

# 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  
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed

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](http://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

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](http://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 **Weiß**en Substanz“

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

Search Engine

Entrez PubMed - Mozilla

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](http://stroke.ahajournals.org)

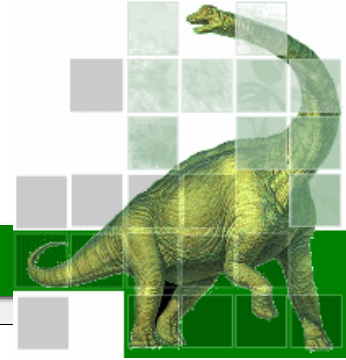
**Interaction between hypertension, apoE, and cerebral white matter lesions.**

de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Giin 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\* semantic indexing system



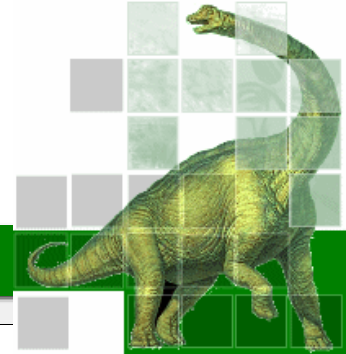
- Subword oriented:
  - Subwords are atomic conceptual or linguistic units
- Multilingual subword lexicon:
  - Good coverage for German, English, Portuguese (manually), French, Spanish, Swedish under construction
- Subword thesaurus
  - Groups synonyms and translations are in equivalence classes.


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

**#female** = { *woman, women, female, frau-, weib-, mulher-* }


↑ MID (MorphoSaurus Identifier)


# Example of MorphoSaurus Indexing




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

**Orthographic Normalization**

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

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

**Orthographic Rules**


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


**Original**


**Segmenter Subword Lexicon**

**Interlingua of MIDs**


**Semantic Normalization**

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

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

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

**Subword Thesaurus**

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

# 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

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](http://stroke.ahajournals.org)

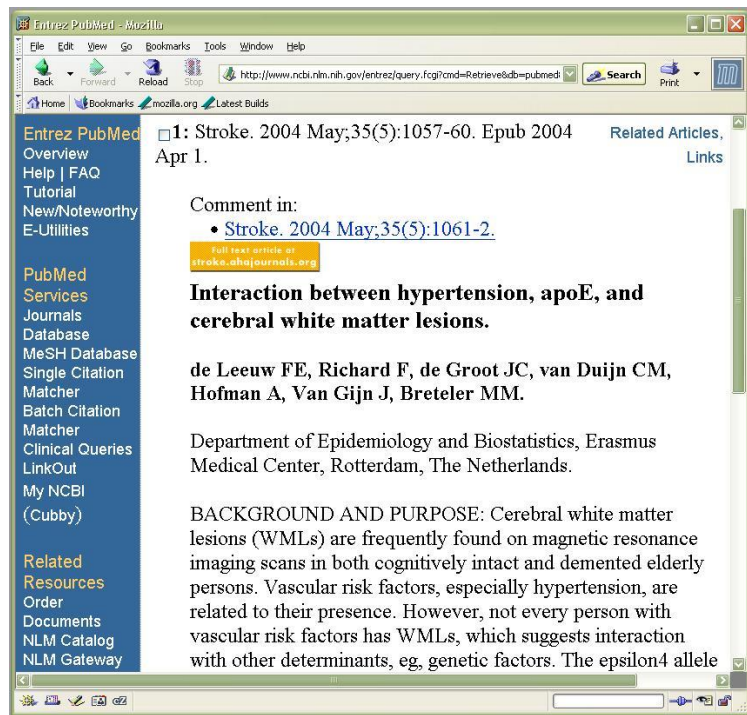
**Interaction between hypertension, apoE, and cerebral white matter lesions.**

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

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed

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

Comment in:

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

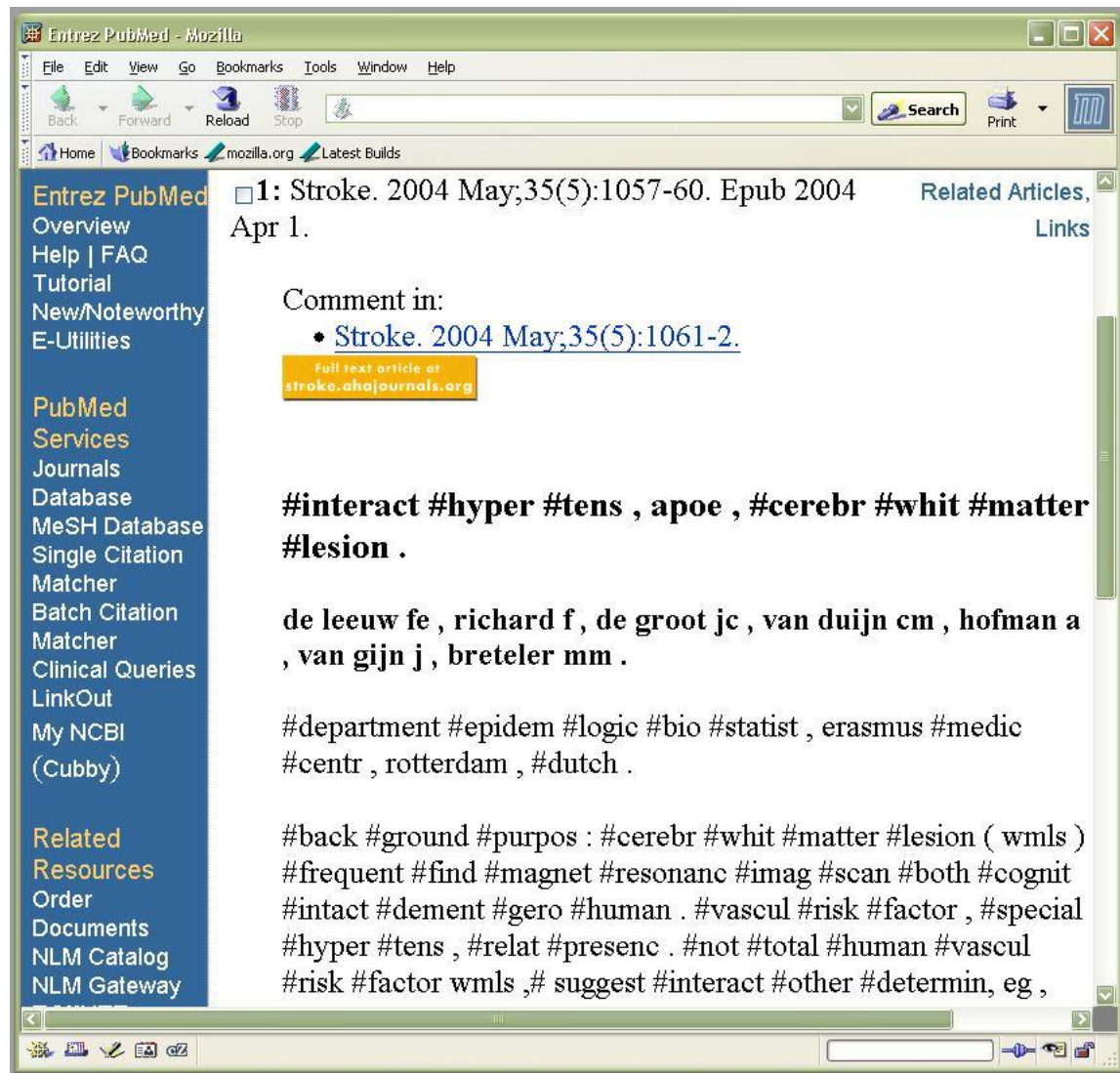
Full text article at [stroke.ahajournals.org](http://stroke.ahajournals.org)

