagnos is of primar y hypo thyroid ism 

er hoeh te tsh wert e erlaub en die diagnos e einer primaer en hypo thyre ose 



MorphoSaurus Indexing

Entrez PubMed - Mozilla Firefox

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed Search Print

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed Overview Help | FAQ Tutorial New/Noteworthy E-Utilities

PubMed Services Journals Database MeSH Database Single Citation Matcher Batch Citation Matcher Clinical Queries LinkOut My NCBI (Cubby)

Related Resources Order Documents NLM Catalog NLM Gateway

□1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1.

Related Articles Links

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

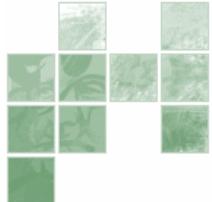
[Full text article at stroke.ahajournals.org](#)

Interaction between hypertension, apoE, and cerebral white matter lesions.

de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele



MorphoSaurus Indexing

Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed Search Print

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed Overview Help | FAQ Tutorial New/Noteworthy E-Utilities

PubMed Services Journals Database MeSH Database Single Citation Matcher Batch Citation Matcher Clinical Queries LinkOut My NCBI (Cubby) Related Resources Order Documents NLM Catalog NLM Gateway

Comment in:
• [Stroke. 2004 May;35\(5\):1061-2.](#)

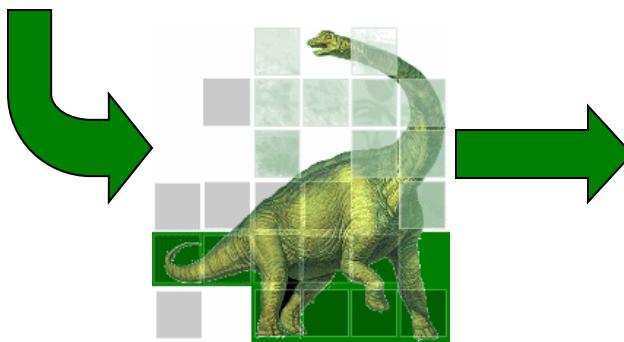
[Full text article at stroke.ahajournals.org](#)

Interaction between hypertension, apoE, and cerebral white matter lesions.

de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele



Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop mozilla.org Latest Builds

Entrez PubMed Overview Help | FAQ Tutorial New/Noteworthy E-Utilities

PubMed Services Journals Database MeSH Database Single Citation Matcher Batch Citation Matcher Clinical Queries LinkOut My NCBI (Cubby) Related Resources Order Documents NLM Catalog NLM Gateway

Comment in:
• [Stroke. 2004 May;35\(5\):1061-2.](#)

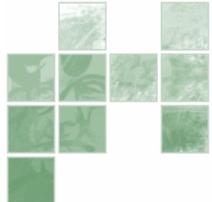
[Full text article at stroke.ahajournals.org](#)

#interact #hyper #tens , apoE , #cerebr #whit #matter #lesion .

de leeuw fe , richard f , de groot jc , van duijn cm , hofman a , van gijn j , breteler mm .

#department #epidem #logic #bio #statist , erasmus #medic #centr , rotterdam , #dutch .

#back #ground #purpos : #cerebr #whit #matter #lesion (wmls) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special #hyper #tens , #relat #presenc . #not #total #human #vascul #risk #factor wmls ,# suggest #interact #other #determin , eg ,



MorphoSaurus Search



„Korrelation von
Hypertonie und
Läsion der
Weißen Substanz“

Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Forward Reload Stop

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed

Overview Help | FAQ Tutorial New/Noteworthy E-Utilities

PubMed Services Journals Database MeSH Database Single Citation Matcher Batch Citation Matcher Clinical Queries LinkOut My NCBI (Cubby)

Related Resources Order Documents NLM Catalog NLM Gateway

□ 1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1.

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

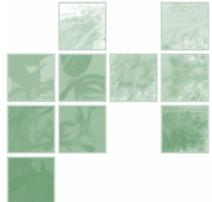
Full text article at stroke.ahajournals.org

#interact #hyper #tens , apoe , #cerebr #whit #matter #lesion .

de leeuw fe , richard f , de groot jc , van duijn cm , hofman a , van gijn j , breteler mm .

#department #epidem #logic #bio #statist , erasmus #medic #centr , rotterdam , #dutch .

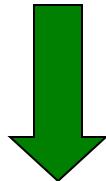
#back #ground #purpos : #cerebr #whit #matter #lesion (wmls) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special #hyper #tens , #relat #presenc . #not #total #human #vascul #risk #factor wmls ,# suggest #interact #other #determin, eg ,



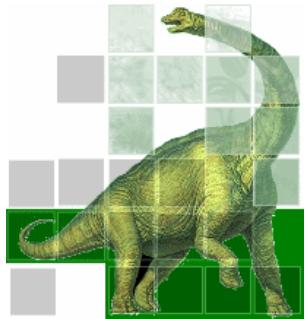
MorphoSaurus Search



„Korrelation von
Hypertonie und
Läsion der
Weißen Substanz“



„#correl #hyper
#tens #lesion
#whit #matter“



Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Forward Reload Stop

Home Bookmarks mozilla.org Latest Builds

Entrez PubMed

Overview Help | FAQ Tutorial New/Noteworthy E-Utilities

PubMed Services Journals Database MeSH Database Single Citation Matcher Batch Citation Matcher Clinical Queries LinkOut My NCBI (Cubby)

Related Resources Order Documents NLM Catalog NLM Gateway

□ 1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1.

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

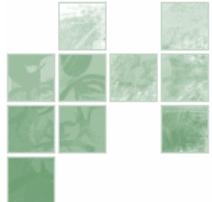
Full text article at stroke.ahajournals.org

#interact #hyper #tens , apoe , #cerebr #whit #matter #lesion .

de leeuw fe , richard f , de groot jc , van duijn cm , hofman a , van gijn j , breteler mm .

#department #epidem #logic #bio #statist , erasmus #medic #centr , rotterdam , #dutch .

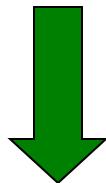
#back #ground #purpos : #cerebr #whit #matter #lesion (wmls) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special #hyper #tens , #relat #presenc . #not #total #human #vascul #risk #factor wmls ,# suggest #interact #other #determin , eg ,



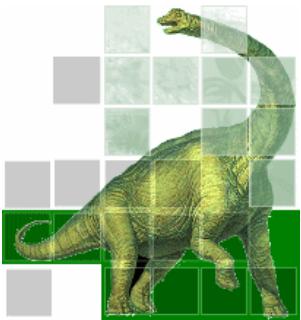
MorphoSaurus Search



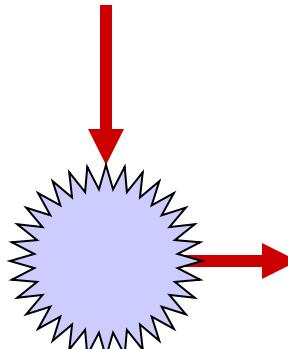
„Korrelation von
Hypertonie und
Läsion der
Weißen Substanz“



„#correl #hyper
#tens #lesion
#whit #matter“



Search Engine



The screenshot shows a Mozilla Firefox browser window displaying a PubMed search result. The URL bar shows "Entrez PubMed - Mozilla". The main content area displays the following text:

□1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1.

