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

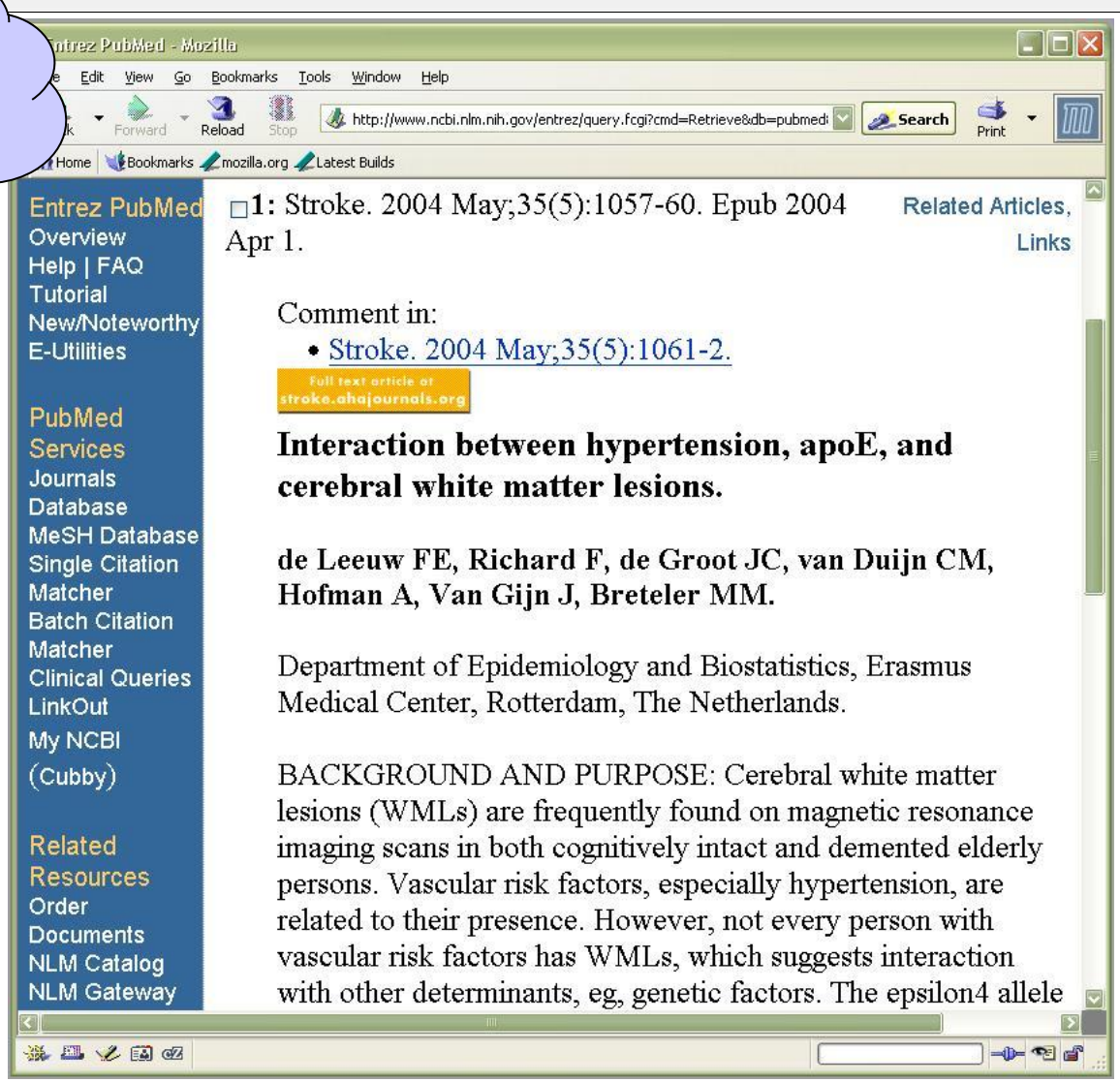
Kornél Markó, Stefan Schulz, Udo Hahn

Freiburg University Hospital, Medical Informatics Department, Germany
Jena University, Language & Information Engineering (JULIE), Germany