**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

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed

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

Comment in:

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

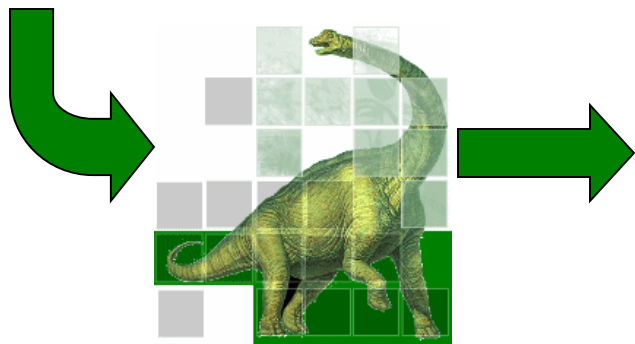
Full text article at [stroke.ahajournals.org](http://stroke.ahajournals.org)

**#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 .**

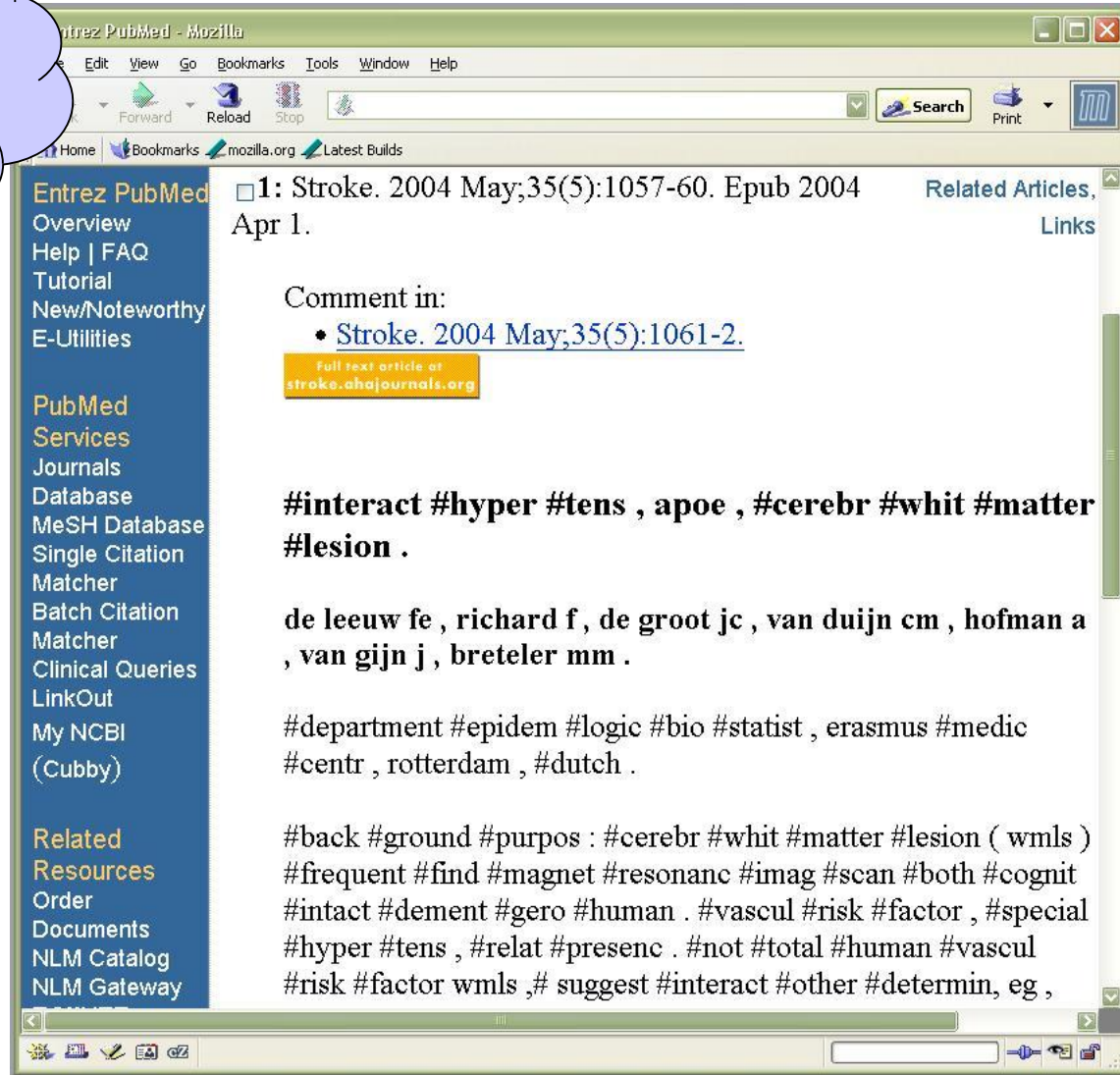
**#back #ground #purpos : #cerebr #whit #matter #lesion ( wmls ) #frequent #find #magnet #resonanc #imag #scan #both #cognit #intact #dement #gero #human . #vascul #risk #factor , #special #hyper #tens , #relat #presenc . #not #total #human #vascul #risk #factor wmls ,# suggest #interact #other #determin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