Comment in:
• [Stroke. 2004 May;35\(5\):1061-2.](#)

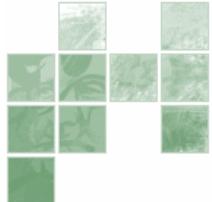
Full text article at
stroke.ahajournals.org

#interact #hyper #tens , apoe , #cerebr #whit #matter #lesion .

de leeuw fe , richard f , de groot jc , van duijn cm , hofman a , van gijn j , breteler mm .

#department #epidem #logic #bio #statist , erasmus #medic #centr , rotterdam , #dutch .

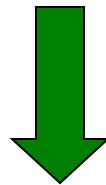
#back #ground #purpos : #cerebr #whit #matter #lesion (wmls) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special #hyper #tens , #relat #presenc . #not #total #human #vascul #risk #factor wmls ,# suggest #interact #other #determin , eg ,



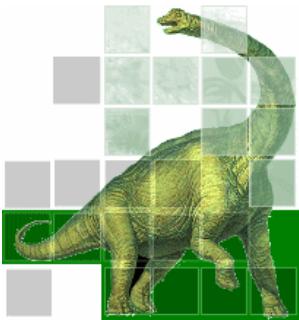
MorphoSaurus Search



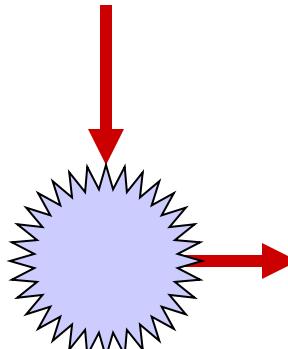
„Korrelation von Hypertonie und Läsion der Weißen Substanz“



„#correl #hyper
#tens #lesion
#whit #matter“



Search Engine



Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

Forward Reload Stop

Home Bookmarks mozilla.org Latest Builds

Related Articles, Links

□ 1: Stroke. 2004 May;35(5):1057-60. Epub 2004 Apr 1.

Comment in:

- [Stroke. 2004 May;35\(5\):1061-2.](#)

Full text article at [stroke.ahajournals.org](#)

#interact #hyper #tens , apoe , #cerebr #whit #matter #lesion .

de leeuw fe , richard f , de groot jc , van duijn cm , hofman a , van gijn j , breteler mm .

#department #epidem #logic #bio #statist , erasmus #medic #centr , rotterdam , #dutch .

#back #ground #purpos : #cerebr #whit #matter #lesion (wmls) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special #hyper #tens , #relat #presenc . #not #total #human #vascul #risk #factor wmls , # suggest #interact #other #determin , eg ,

MorphoSaurus Database

DE Deutsch (Deutschland)

New Lexeme Thesaurus: Join Rel UnRel Tools: Mesh UMLS WordStat Export Report a Bug Exit

ALL German English Portuguese Spanish French Swedish

Order 1: Lexeme Order 2: Lexeme surg Search Order 1: Lexeme Order 2: Lexeme ope Search

Lexicon

| Lexeme | EqClass | Type | Total: 12 |
|-------------|---------|------|-----------|
| surg | 5654 | Stem | |
| surg | 5654 | Stem | |
| surge | 500107 | Stem | |
| surge | 5654 | Stem | |
| surge | 5654 | Stem | |
| surgeon | 499 | Stem | |
| surgery | 499 | Stem | |
| surgery | 499 | Stem | |
| surgic | 499 | Stem | |
| surgicenter | 33416 | Stem | |
| surgir | 5654 | Stem | |
| surgir | 5654 | Stem | |

Thesaurus

| Eq Class 499 | for indexing | Unjoin |
|--------------|--------------|--------|
| eingriff | | |
| ingrepp | | |
| op | | |
| op | | |

| sense_of | EqClass: 1208 |
|----------|---------------|
| eingriff | |
| ingrepp | |

| Eq Class 499 | for indexing | Unjoin |
|--------------|--------------|--------|
| chirurg | | |
| chirurgen | | |
| chirurgie | | |
| chirurgin | | |
| chirurgisch | | |
| operat | | |
| operation | | |
| operations | | |
| operier | | |
| operat | | |
| surgeon | | |
| surgery | | |
| surgic | | |
| cirurg | | |

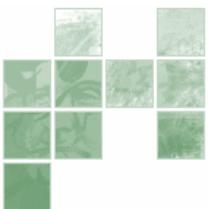
| sense_of | EqClass: 1208 |
|----------|---------------|
| eingriff | |
| ingrepp | |

| Eq Class 499 | for indexing | Unjoin |
|--------------|--------------|--------|
| chirurg | | |
| chirurgen | | |
| chirurgie | | |
| chirurgin | | |
| chirurgisch | | |
| operat | | |
| operation | | |
| operations | | |
| operier | | |
| operat | | |
| surgeon | | |
| surgery | | |
| surgic | | |
| cirurg | | |

| sense_of | EqClass: 501250 |
|----------|-----------------|
| op | |

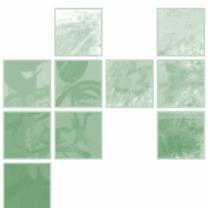
| word_part_of | EqClass: 501250 |
|--------------|-----------------|
| op | |

MorphoEdit WEB © PUC-PR Preferred Resolution: 1024x768



Lexicon Construction

- Manual construction of lexicon labor-intensive
- Many medical terms exhibit high degree of similarity across languages (“cognates”): *surgical, chirurgisch, chirurgique, cirúrgico, quirurgico, kirurgisk*
- Others don’t (“non-cognate translations”): *spleen, Milz, rate, baço, bazo, mjälten*
- Project: development of automated technique for lexicon acquisition for new languages
- Case study:
Acquisition of medical subword lexemes for target languages Spanish, French and Swedish



Automatic Lexicon Acquisition

Two steps of lexicon acquisition:

1. *Generation and Validation* of trusted subword cognates for the target languages
2. *Bootstrapping*: iterative learning of non cognate subword translations



Resources for Cognate Acquisition

- Manually constructed subword lexicons in the source languages:
 - German (~22,000 stems)
 - English (~22,000 stems)
 - Portuguese (~14,000 stems)
- Manually created list of prefixes and suffixes for the target languages
- Medical corpora for all languages
- Word frequency lists generated from these corpora
- Language pair specific string substitution rules



Generating Cognate Candidates

List of Portuguese
Subwords
(14,004 stems):

...
estomag-
mulher-
...