UNIVERSITÄTS FREIBURG KLINIKUM

Multilingual Textretrieval

„Korrelation von Hypertonie und Läsion der Weißen Substanz“

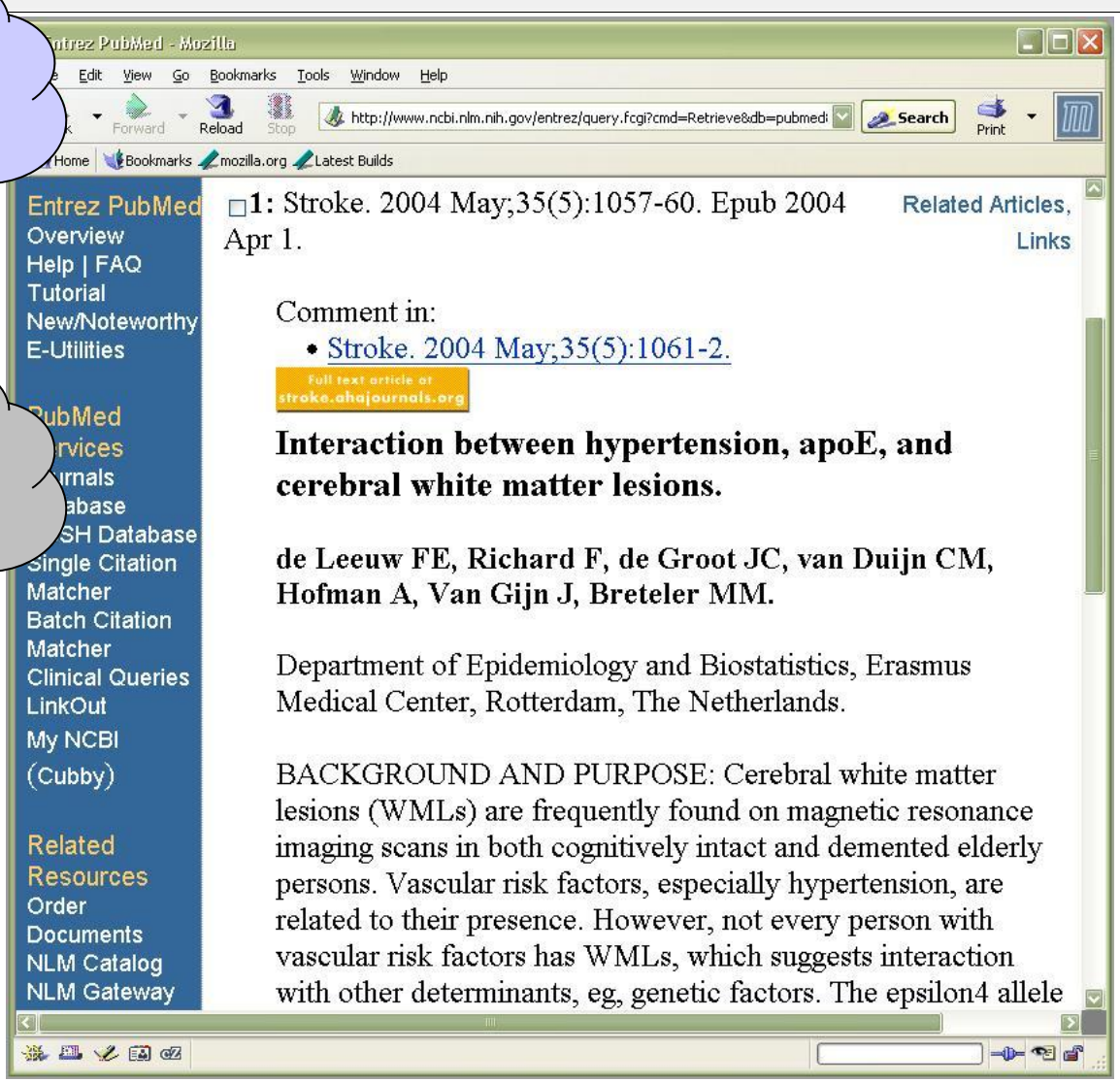


Multilingual Textretrieval

„Korrelation von Hypertonie und Läsion der Weißen Substanz“

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

Search Engine

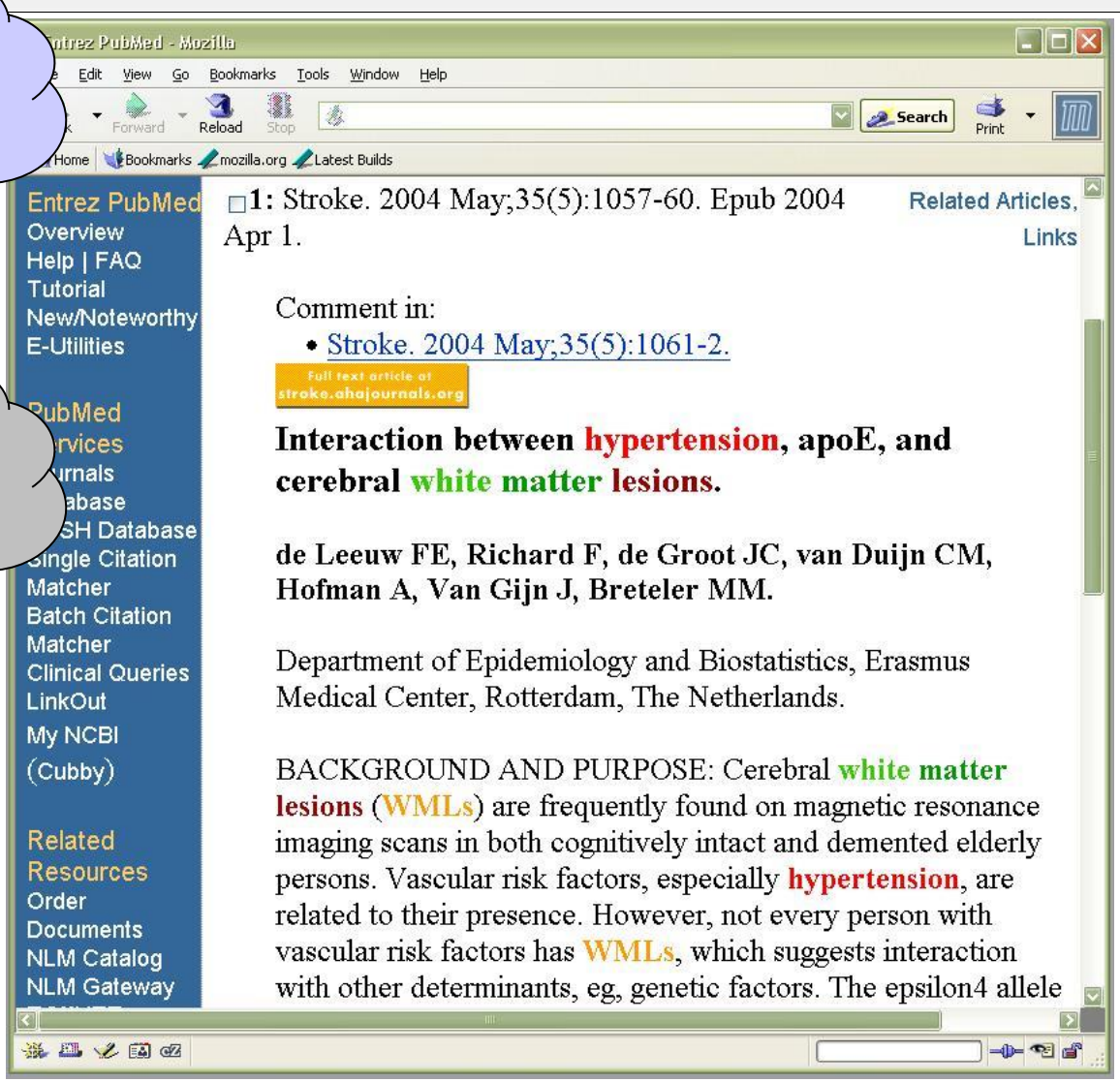


Multilingual Textretrieval

„Korrelation von Hypertonie und Läsion der Weißen Substanz“

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

Search Engine

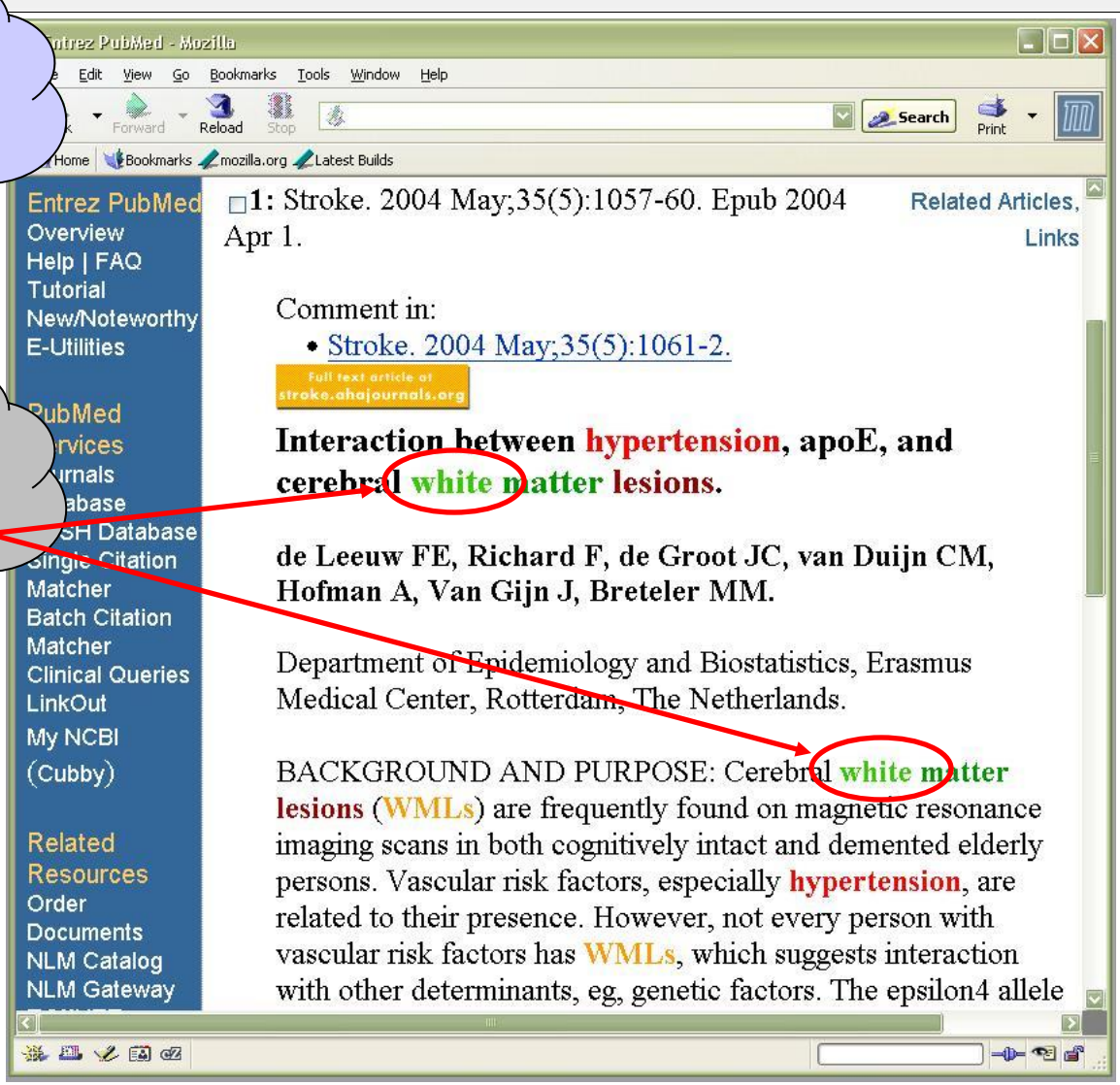


Multilingual Textretrieval

„Korrelation von Hypertonie und Läsion der Weißen Substanz“

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

Search Engine



Linguistic Phenomena

- Morphological processes:
  - *Inflection: leukocyt**e** <> leukozyte**s**,*
    *appendi**x** <> appendi**ces***
  - *Derivation: leukocyt**e** <> leukocyt**ic***
  - *Composition: leuk|em|ia, para|sympath|ectomy,*
    Magen|schleim|haut|entzünd|ung
- Synonymy:
  - *ascorbic acid <> vitamin C, hemorrhage <> bleeding*
- *Spelling variants:*
  - ***oe**sophagus <> **e**sophagus,*
  - ***K**ar**z**inom <> **C**ar**c**inom <> **C**ar**z**inom (carcinoma)*