# MorphoSaurus Search

„Korrelation von  
Hypertonie und  
Läsion der  
Weißen Substanz“



Entrez PubMed - Mozilla

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 ,

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

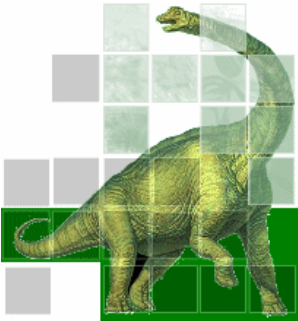
# MorphoSaurus Search



„Korrelation von  
Hypertonie und  
Läsion der  
Weißen Substanz“



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



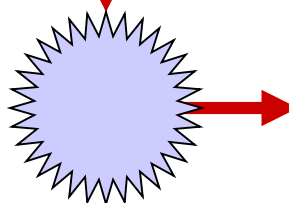
# MorphoSaurus Search



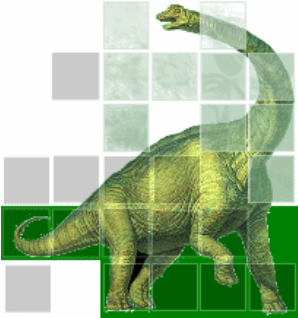
„Korrelation von Hypertonie und Läsion der Weißen Substanz“



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



Search Engine



Entrez PubMed - Mozilla

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](http://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 ,

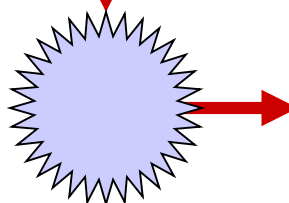
# MorphoSaurus Search



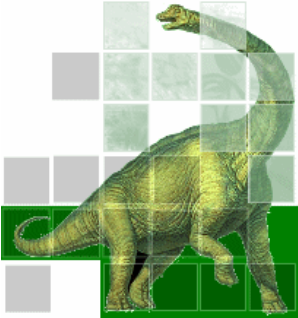
„Korrelation von Hypertonie und Läsion der Weißen Substanz“



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



Search Engine



Entrez PubMed - Mozilla

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](http://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

The screenshot displays the MorphoSaurus Database interface. At the top, there is a navigation bar with buttons for 'New Lexeme', 'Thesaurus: Join Rel UnRel', 'Tools: Mesh UMLS WordStat Export', and 'Report a Bug Exit'. Below this, there are language selection checkboxes for ALL, German, English, Portuguese, Spanish, French, and Swedish. The search interface includes two search boxes: 'Order 1: Lexeme Order 2: Lexeme surg Search' and 'Order 1: Lexeme Order 2: Lexeme ope Search'. The main area is divided into two columns of search results. The left column shows results for 'surg' with a total of 12 items, and the right column shows results for 'ope' with a total of 24 items. Each result row includes a radio button, a flag icon, the lexeme, the EqClass, the Type, and a 'Total' count. Below the search results, there are four panels showing detailed views for 'Eq Class 499' and 'sense\_of EqClass: 1208'. The 'Eq Class 499' panels are labeled 'for indexing' and contain a list of related terms with radio buttons. The 'sense\_of EqClass: 1208' panels show hierarchical relationships like 'sense\_of' and 'word\_part\_of' with their respective EqClasses.

Lexeme	EqClass	Type	Total
<input type="radio"/> <a href="#">surg</a>	5654	Stem	12
<input type="radio"/> <a href="#">surg</a>	5654	Stem	12
<input type="radio"/> <a href="#">surge</a>	500107	Stem	12
<input type="radio"/> <a href="#">surge</a>	5654	Stem	12
<input type="radio"/> <a href="#">surge</a>	5654	Stem	12
<input type="radio"/> <a href="#">surgeon</a>	499	Stem	12
<input checked="" type="radio"/> <a href="#">surgery</a>	499	Stem	12
<input type="radio"/> <a href="#">surgery</a>	499	Stem	12
<input type="radio"/> <a href="#">surgic</a>	499	Stem	12
<input type="radio"/> <a href="#">surgicenter</a>	33416	Stem	12
<input type="radio"/> <a href="#">surgir</a>	5654	Stem	12
<input type="radio"/> <a href="#">surgir</a>	5654	Stem	12

Lexeme	EqClass	Type	Total
<input type="radio"/> <a href="#">open</a>	503850	Invariant	24
<input type="radio"/> <a href="#">open</a>	503850	Stem	24
<input type="radio"/> <a href="#">openbite</a>	31354	Stem	24
<input type="radio"/> <a href="#">opening</a>	6007	Stem	24
<input type="radio"/> <a href="#">oper</a>	499	Stem	24
<input type="radio"/> <a href="#">operabel</a>	7856	Stem	24
<input type="radio"/> <a href="#">operabil</a>	7856	Stem	24
<input type="radio"/> <a href="#">operabil</a>	7856	Stem	24
<input type="radio"/> <a href="#">operabl</a>	7856	Stem	24
<input type="radio"/> <a href="#">operat</a>	499	Stem	24
<input type="radio"/> <a href="#">operat</a>	499	Stem	24
<input type="radio"/> <a href="#">operat</a>	499	Stem	24
<input type="radio"/> <a href="#">operat</a>	499	Stem	24
<input type="radio"/> <a href="#">operation</a>	499	Stem	24