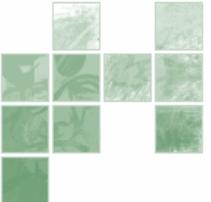
Generating Cognate Candidates

List of Portuguese
Subwords
(14,004 stems):

...
estomag-
mulher-
...

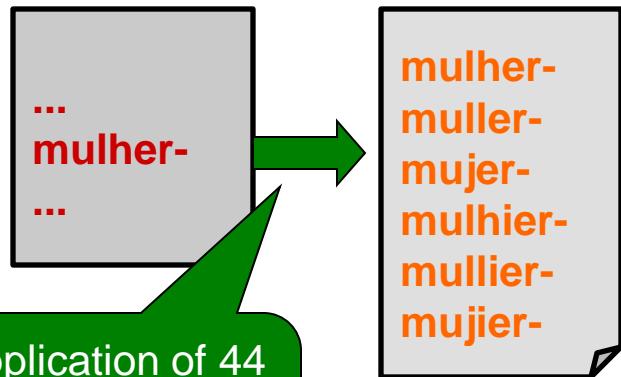
Application of 44
string
substitution
rules

| Rule (Port. » Span.) |
|-------------------------|
| qua » cua |
| eia » ena |
| ss » s |
| lh » j |
| lh » ll |
| l » ll |
| f » h |
| ... |

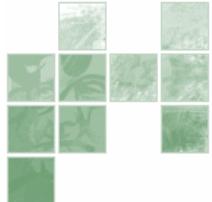


Generating Cognate Candidates

List of Portuguese
Subwords
(14,004 stems):

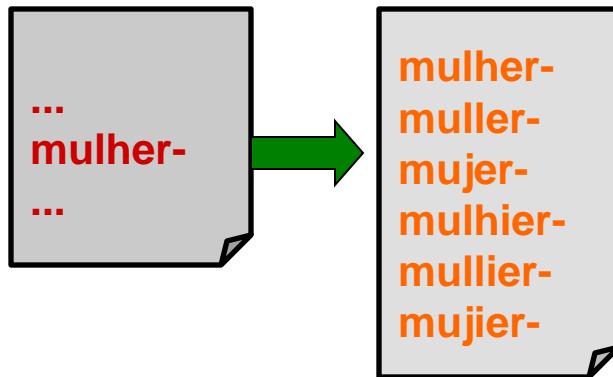


Application of 44
string
substitution
rules

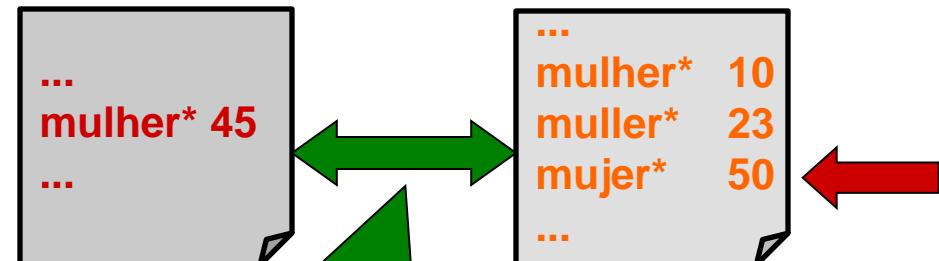


Generating Cognate Candidates

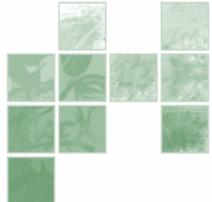
List of Portuguese
Subwords
(14,004 stems):



Word frequency lists
derived from unrelated corpora:
Size (Portuguese) ~ Size(Spanish)

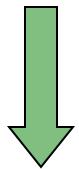


Comparison between word frequency lists:
Choose that cognate alternative with the *most similar* corpus frequency

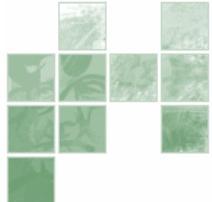


Semantic Mapping

mulher-  mujer-

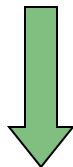


MID: #female = { woman, women, female-, frau-, weib-, mulher-, mujer- }



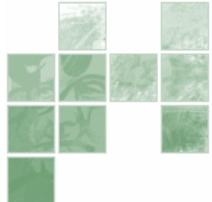
Semantic Mapping

mulher → mujer



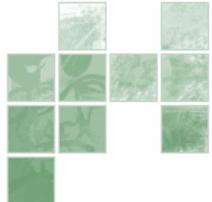
MID: #female = { woman, women, female-, frau-, weib-, mulher-, mujer-}

| Language Pair | Source Lexicon | Cognates acquired |
|--------------------|----------------|-------------------|
| Portuguese-Spanish | 14,004 | 8,644 |
| German-French | 21,705 | |
| English-French | 21,501 | 9,536 |
| German-Swedish | 21,705 | |
| English-Swedish | 21,501 | 6,086 |



Use of parallel corpora to identify false cognates:

- Example:
 - Portuguese *crianc-* (*child*) ↔ Spanish *crianz-* (*breed*)
 - Portuguese *crianc-* (*child*) ↔ Spanish *nin-* (*child*)
- UMLS Metathesaurus as parallel corpus
 - English-Spanish: 60,526 translations
 - English-French: 17,130 translations
 - English-Swedish: 10,953 translations
- English-Spanish Example
 - „*Cell Growth*“ ↔ „*Crecimiento Celular*“



Cognate Validation

- Use generated cognate seed lexicons to process the UMLS translations with the MorphoSaurus



Cognate Validation

- Use generated cognate seed lexicons to process the UMLS translations with the MorphoSaurus
- Whenever a MID co-occurs on both sides of a translation pair, the corresponding lexicon entry is taken to be valid
- Discard candidates that never matched

„Abdominal wall procedure“:

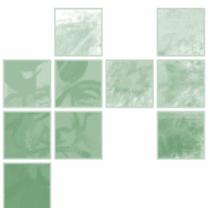
|abdomin|all|
|wall|
|proced|ure|

#abdom
#wall
#operat

„Cirugia de la pared abdominal“:

|cirug|ia|
|pared|
|abdomin|al|

#wall
#abdom



Cognate Validation

- Use generated cognate seed lexicons to process the UMLS translations with the MorphoSaurus
- Whenever a MID co-occurs on both sides of a translation pair, the corresponding lexicon entry is taken to be valid
- Discard candidates that never matched

„Abdominal wall procedure“:

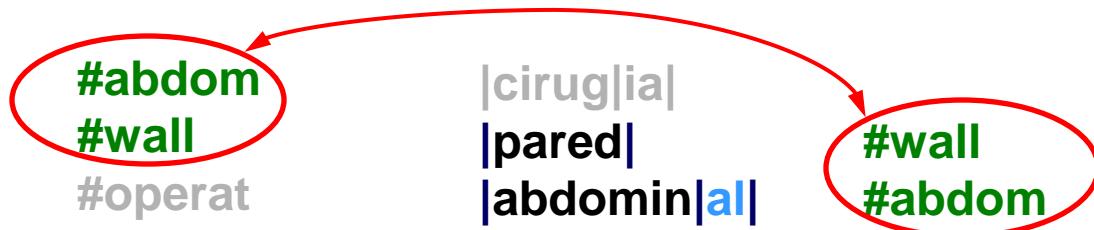
|abdomin|all
|wall|
|proced|ure|

#abdom
#wall
#operat

„Cirugia de la pared abdominal“:

|cirug|ia
|pared|
|abdomin|all|

#wall
#abdom





Cognate Validation

- Use generated cognate seed lexicons to process the UMLS translations with the MorphoSaurus
- Whenever a MID co-occurs on both sides of a translation pair, the corresponding lexicon entry is taken to be valid
- Discard candidates that never matched

„Abdominal wall procedure“:

|abdomin|all
|wall|
|proced|ure|

#abdom
#wall
#operat

„Cirugia de la pared abdominal“:

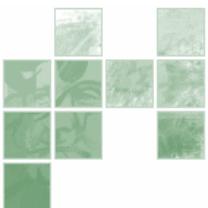
|cirug|ia
pared
abdomin|all

#wall
#abdom



Cognate Validation

| Language Pair | Source Lexicon | Cognates acquired | Cognates validates |
|---------------------------------------|----------------------------|--|--------------------|
| Portuguese-Spanish | 14,004 | 8,644 | 3,230 |
| German-French | 21,705 | 9,536 | 3,540 |
| English-French | 21,501 | | |
| German-Swedish | 21,705 | 6,086 | 1,565 |
| English-Swedish | 21,501 | | |
| „Abdominal wall procedure“: | | „Cirugia de la pared abdominal“: | |
| abdomin all wall proced ure | #abdom #wall #operat | cirug ia <input checked="" type="checkbox"/> pared <input checked="" type="checkbox"/> abdomin all | #wall #abdom |



Step 2: Bootstrapping

- Acquisition of non-cognates uses:
 - validated cognate seed lexicons
 - parallel corpora

„Abdominal wall procedure“:

|abdomin|all| #abdom
|wall| #wall
|proced|ure| #operat

„Cirugia de la pared abdominal“:

|cirug|ia| #wall
|pared| #abdom
|abdomin|all| #abdom



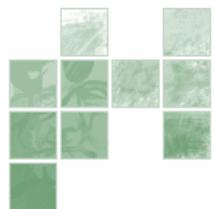
Bootstrapping Algorithm

→ For every UMLS term pair do

→ „*Abdominal wall procedure*“: „*Cirugia de la pared abdominal*“:

|abdomin|all| #abdom
|wall| #wall
|proced|ure| #operat

|cirug|ia| #wall
|pared| #abdom
|abdomin|all| #abdom



Bootstrapping Algorithm

For every UMLS term pair do

→ If there is exactly one invalid segmentation in target language

„Abdominal wall procedure“:

|abdomin|all|
|wall|
|proced|ure|

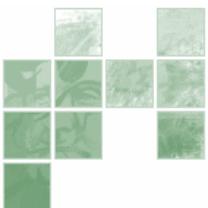
#abdom
#wall
#operat



„Cirugia de la pared abdominal“:

|cirug|ia|
|pared|
|abdomin|all|

#wall
#abdom



Bootstrapping Algorithm

For every UMLS term pair do

- If there is exactly one invalid segmentation in target language
- If there is exactly one more MID in source language



„Abdominal wall procedure“:

|abdomin|all
|wall|
|proced|ure|

#abdom
#wall
#operat

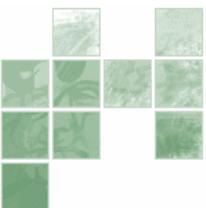


„Cirugia de la pared abdominal“:

|cirug|ia|
|pared|
|abdomin|all|

#wall
#abdom





Bootstrapping Algorithm

For every UMLS term pair do

If there is exactly one invalid segmentation in target language

If there is exactly one more MID in source language

Take supernumerary MID and invalid segmentation from target



„Abdominal wall procedure“:

|abdomin|all
|wall|
|proced|ure|

#abdom
#wall
#operat

„Cirugia de la pared abdominal“:

|cirug|ia
|pared|
|abdomin|all
#wall
#abdom





Bootstrapping Algorithm

For every UMLS term pair do

If there is exactly one invalid segmentation in target language

If there is exactly one more MID in source language

Take supernumerary MID and invalid segmentation from target

Restore invalid segmentation and strip off potential affixes



„Abdominal wall procedure“:

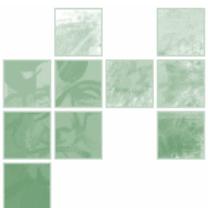
|abdomin|all
|wall|
|proced|ure|

#abdom
#wall
#operat

„Cirugia de la pared abdominal“:

|cirug|ia| → |cirug|ia|
|pared|
|abdomin|all

#wall
#abdom



Bootstrapping Algorithm

For every UMLS term pair do

If there is exactly one invalid segmentation in target language

If there is exactly one more MID in source language

Take supernumerary MID and invalid segmentation from target

Restore invalid segmentation and strip off potential affixes

Add new stem into target lexicon. Link it to source MID.



„Abdominal wall procedure“:

|abdomin|all
|wall|
|proced|ure|

#abdom
#wall
#operat

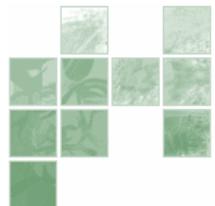
„Cirugia de la pared abdominal“:

|cirug|ia|
|pared|
|abdomin|all|

→ |cirug|**ia**|
#wall
#abdom

#operat = { proced, surgery, operat, prozess, operier, proced, process, metod, cirug }





Bootstrapping Algorithm

For every UMLS term pair do

If there is exactly one invalid segmentation in target language

If there is exactly one more MID in source language

Take supernumerary MID and invalid segmentation from target

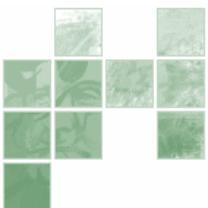
Restore invalid segmentation and strip off potential affixes

Add new stem into target lexicon. Link it to source MID.

→ Repeat all until quiescence

,*Abdominal wall procedure*“:

,*Cirugia de la pared abdominal*“:



Bootstrapping Algorithm

For every UMLS term pair do

If there is exactly one invalid segmentation in target language

If there is exactly one more MID in source language

Take supernumerary MID and invalid segmentation from target

Restore invalid segmentation and strip off potential affixes

Add new stem into target lexicon. Link it to source MID.