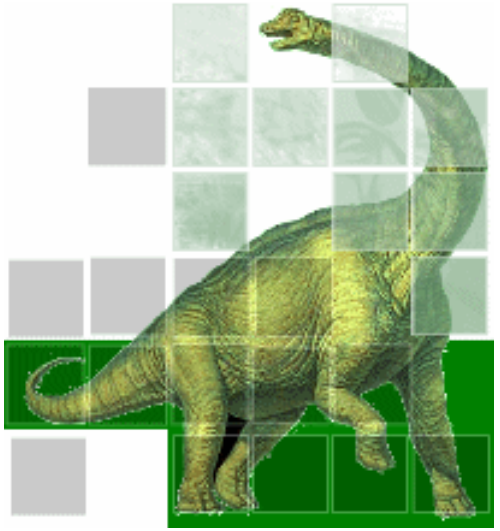
Subword Approach

- Subwords are atomic, conceptual *or* linguistic units:
  - Stems: *stomach*, *gastr*, *diaphys*
  - Prefixes: *anti-, bi-, hyper-*
  - Suffixes: *-ary, -ion, -itis*
  - Infixes: *-o-, -s-*
- Equivalence classes contain synonymous subwords and their translations in a thesaurus:

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

Morphosaurus

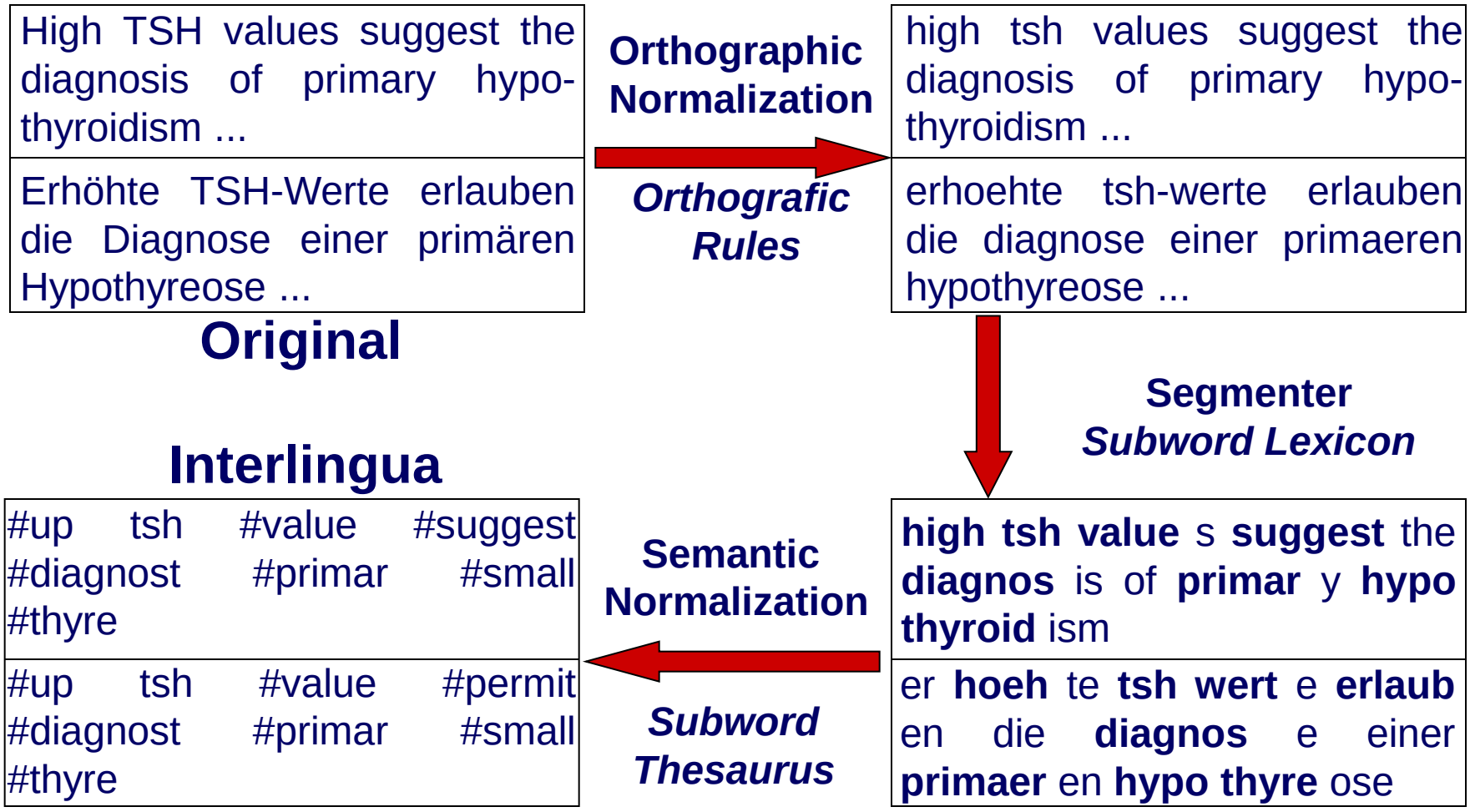
- Subword-Lexicon:
  - Organizes subwords in several languages (English, German, Portuguese)
- Subword-Thesaurus:
  - Groups synonymous subwords (within and between languages)
- Subword-Segmenter:
  - Extraction of Subwords and Assignment of *Equivalence Classes*



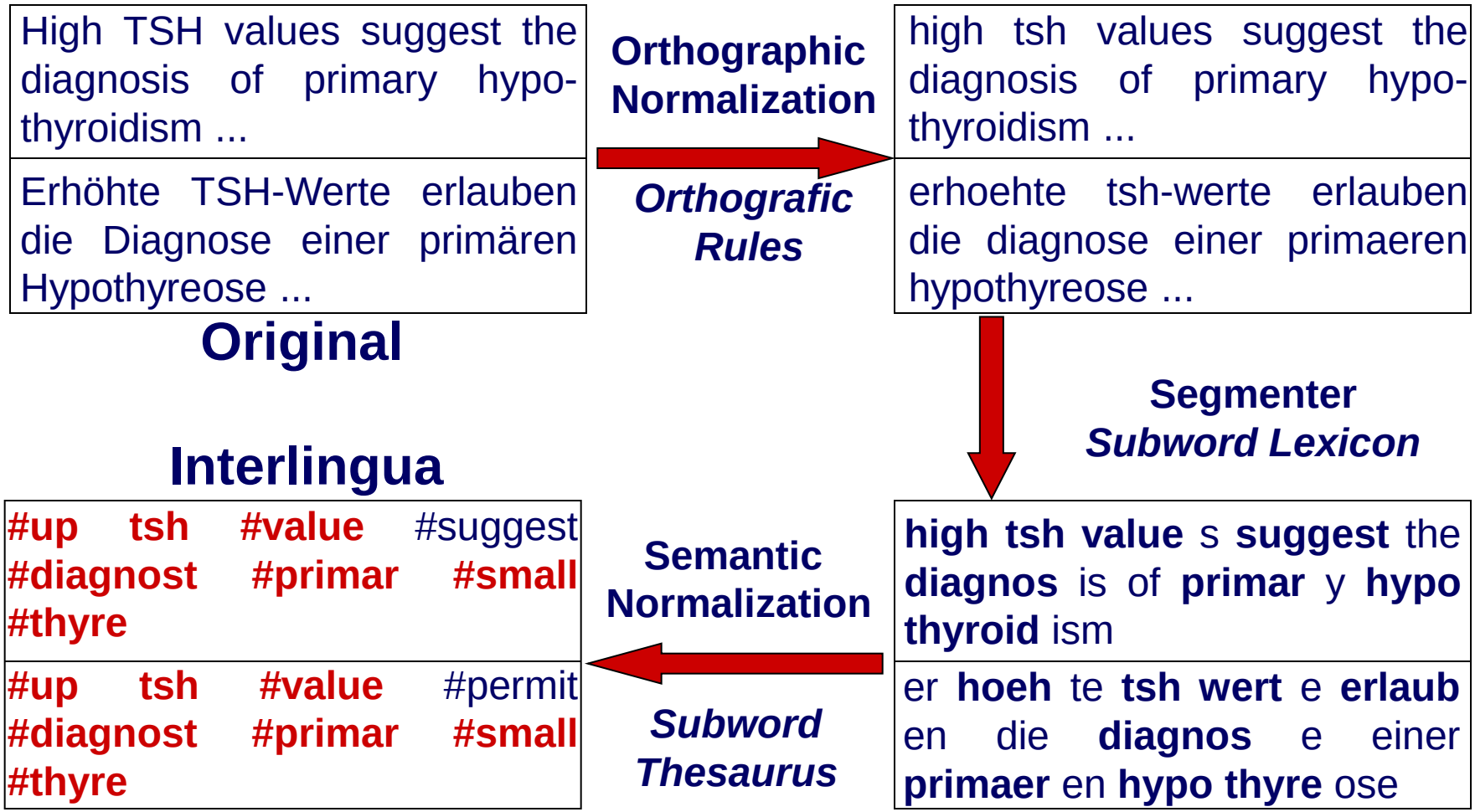
Morphosaurus

Morphosaurus-Identifier (MID)