**Lexicon**

**Thesaurus**



# Lexicon Construction

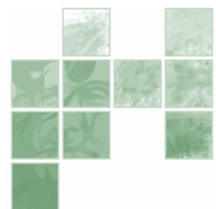
- Manual construction of lexicon labor-intensive
- Many medical terms exhibit high degree of similarity across languages (“cognates”):  
*surgical*, *chirurgisch*, *chirurgique*, *cirúrgico*, *quirurgico*, *kirurgisk*
- Others don’t (“non-cognate translations”):  
*spleen*, *Milz*, *rate*, *baço*, *bazo*, *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:

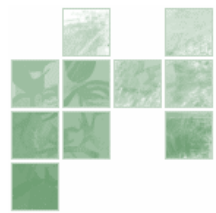
1. *Generation and Validation* of trusted subword cognates for the target languages
2. *Bootstrapping*: iterative learning of non cognate subword translations



# 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):



# Generating Cognate Candidates

List of Portuguese  
Subwords  
(14,004 stems):

...  
**estomag-**  
**mulher-**  
...

Application of 44  
string  
substitution  
rules

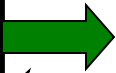
Rule (Port. » Span.)
qua » cua
eia » ena
ss » s
lh » j
lh » ll
l » ll
f » h
...



# Generating Cognate Candidates

List of Portuguese  
Subwords  
(14,004 stems):

...  
**mulher-**  
...

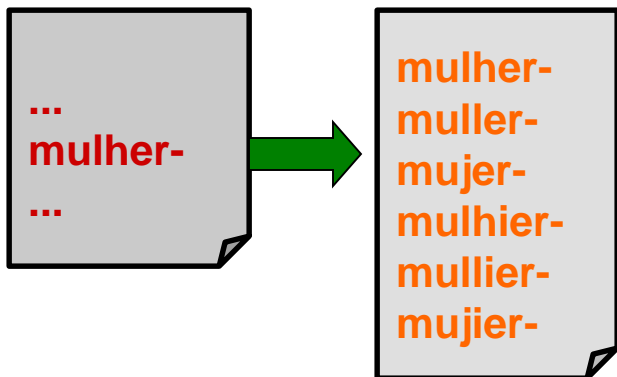


**mulher-**  
**muller-**  
**mujer-**  
**mulhier-**  
**mullier-**  
**mujier-**

Application of 44  
string  
substitution  
rules

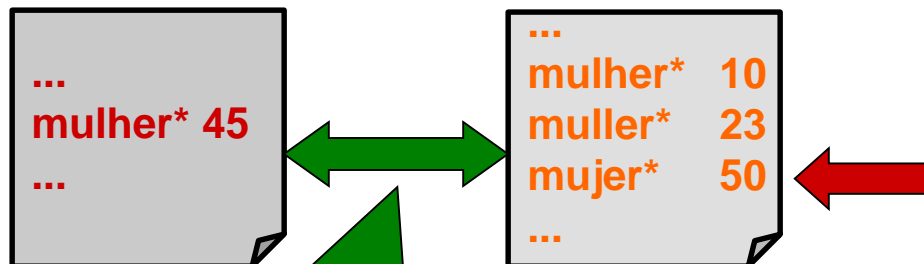
# Generating Cognate Candidates

List of Portuguese  
Subwords  
(14,004 stems):



Word frequency lists  
derived from unrelated corpora:

Size (Portuguese) ~ Size(Spanish)

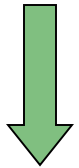


Comparison between word frequency lists:  
Choose that cognate alternative with the *most similar* corpus frequency



# Semantic Mapping

mulher-  mujer-

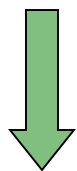


**MID:** #female = { woman, women, female-, frau-, weib-, mulher-, mujer- }



# Semantic Mapping

mulher → mujer



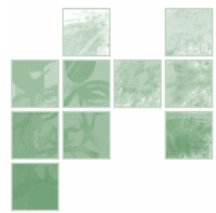
**MID:** #female = { woman, women, female-, frau-, weib-, mulher-, **mujer-**}

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



# 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  <b>al</b>	<b>#abdom</b>
wall	<b>#wall</b>
proced ure	#operat

*„Cirugia de la pared abdominal“:*

cirugia	
pared	<b>#wall</b>
abdomin  <b>al</b>	<b>#abdom</b>



# 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“:

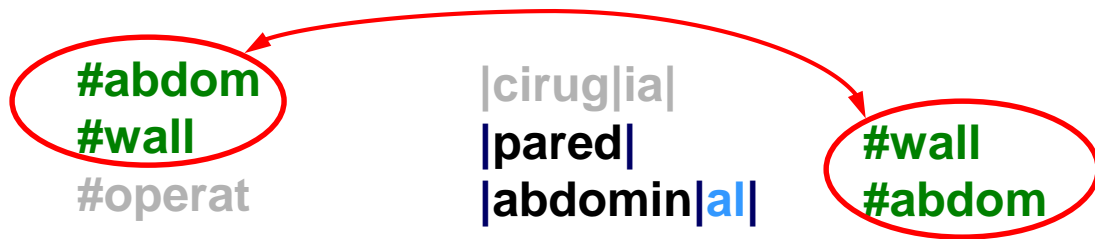
|abdomin|al|  
|wall|  
|proced|ure|

#abdom  
#wall  
#operat

„Cirugia de la pared abdominal“:

|cirugia|  
|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|al|  
|wall|  
|proced|ure|

#abdom  
#wall  
#operat

„Cirugia de la pared abdominal“:

|cirugia|  
 pared|  
 abdomin|al|

#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|al|  
|wall|  
|proced|ure|