Repeat all until quiescence

„Abdominal wall procedure“:

„Skin operations“:

|skin|

|operat|ions|

#derma

#operat

„Cirugia de la pared abdominal“:

„Cirugia de piel“:

|cirug|ia|

|piel|

#operat



Bootstrapping Algorithm

For every UMLS term pair do

If there is exactly one invalid segmentation in target language

If there is exactly one more MID in source language

Take supernumerary MID and invalid segmentation from target

Restore invalid segmentation and strip off potential affixes

Add new stem into target lexicon. Link it to source MID.

Repeat all until quiescence

„Abdominal wall procedure“:

„Skin operations“:

|skin|

|operat|ions|

#derma

#operat

„Cirugia de la pared abdominal“:

„Cirugia de piel“:

|cirug|ia|

|piel|

#operat

|piell|

#derma = { derm, cutis, skin, haut, kutis, pele, cutis, piel } ←



Bootstrapping Algorithm

For every UMLS term pair do

If there is exactly one invalid segmentation in target language

If there is exactly one more MID in source language

Take supernumerary MID and invalid segmentation from target

Restore invalid segmentation and strip off potential affixes

Add new stem into target lexicon. Link it to source MID.

Repeat all until quiescence

„Abdominal wall procedure“:

„Skin operations“:

„Skin abnormalities“:

„Cirugia de la pared abdominal“:

„Cirugia de piel“:

„Malformacion de la piel“:

|skin| #derma
|abnorm|alities| #anomal

|malformation| #derma
|piel|



Bootstrapping Algorithm

For every UMLS term pair do

If there is exactly one invalid segmentation in target language

If there is exactly one more MID in source language

Take supernumerary MID and invalid segmentation from target

Restore invalid segmentation and strip off potential affixes

Add new stem into target lexicon. Link it to source MID.

Repeat all until quiescence

„Abdominal wall procedure“:

„Skin operations“:

„Skin abnormalities“:

|skin|

#derma

|abnorm|alities| #anomal

„Cirugia de la pared abdominal“:

„Cirugia de piel“:

„Malformacion de la piel“:



|malformation|

|piel|



#derma

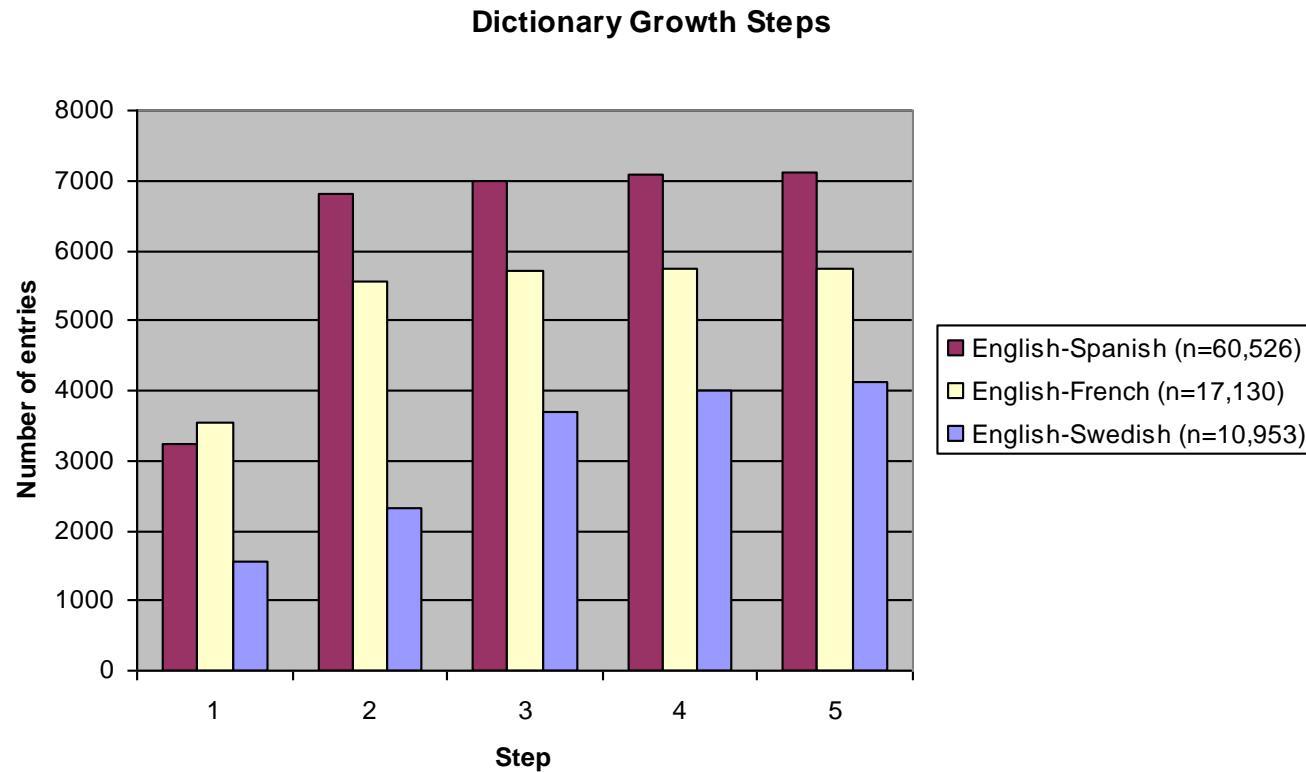
|malform|ation|



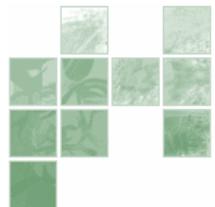
#anormal = { abnorm, anomal, abnorm, anomal, abnorm, anomal, malform }