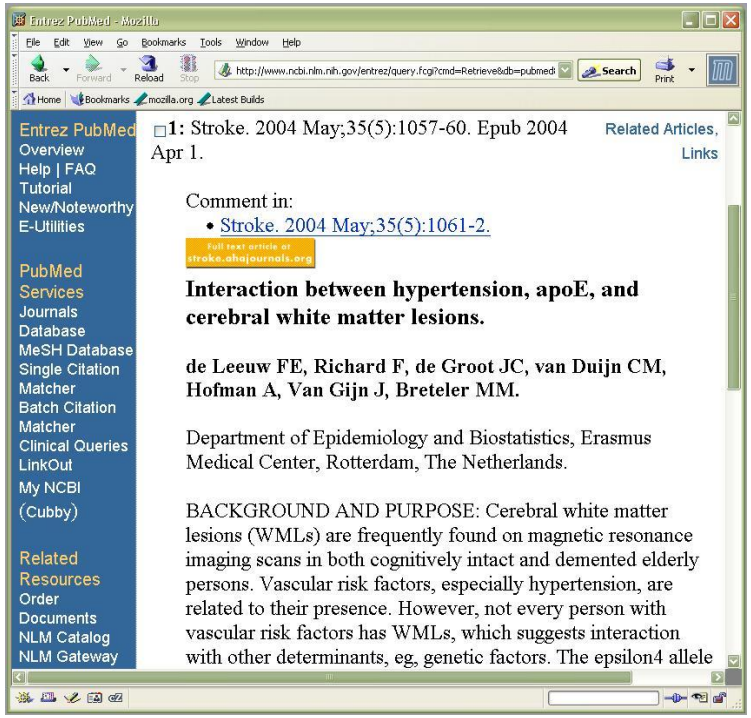
Example


High TSH values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose ...
Original
Orthographic Normalization
Orthografic Rules
high tsh values suggest the diagnosis of primary hypo-thyroidism ...
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
Semantic Normalization
Subword Thesaurus
Interlingua
#up tsh #value #suggest #diagnost #primar #small #thyre
#up tsh #value #permit #diagnost #primar #small #thyre

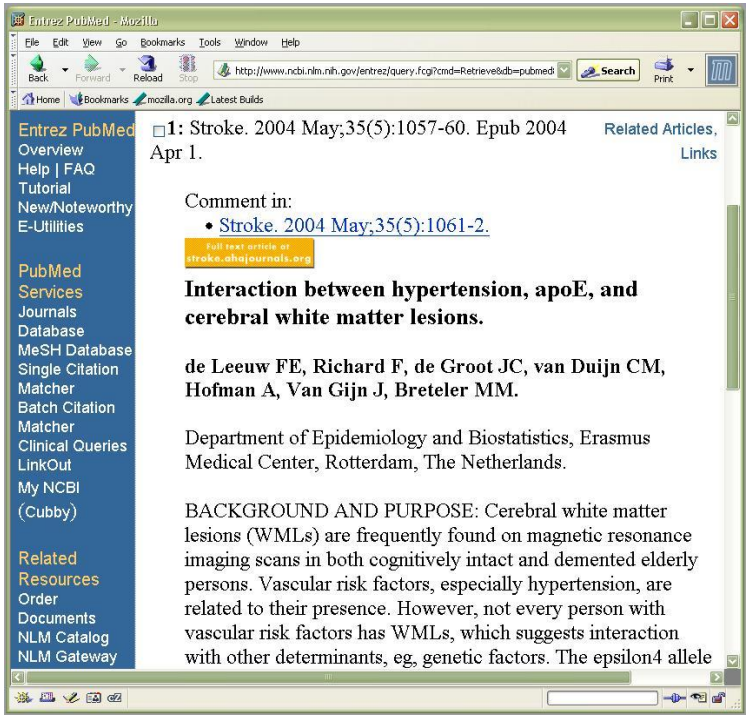
Example

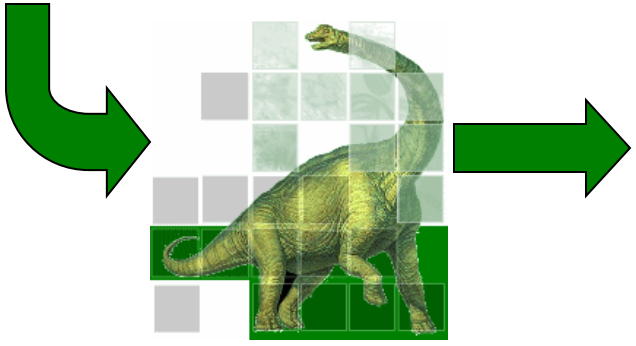

High TSH values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose ...
Original
Orthographic Normalization
Orthografic Rules
high tsh values suggest the diagnosis of primary hypo-thyroidism ...
erhoehte tsh-werte erlauben die diagnose 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
Semantic Normalization
Subword Thesaurus
Interlingua
#up tsh #value #suggest #diagnost #primar #small #thyre
#up tsh #value #permit #diagnost #primar #small #thyre