#abdom  
#wall  
#operat

|cirugia|  
 |pared|  
 |abdomin|al|

#wall  
#abdom





# Step 2: Bootstrapping

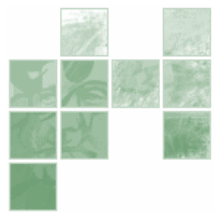
- Acquisition of non-cognates uses:
  - validated cognate seed lexicons
  - parallel corpora

*„Abdominal wall procedure“:*

abdomin al	#abdom
wall	#wall
proced ure	#operat

*„Cirugia de la pared abdominal“:*

cirug ia	
pared	#wall
abdomin al	#abdom



# Bootstrapping Algorithm

 For every UMLS term pair do

 „Abdominal wall procedure“:

abdomin al	#abdom
wall	#wall
proced ure	#operat

„Cirugia de la pared abdominal“:

cirug ia	
pared	#wall
abdomin al	#abdom



# Bootstrapping Algorithm

For every UMLS term pair do

 If there is exactly one invalid segmentation in target language

*„Abdominal wall procedure“:*

abdomin al	#abdom
wall	#wall
proced ure	#operat



*„Cirugia de la pared abdominal“:*

cirug ia	
pared	#wall
abdomin al	#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|al|  
|wall|  
|proced|ure|

#abdom  
#wall  
#operat



*„Cirugia de la pared abdominal“:*

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

#wall  
#abdom







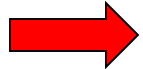
# Bootstrapping Algorithm

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“:*

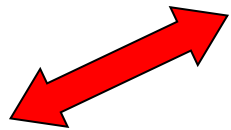
|abdomin|al|  
|wall|  
|proced|ure|

#abdom  
#wall  
#operat

*„Cirugia de la pared abdominal“:*

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

#wall  
#abdom





# Bootstrapping Algorithm

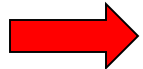
For every UMLS term pair do

    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 al	#abdom
wall	#wall
proced ure	#operat

*„Cirugia de la pared abdominal“:*

cirug ia		cirug ia
pared		#wall
abdomin al		#abdom



# Bootstrapping Algorithm

For every UMLS term pair do

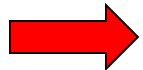
If there is exactly one invalid segmentation in target language

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 al	#abdom
wall	#wall
proced ure	#operat

„Cirugia de la pared abdominal“:

cirug ia	→	cirug ia
pared		#wall
abdomin al		#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|           #derma  
|operat|ions|   #operat

„Cirugia de la pared abdominal“:

„Cirugia de piel“:

|cirug|ia|       #operat  
|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“:

„Cirugia de la pared abdominal“:

„Cirugia de piel“:

|skin|           #derma  
|operat|ions|   #operat

|cirug|ia|           #operat  
|piel|           →   |piel|

#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“:

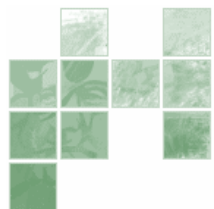
„Cirugia de la pared abdominal“:

„Cirugia de piel“:

„Malformacion de la piel“:

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

|malformation|  
|piel|                    #derma



# 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

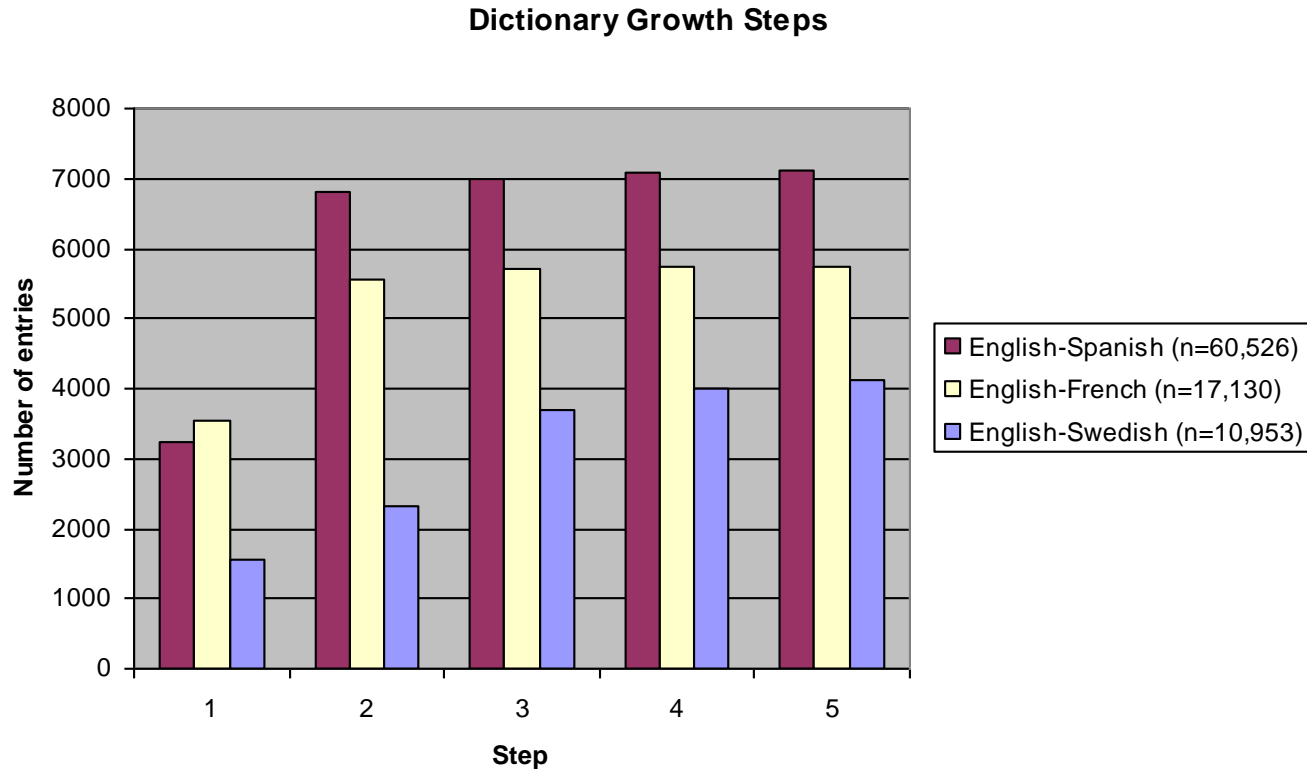
|malformation|     |malform|ation|  
|piel|                    #derma