Bootstrapping Results



Total: 7,154 Spanish,
5,734 French and
4,148 Swedish entries acquired

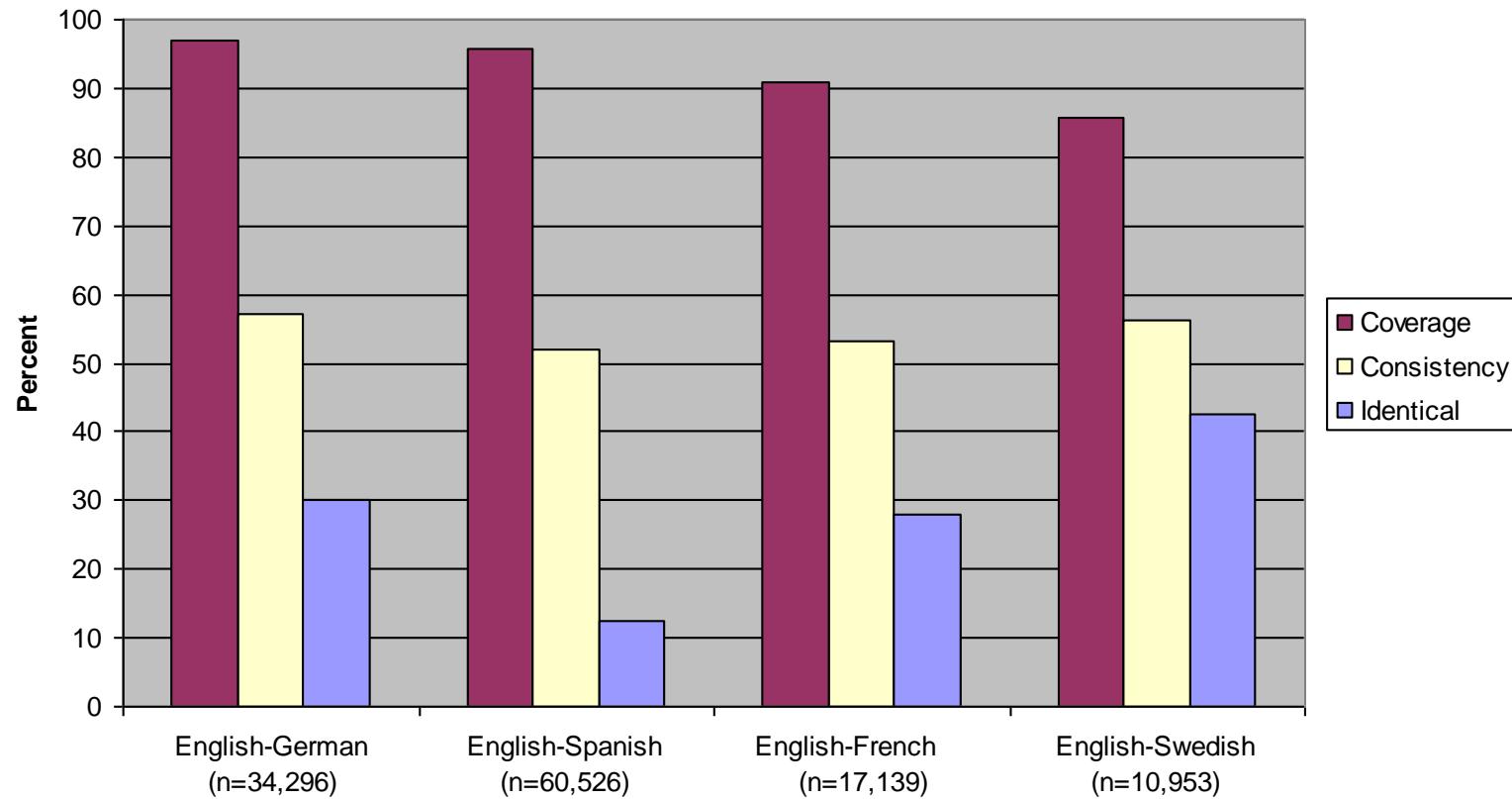


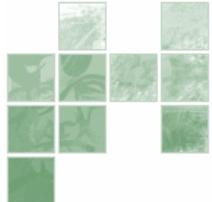
Evaluation

- Process the English-Spanish, English-French and English-Swedish UMLS translation pairs with the MorphoSaurus system
- Additionally process Spanish-French, Spanish-Swedish, and French-Swedish UMLS translation pairs
- Measures:
 - Coverage: At least one MID co-occurs on both sides
 - Consistency
$$C_{AU(i)} = \frac{(100 * A)}{(A + N + M)}$$
 - A : Number of MIDs co-occurring on both sides
 - N, M : Number of MIDs occurring on only one side
 - Identical Indexes