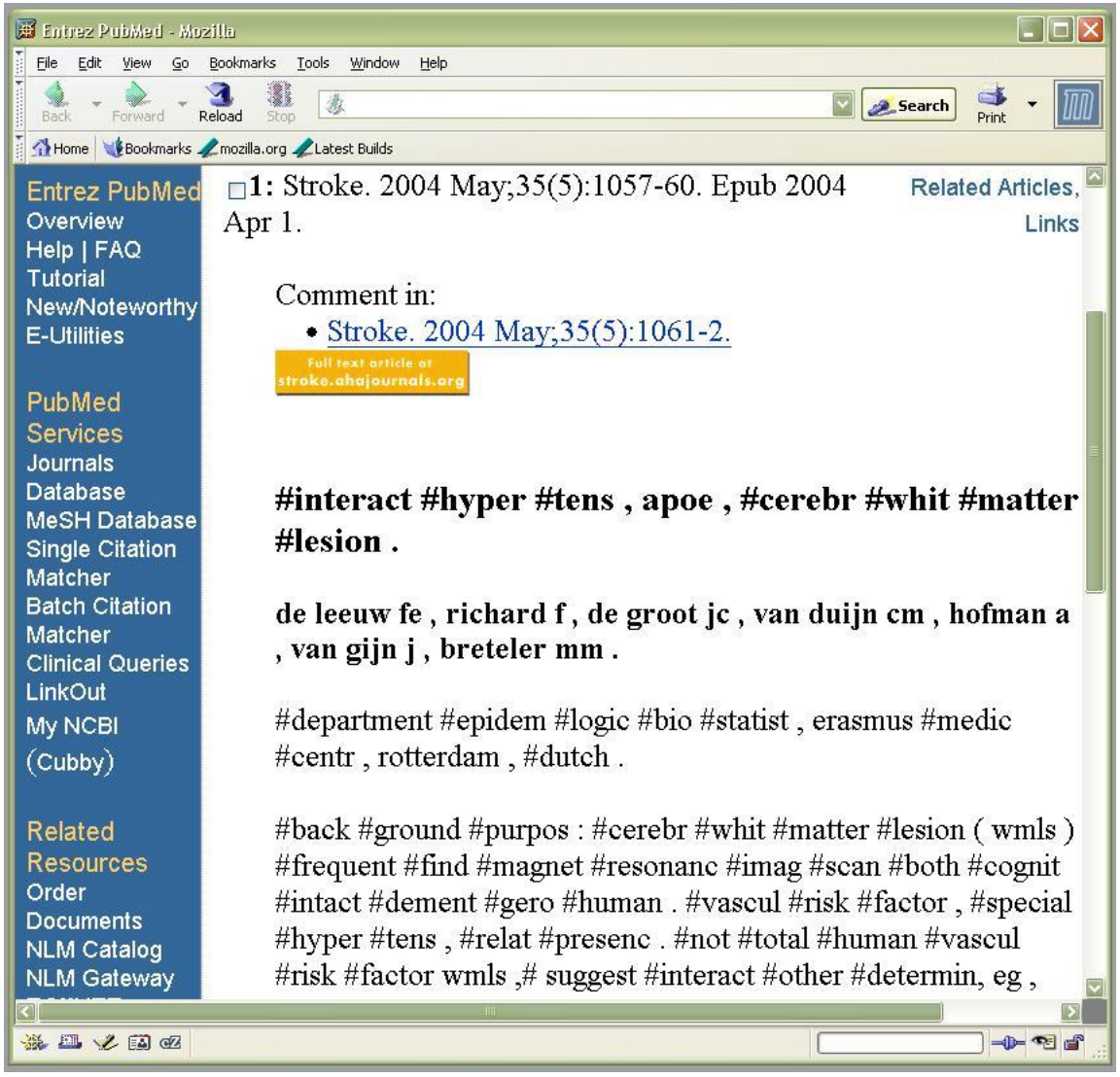
Morphosaurus Search



Morphosaurus Search

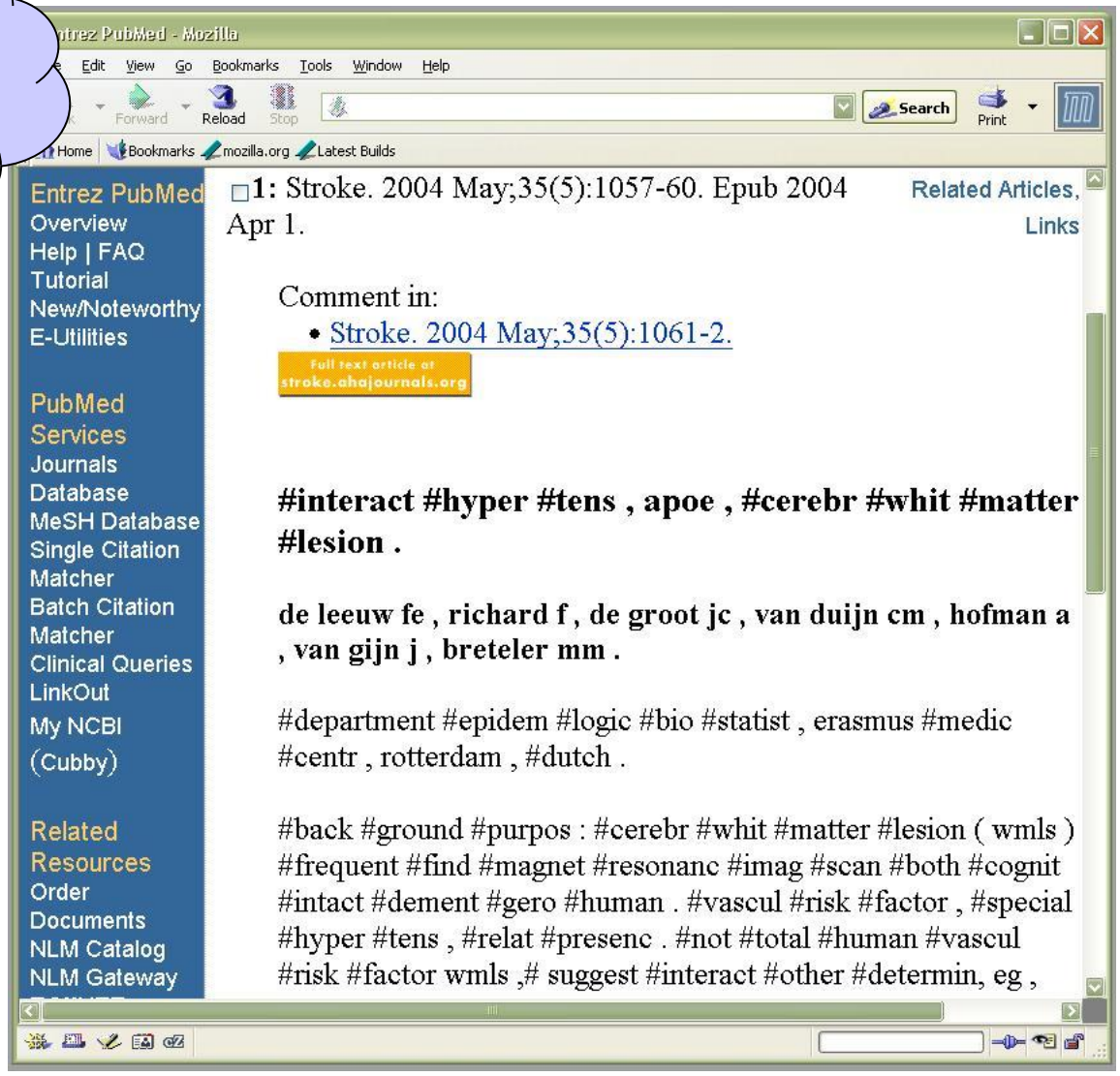






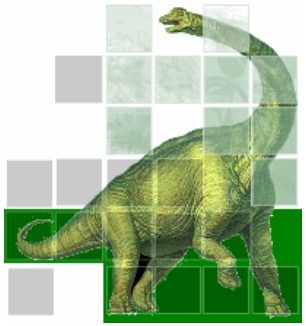
Morphosaurus Search

„Korrelation von Hypertonie und Läsion der Weißen Substanz“

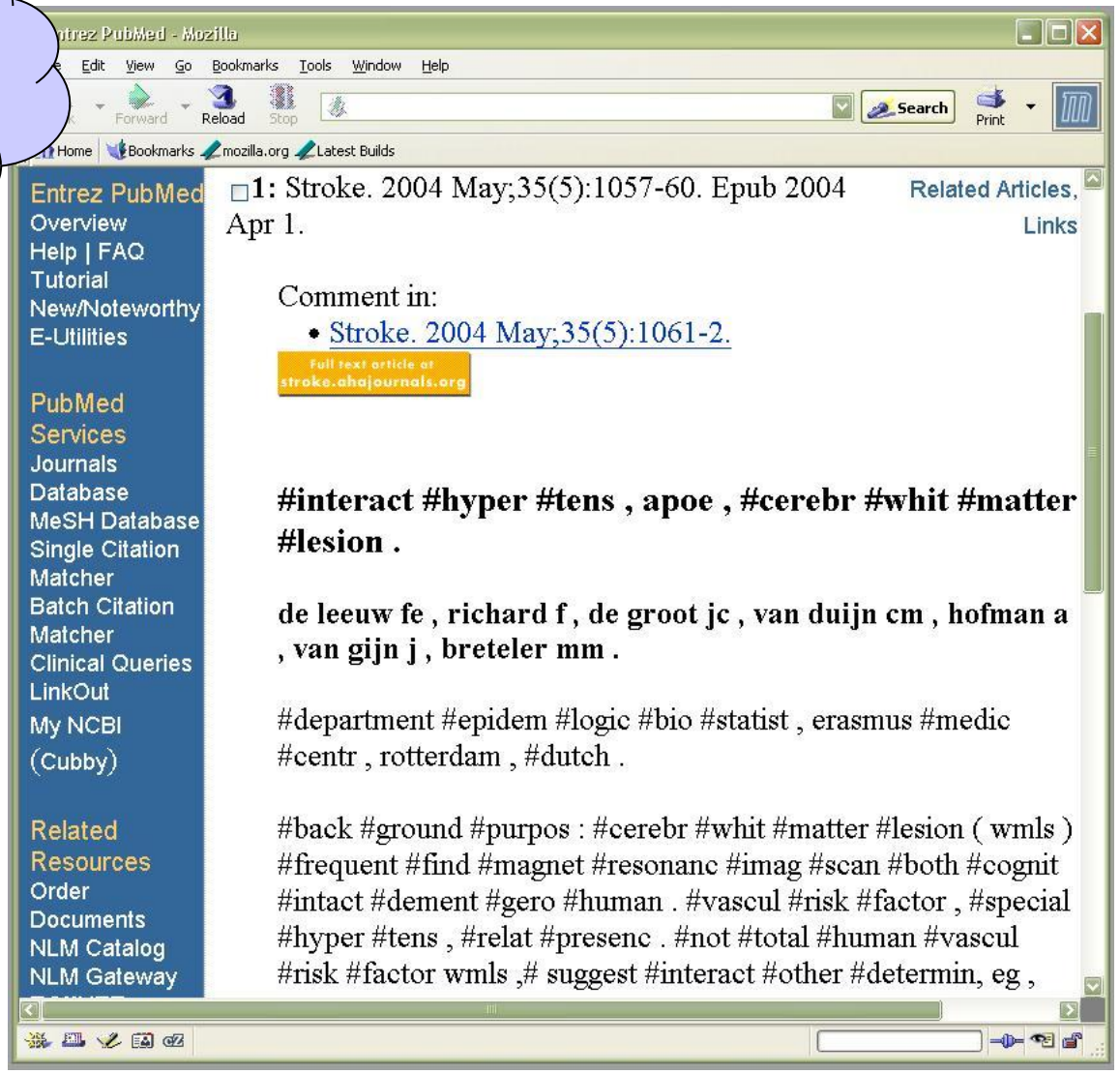


Morphosaurus Search