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



# Bootstrapping Results



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

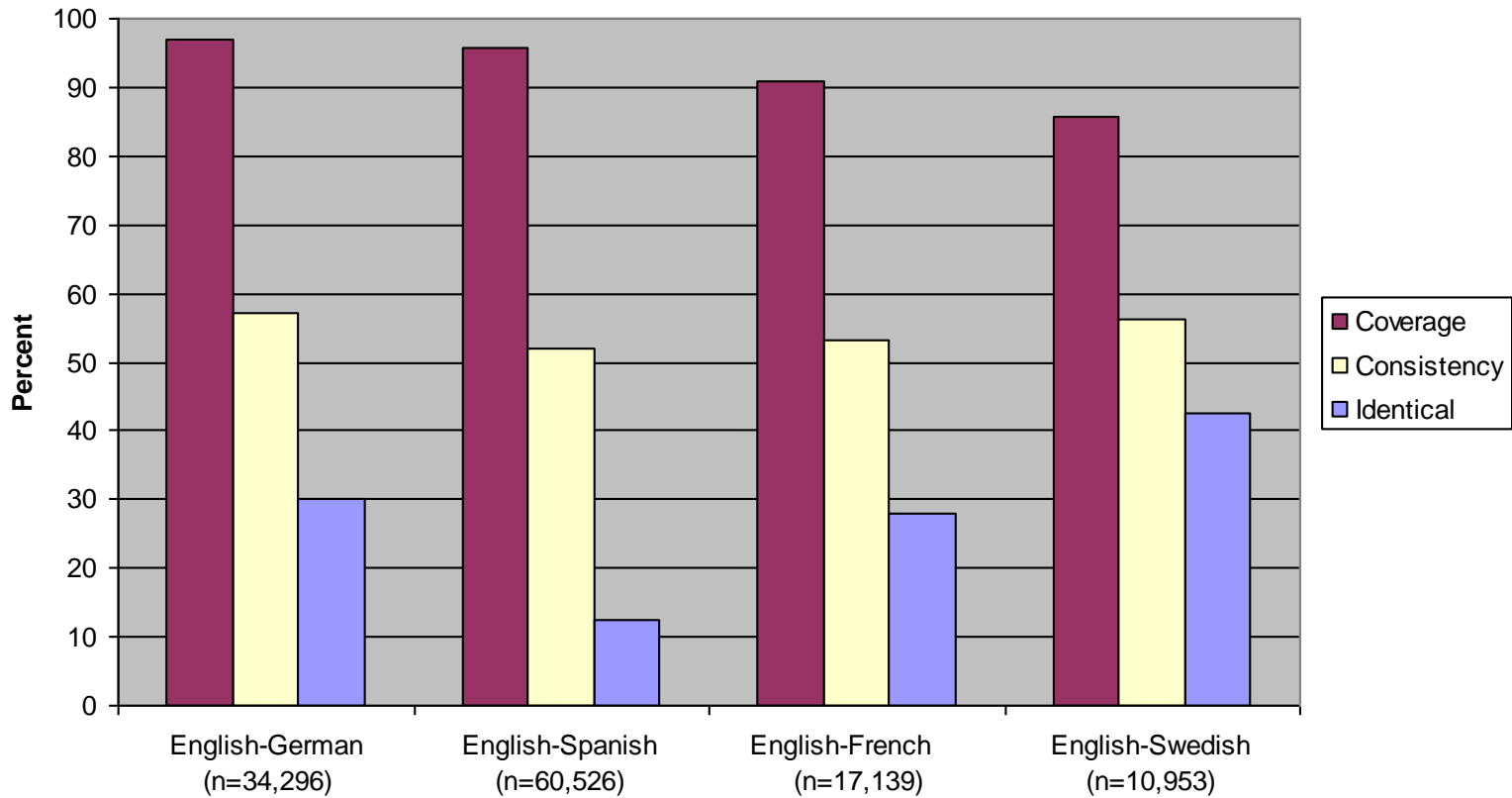


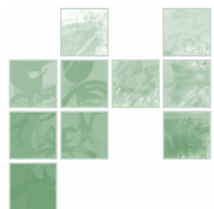
# Evaluation

- 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

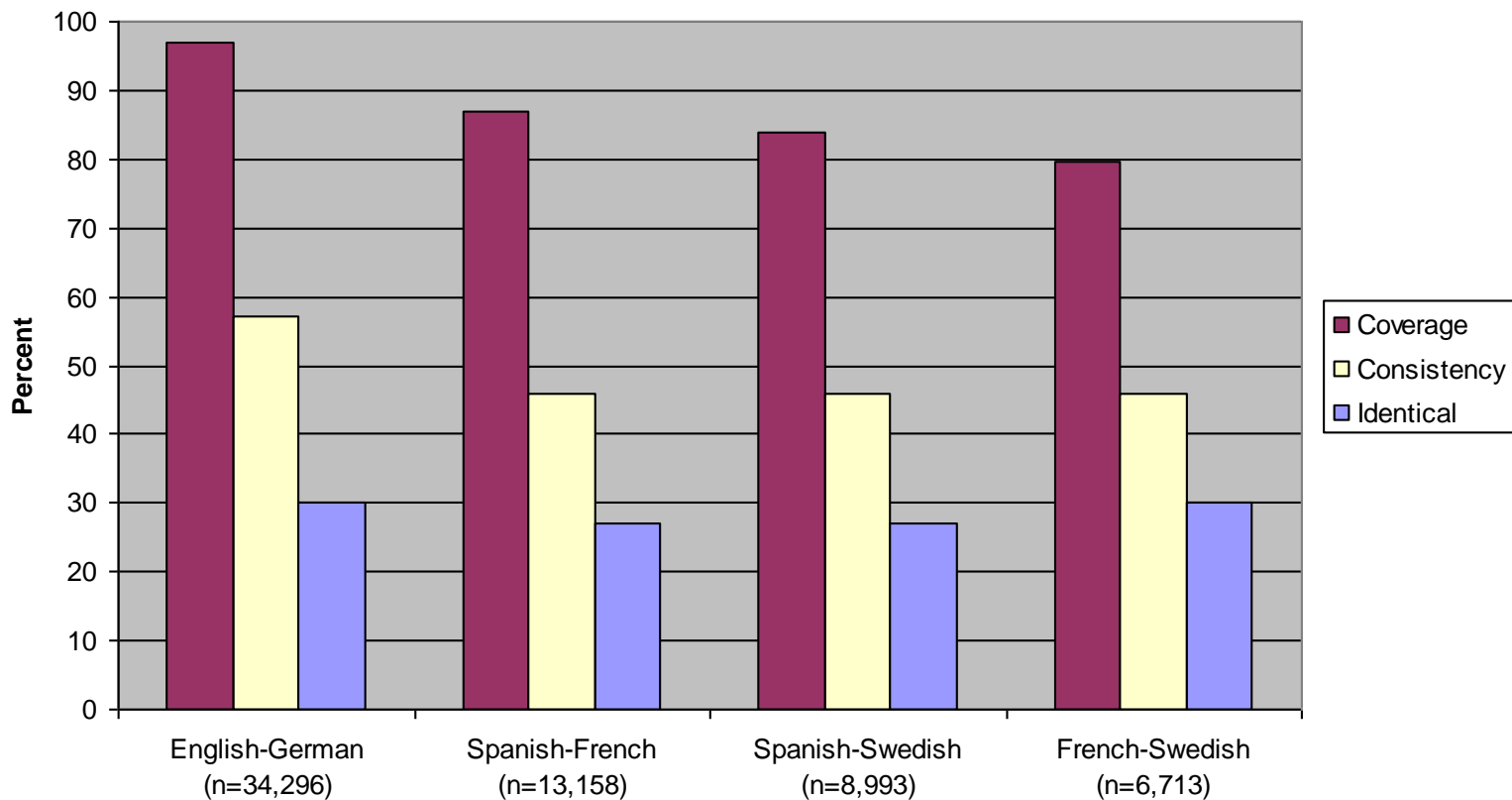


# Results





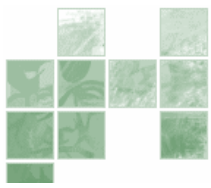
# Results





# 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

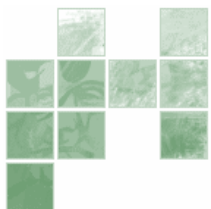
Kornél Markó

Stefan Schulz

Udo Hahn

Medical Informatics,  
Freiburg University Hospital  
(Germany)

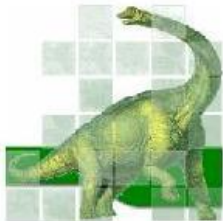
Jena University, Language &  
Information Engineering  
(Germany)