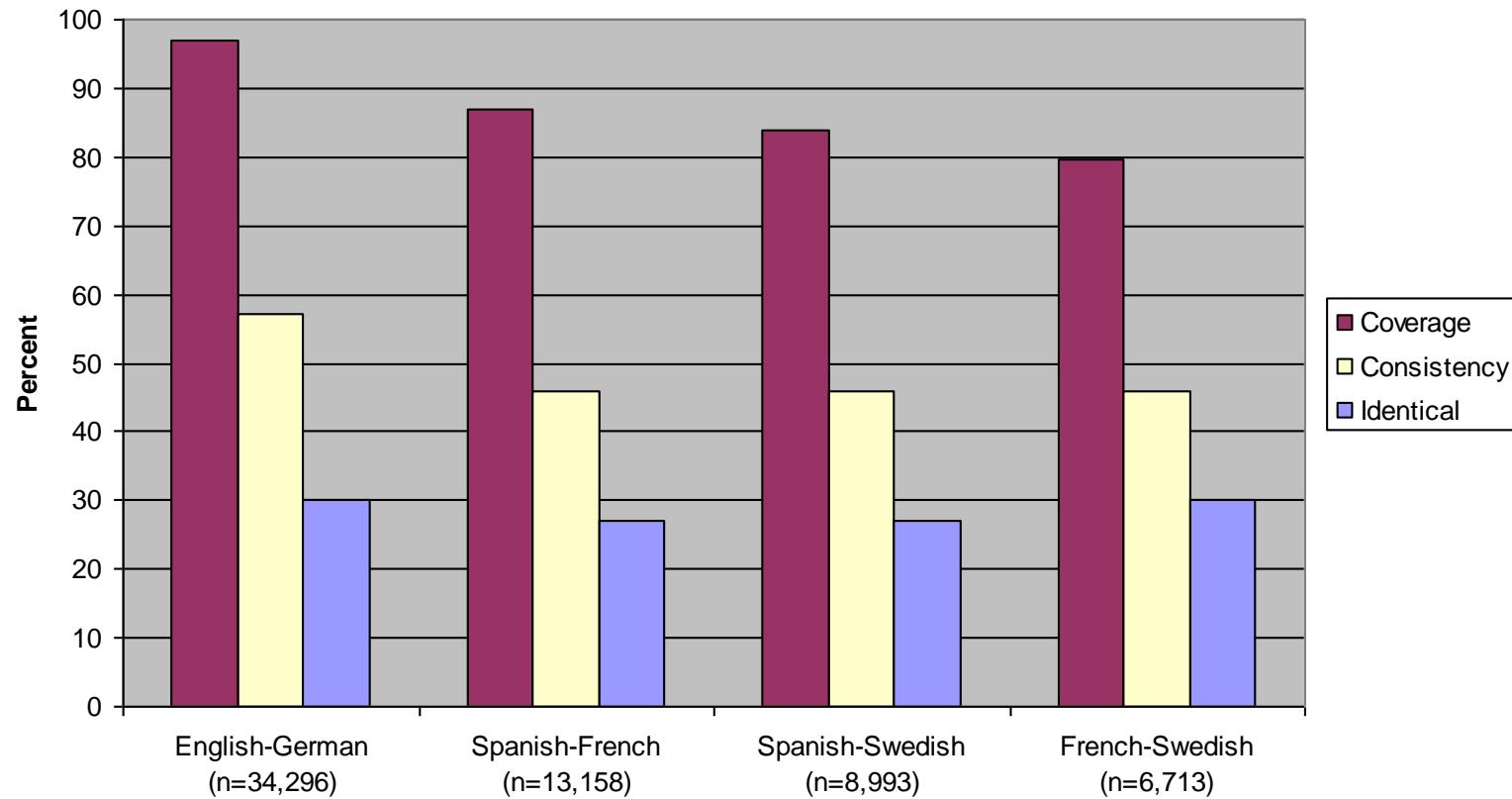


Results





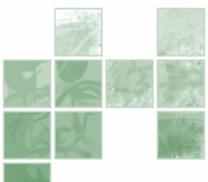
Results





Conclusion

- Cross-Language Document Retrieval based on a language-independent, interlingual layer.
- Automated approach for acquiring lexicon entries for new languages
- Significant amount cognate subwords can be acquired using simple string substitution rules.
- These seed lexicons are further enlarged by subword translations which are *not* cognates by bootstrapping and using parallel corpora.
- Current limitation: size of parallel corpora for bootstrapping step



Automatic Lexicon Acquisition for a Medical Cross-Language Information Retrieval System

Kornél Markó

Stefan Schulz

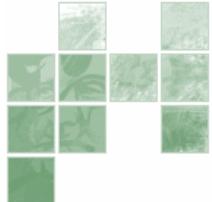
Udo Hahn

Medical Informatics,
Freiburg University Hospital
(Germany)

Jena University, Language &
Information Engineering
(Germany)

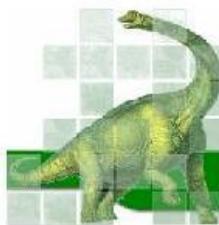


www.MorphoSaurus.net



MorphoSaurus Search

Search Engine



Search: (English/German/Portuguese)

Language:

Information Level:

MorphoSaurus:

Search results for 'darmkrebsrisiko':

Best 20 Documents of 2095 matches (298 msec)

Entrez PubMed (94%) Expert information

Keywords: MEDLINE, NCBI, National Center for Biotechnology Information, National Library of Medicine, NLM, PubMed

Description: PubMed is the National Library of Medicine's search service that provides access to over 11 million citations in MEDLINE, PreMEDLINE, and other related databases, with links to participating online journals.

... almost unaltered **risk** of all other **cancers** (SIR, 1.2; 95% CI, 1.0-1.4), including nonelevated **risks** for several **gastrointestinal** tract **cancers**. At 10 years of follow up, the absolute **risk** of liver **cancer** was 6% among men and 1.5% among women. With 21 liver **cancers** and 508 nonhepatobiliary **cancers**, first degree ... (Cached)

http://supreme.coling.uni-jena.de/~coling/search_engine_docs/original/medline/14724826.html

Deutsches Ärzteblatt (91%) Expert information

... DEUTSCHES ÄRZTEBLATT PRINT wErhöhtes **Risiko für gastrointestinale Karzinome** nach Cholezystektomie Deutsches Ärzteblatt 99, Ausgabe 26 vom 28.06.2002, Seite A 1824 B 1541 C 1437 MEDIZIN: Referiert Eine Cholezystektomie führt möglicherweise über toxische Effekte des alkalischen Refluats auf die Speiseröhrenschleimhaut zu einem mäßiggradigen Anstieg des **Adenokarzinomrisikos** der Speiseröhre (**Barrett Karzinom**). Aber ... (Cached)

http://supreme.coling.uni-jena.de/~coling/search_engine_docs/original/aerzteblatt/artikeldruck.asp?3id=32191.html

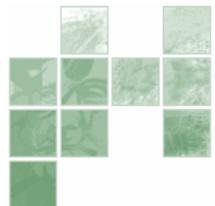
Dickdarm und Mastdarmkrebs Kolorektales Karzinom (91%) Patient information

Keywords: Dickdarm, Mastdarmkrebs

Description: Dickdarm und Mastdarmkrebs ist eine bösartige Schleimhautwucherung im Dickdarm oder Mastdarm. Ärzte bezeichnen diese Krebsart auch als kolorektales Karzinom (von griechisch kolon, Darm und lateinisch intestinum rectum, Enddarm). Häufig wird der Dickdarm und Mastdarmkrebs auch nur als Kolonkarzinom bezeichnet, obwohl ein KOLONkarzinom im eigentlichen Sinne nur Dickdarmkrebs ist.

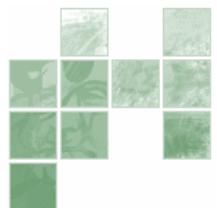
... Werbung Sponsoring NetDoctor.com Dickdarm und **Mastdarmkrebs** (Kolorektales **Karzinom**) Prof. Dr. med. Stefan Endres, Facharzt für Innere Medizin und **Gastroenterologie** Was ist Dickdarm und **Mastdarmkrebs**? Dickdarm und **Mastdarmkrebs** ist eine **bösartige** Schleimhautwucherung im Dickdarm oder Mastdarm. Ärzte bezeichnen diese **Krebsart** auch als kolorektales **Karzinom** (von griechisch kolon, **Darm** und lateinisch **intestinum** ... (Cached)

http://www.netdoktor.de/krankheiten/Fakta/dickdarm_mastdarmkrebs.htm



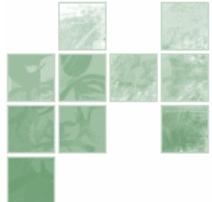
Evaluation

- OHSUMED-Corpus (Hersh et al., 1994)
 - Subset of MEDLINE
 - ~233,000 English documents
 - 106 English user queries, additionally translated to German, Portuguese, Spanish and Swedish by medical experts
 - query-document pairs have been manually judged for relevance
- Search Engine: Lucene
 - <http://lucene.apache.org/>