„Korrelation von Hypertonie und Läsion der Weißen Substanz“

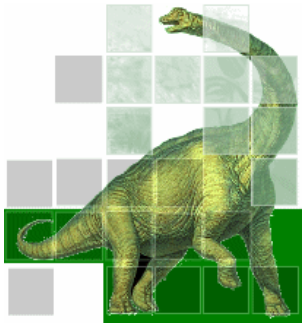


„#correl #hyper #tens #lesion #whit #matter“



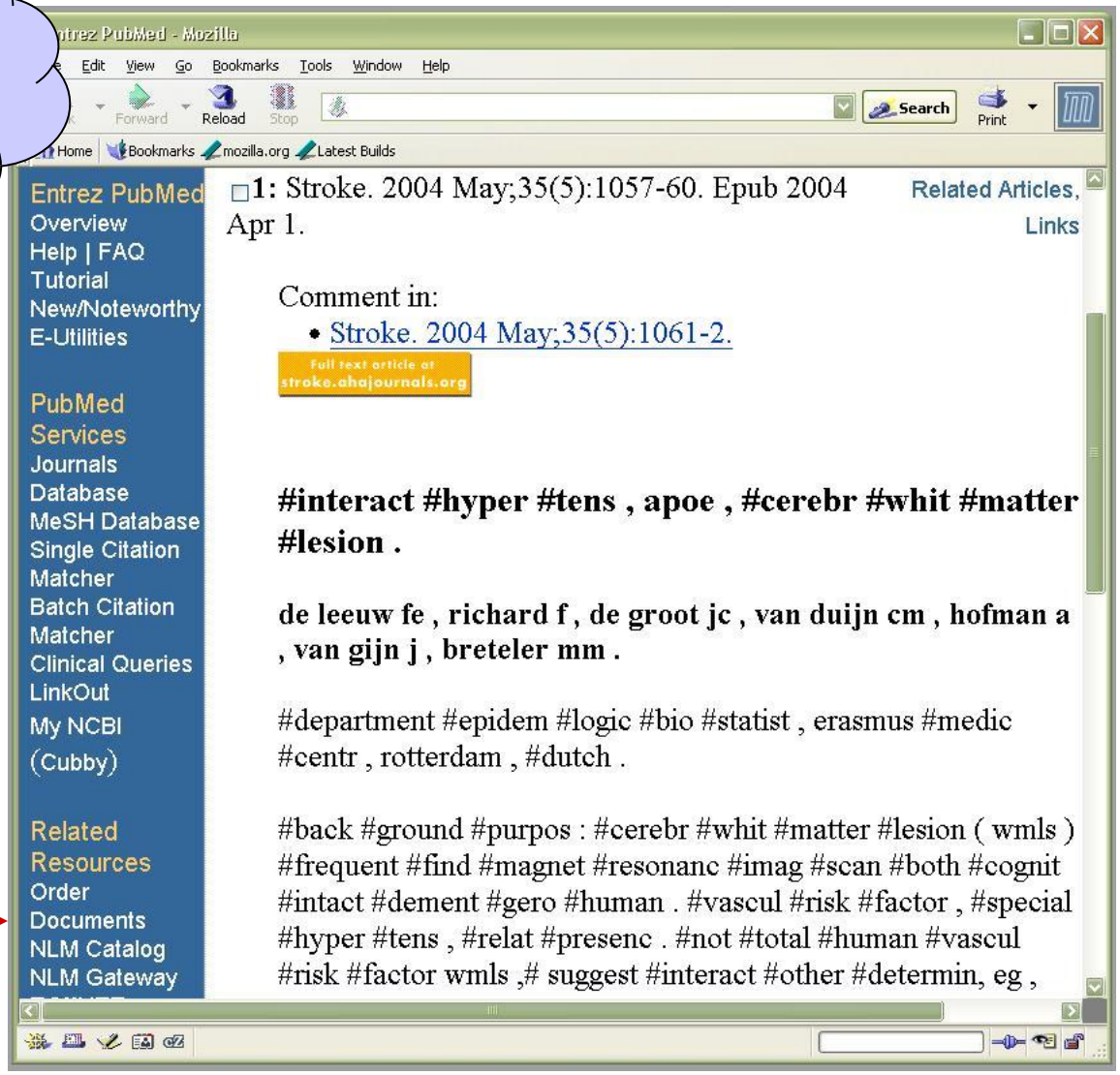
Morphosaurus Search

„Korrelation von Hypertonie und Läsion der Weißen Substanz“



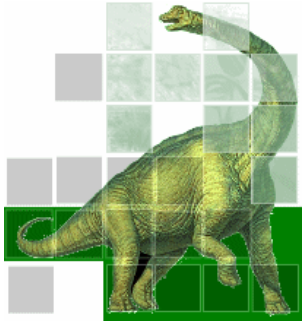
„#correl #hyper #tens #lesion #whit #matter“