[www.MorphoSaurus.net](http://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](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%3fid=32191.html](http://supreme.coling.uni-jena.de/~coling/search_engine_docs/original/aerzteblatt/artikeldruck.asp%3fid=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](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

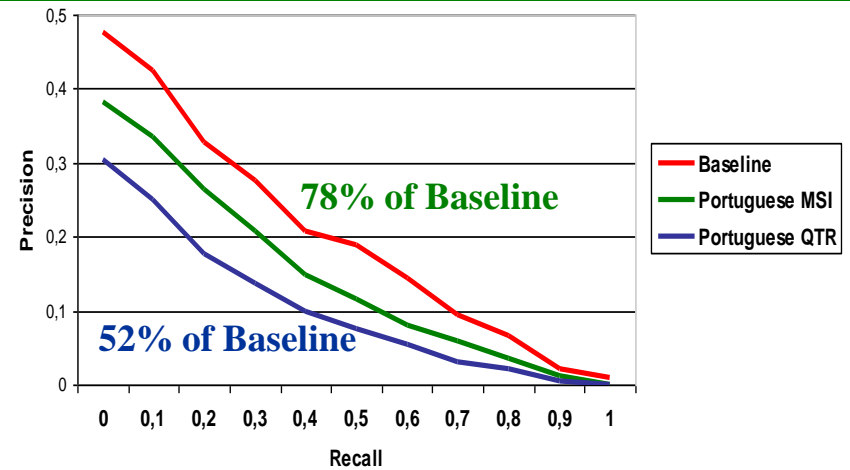
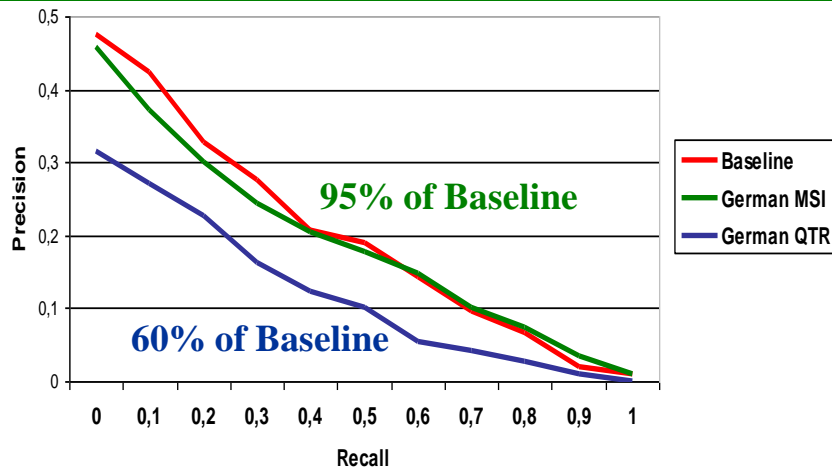
- **Baseline:** monolingual text retrieval
  - (stemmed) English user queries
  - (stemmed) English texts
- Query translation (QTR)
  - Google translator
  - Multilingual dictionary compiled from UMLS
- **MorphoSaurus Indexing (MSI)**
  - Interlingual representation of both user queries and documents