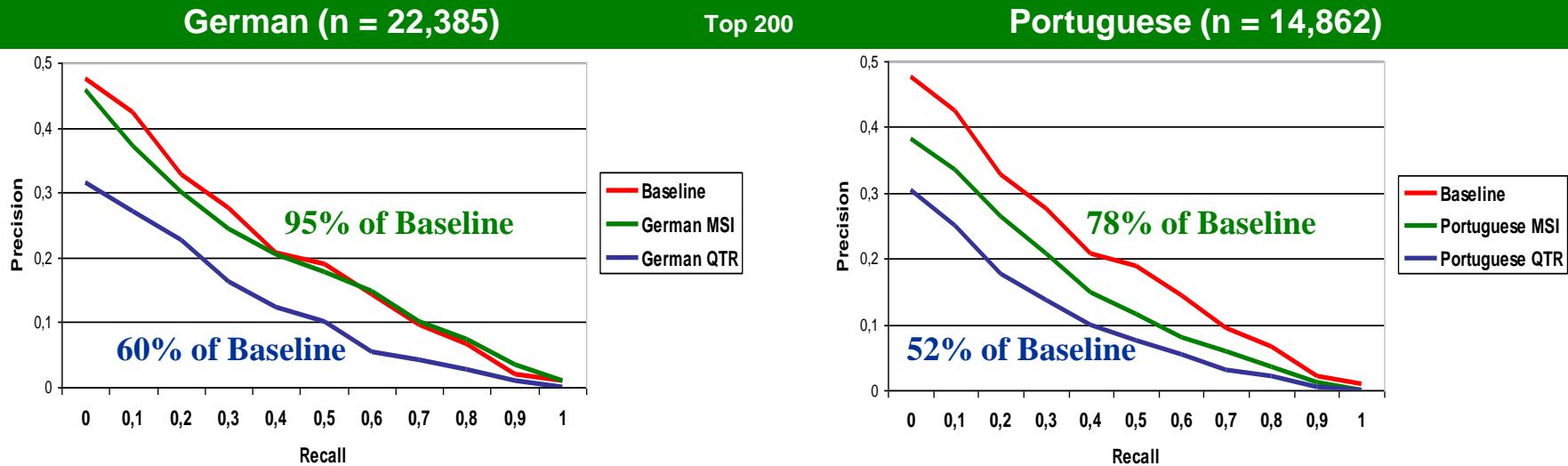


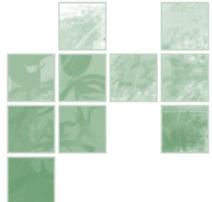
Evaluation

- **Baseline:** monolingual text retrieval
 - (stemmed) English user queries
 - (stemmed) English texts
- **Query translation (QTR)**
 - Google translator
 - Multilingual dictionary compiled from UMLS
- **MorphoSaurus Indexing (MSI)**
 - Interlingual representation of both user queries and documents

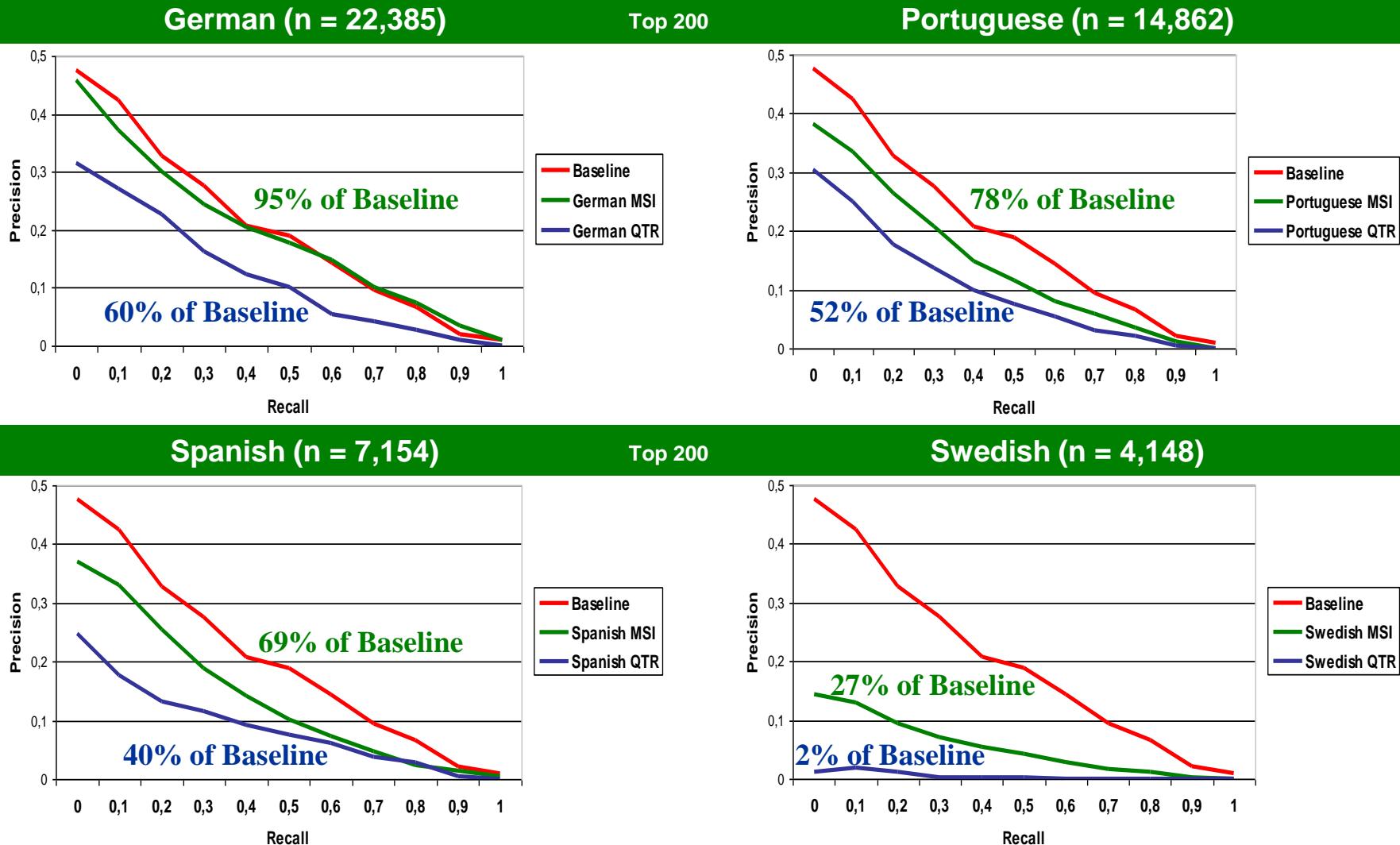


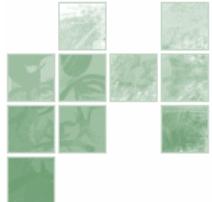
Evaluation Results





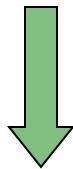
Evaluation Results





Semantic Mapping

mulher → mujer



#female = { woman, women, female, frau, weib, mulher, mujer }

| Language Pair | Source Lexicon | Selected Cognates | Linked MIDs |
|--|----------------|-------------------|-------------|
| Portuguese-Spanish | 14,004 | 8,644 | 6,036 |
| German-Swedish | 21,705 | 4,249 | 3,308 |
| English-Swedish | 21,501 | 4,140 | 3,208 |
| Combined Swedish Evidence (set union) | | 6,086 | 4,157 |