# Evaluation Results

German (n = 22,385)

Top 200

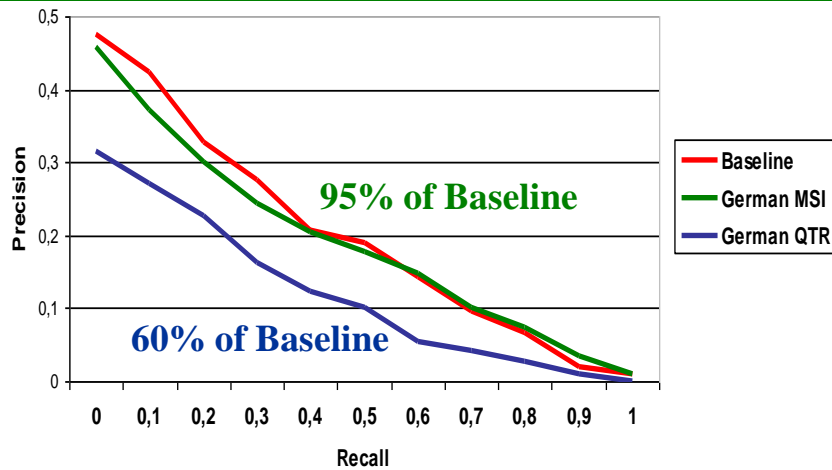
Portuguese (n = 14,862)



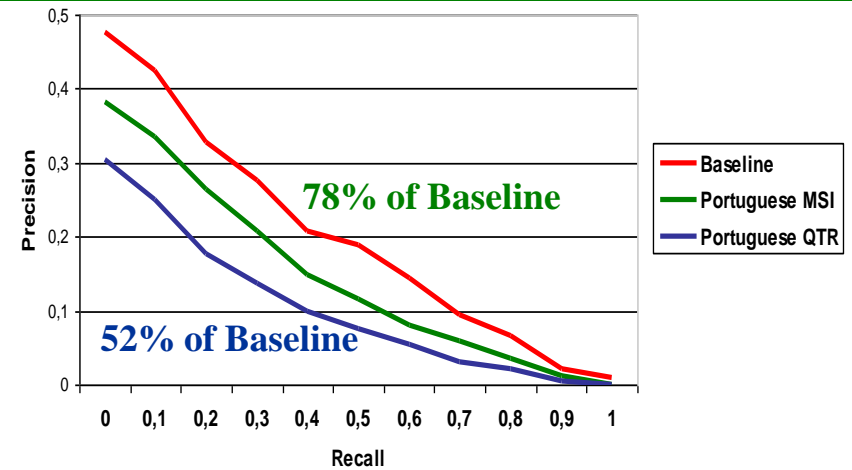
# Evaluation Results

German (n = 22,385)

Top 200

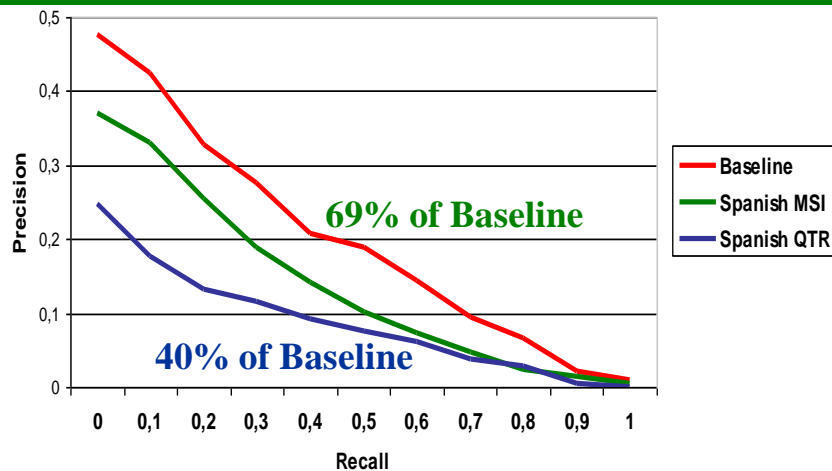


Portuguese (n = 14,862)

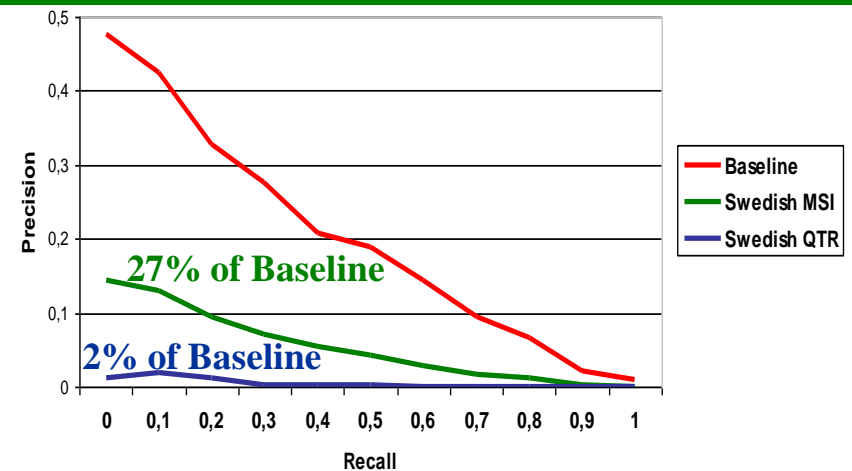


Spanish (n = 7,154)

Top 200



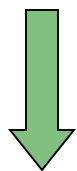
Swedish (n = 4,148)





# 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