Search Engine



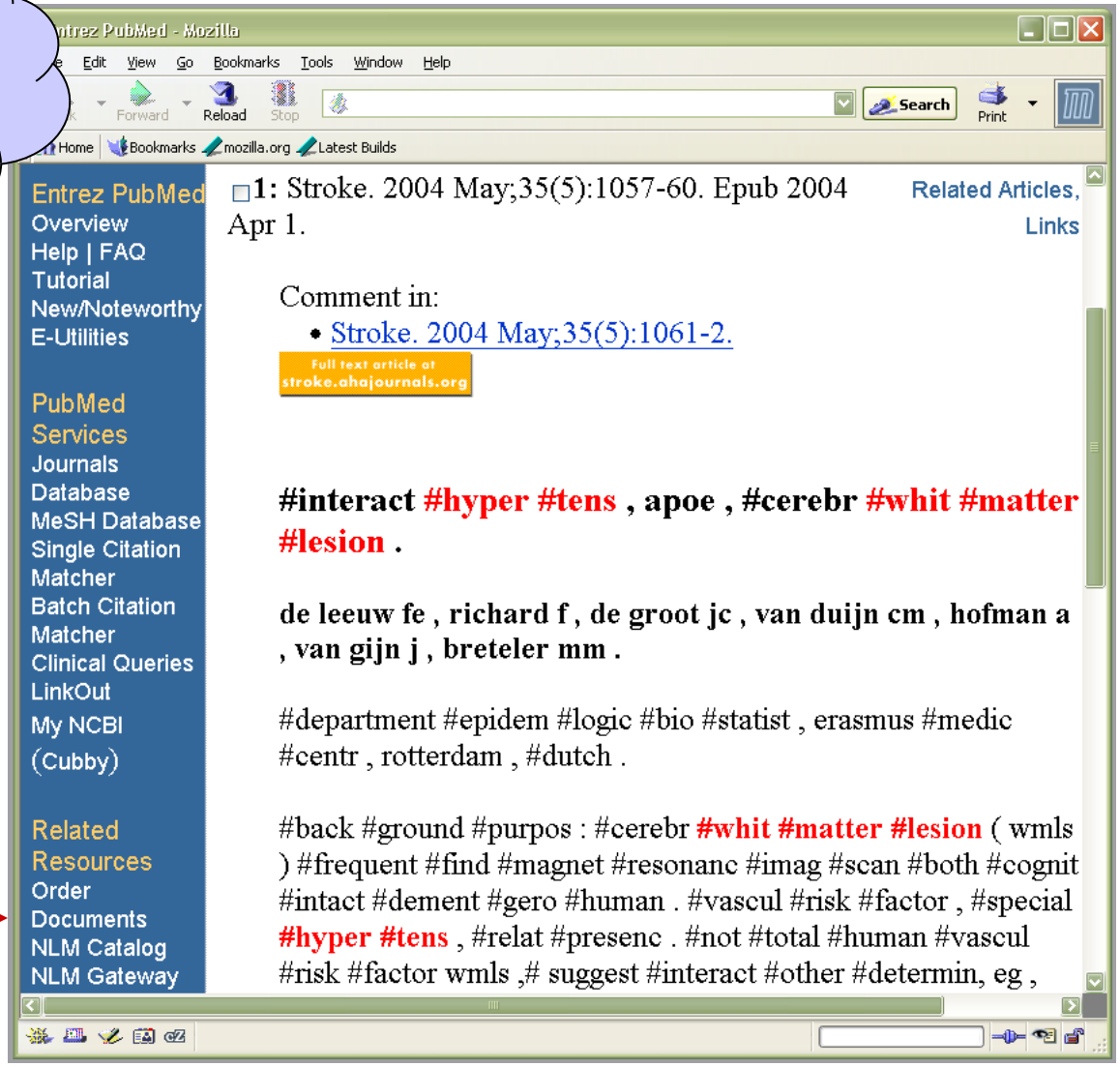
Morphosaurus Search

„Korrelation von Hypertonie und Läsion der Weißen Substanz“



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

Search Engine



Automatic Language Acquisition

- Automatic Acquisition of ***Spanish*** and ***Swedish*** subword lexicons

- Step 1: Generation of *cognate* seed lexicons:
  - Automatic generation of cognate subword candidates
    - Spanish cognates from Portuguese
    - Swedish cognates from English and German
  - Selection of subword candidates
  - Semantic mapping (linkage to equivalence classes)
  - Validation of semantic mappings

- Step 2: Use cognate lexicons as a seed for iteratively learning *non-cognates*

Step 1: Cognate Acquisition

- **Resources** for cognate acquisition for

  Spanish (from Portuguese) and

  Swedish (from English and German):

  - Portuguese (~14,000 stems), German and English (~22,000 stems each) subword lexicons
  - Medical corpora for Portuguese, German, English, Spanish and Swedish acquired from the Web
  - Word frequency lists generated from these corpora
  - Manually created list of Spanish and Swedish affixes

Generating Cognate Candidates

List of Portuguese Subwords (14,004 stems):

Generating Cognate Candidates

List of Portuguese Subwords (14,004 stems):

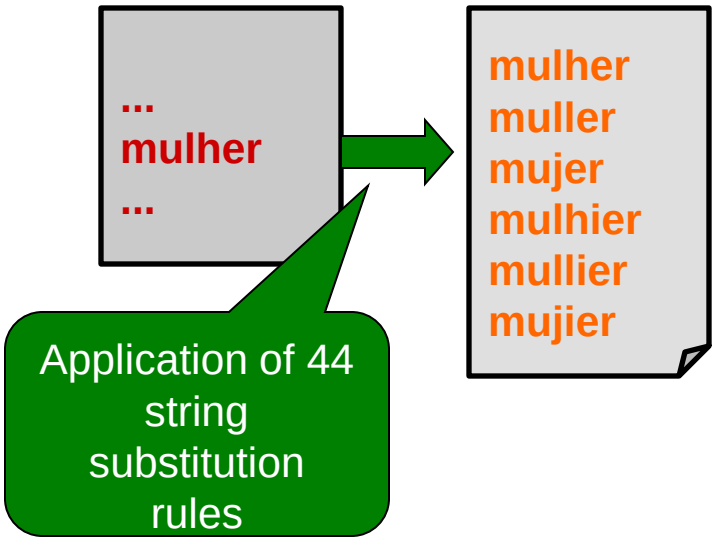
...
estomag
mulher
...

Application of 44 string substitution rules

Rule (Port. » Span.)	Portuguese example	Spanish example	English equivalent
qua » cua	**quadr**	**cuadr**	*frame*
eia » ena	**veia**	**vena**	*vein*
ss » s	**fracass**	**fracas**	*fail*
lh » j	**mulher**	**mujer**	*woman*
l » ll	**lev**	**llev**	*take*
i » y	**ensai**	**ensay**	*trial*
f » h	**formig**	**hormig**	*ant*
...	**...**	**...**	...

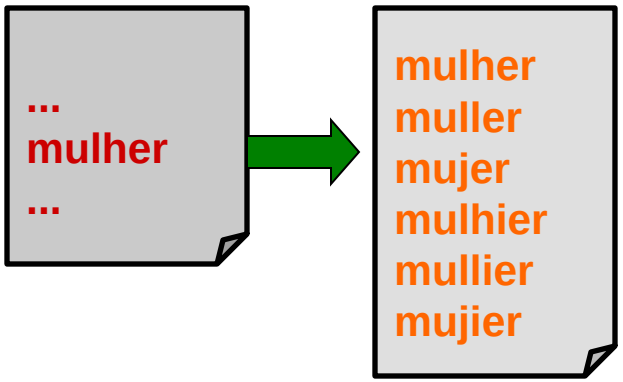
Generating Cognate Candidates

List of Portuguese Subwords (14,004 stems):

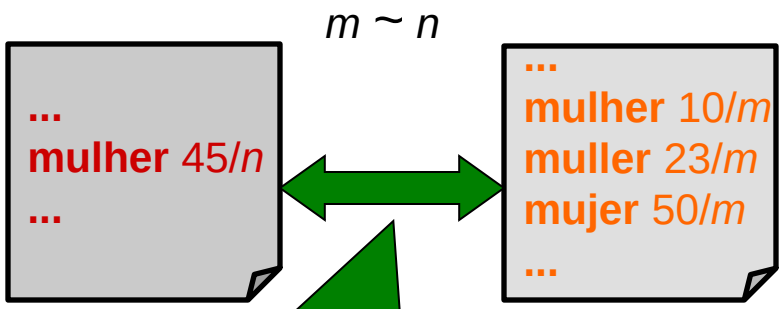


Selecting Cognate Candidates

List of Portuguese Subwords (14,004 stems):



Word frequency lists derived from unrelated corpora:
Portuguese (size = *n*) Spanish (size = *m*)



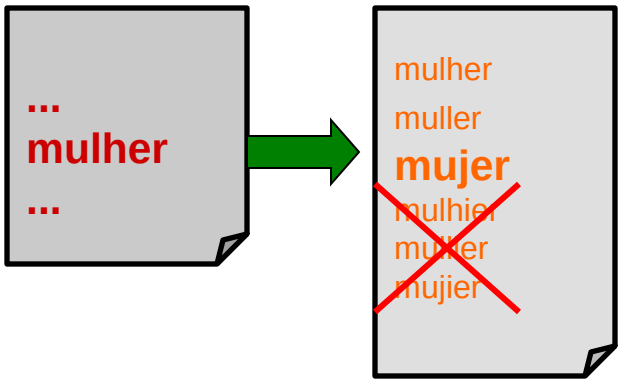
Comparison between word frequency lists:

– Elimination of non-matching subwords

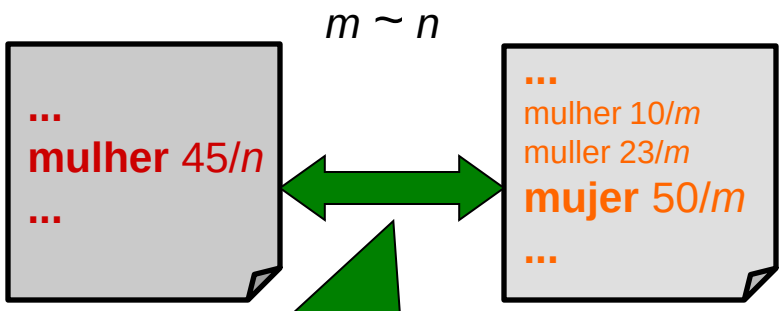
– Choose that cognate alternative with the *most similar* corpus frequency

Selecting Cognate Candidates

List of Portuguese Subwords (14,004 stems):



Word frequency lists derived from unrelated corpora:
Portuguese (size = *n*) Spanish (size = *m*)



Comparison between word frequency lists:

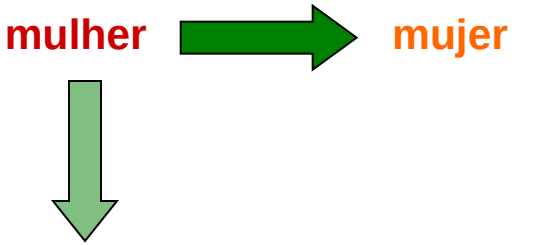
– Elimination of non-matching subwords

– Choose that cognate alternative with the *most similar* corpus frequency

Semantic Mapping

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

Semantic Mapping



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

Language Pair	Source Lexicon	Selected Cognates
Portuguese-Spanish	14,004	**8,644**
German-Swedish English-Swedish	21,705 21,501	**6,086**

Cognate Validation

- Use parallel corpora to identify *false friends*:
  - Portuguese *crianc* (*child*) ↔ Spanish *crianz (breed)*
  - Portuguese *crianc* (*child*) ↔ Spanish *nin* (*child*)
  - Portuguese *criac* (*breed*) ↔ Spanish *crianz* (*breed*)

- UMLS Metathesaurus
  - Contains over 2M medical terms and phrases, aligned in various languages
  - English has the broadest coverage
    - English-Spanish: 60,526 alignments
    - English-Swedish: 10,953 alignmens

- English-Spanish Examples
  - „*Cell Growth*“ ↔ „*Crecimiento Celular*“
  - „*Heart transplantation, with or without recipient cardiectomy*“ ↔ „*Transplante cardiaco, con o sin cardiectomia en el receptor*.“

Cognate Validation

- Use generated cognate seed lexicons to process the UMLS alignments with the Morphosaurus system
- Whenever a MID co-occurs on both sides of an alignment unit, the lexicon entry that led to that particular MID is taken to be valid
- Candidates that never matched this procedure are discarded

Cognate Validation

- Use generated cognate seed lexicons to process the UMLS alignments with the Morphosaurus system
- Whenever a MID co-occurs on both sides of an alignment unit, the lexicon entry that led to that particular MID is taken to be valid
- Candidates that never matched this procedure are discarded

„Abdominal wall procedure“:

|abdomin|al| **#abdom**
|wall| **#wall**
|proced|ure| **#operat**

Cognate Validation

- Use generated cognate seed lexicons to process the UMLS alignments with the Morphosaurus system
- Whenever a MID co-occurs on both sides of an alignment unit, the lexicon entry that led to that particular MID is taken to be valid
- Candidates that never matched this procedure are discarded

„Abdominal wall procedure“:

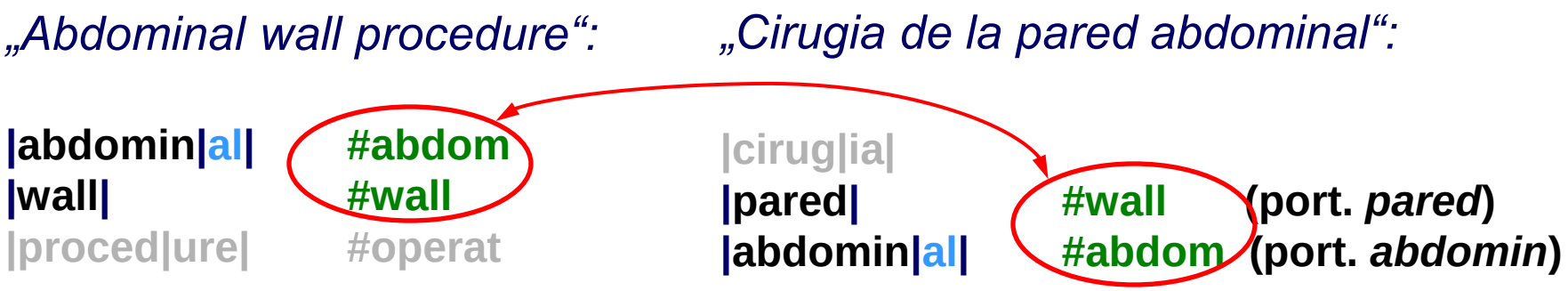
|abdomin|al| **#abdom**
|wall| **#wall**
|proced|ure| **#operat**

„Cirugia de la pared abdominal“:

|cirug|ia|
|pared| **#wall** **(port. *pared*)**
|abdomin|al| **#abdom** **(port. *abdomin*)**

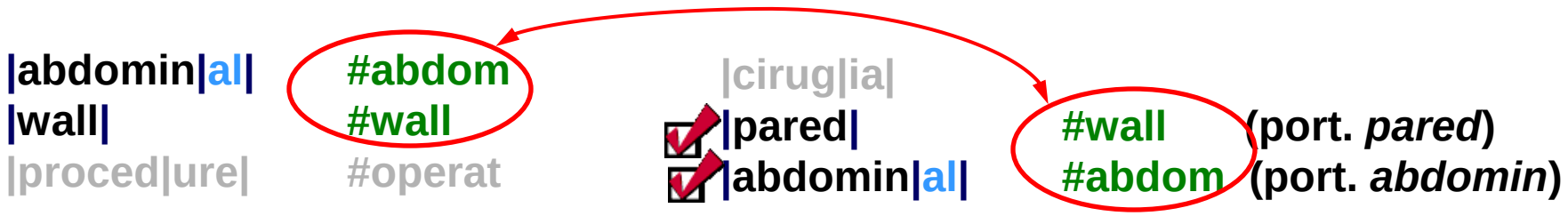
Cognate Validation

- Use generated cognate seed lexicons to process the UMLS alignments with the Morphosaurus system
- Whenever a MID co-occurs on both sides of an alignment unit, the lexicon entry that led to that particular MID is taken to be valid
- Candidates that never matched this procedure are discarded



Cognate Validation

- Use generated cognate seed lexicons to process the UMLS alignments with the Morphosaurus system
- Whenever a MID co-occurs on both sides of an alignment unit, the lexicon entry that led to that particular MID is taken to be valid
- Candidates that never matched this procedure are discarded



Cognate Validation

Language Pair	Hypotheses	Valid
English-**Spanish**	8,644	3,230 (37%)
English-**Swedish**	6,086	1,565 (26%)

„Abdominal wall procedure“:

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

#abdom
#wall
#operat

„Cirugia de la pared abdominal“:

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

#wall (port. ***pared***)
#abdom (port. ***abdomin***)

Step 2: Bootstrapping

- Bootstrapping dictionaries using
  - validated cognate seed lexicons and
  - parallel corpora
  - for acquiring *non-cognates*

„Abdominal wall procedure“:

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

„Cirugia de la pared abdominal“:

|cirug|ia|
|pared| #wall (port. *pared*)
|abdomin|al| #abdom (port. *abdomin*)

Bootstrapping Algorithm

For every alignment in the UMLS 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 alignment in the UMLS do

If there is exactly one invalid segmentation in target 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 alignment in the UMLS 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 alignment in the UMLS 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 alignment in the UMLS 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 alignment in the UMLS 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 alignment in the UMLS 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 alignment in the UMLS do

If there is exactly one invalid segmentation in target language

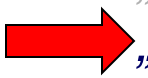
If there is exactly one more MID in source language

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 alignment in the UMLS 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| → **|piel|**

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

Bootstrapping Algorithm

For every alignment in the UMLS 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**

Bootstrapping Algorithm

For every alignment in the UMLS 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“:

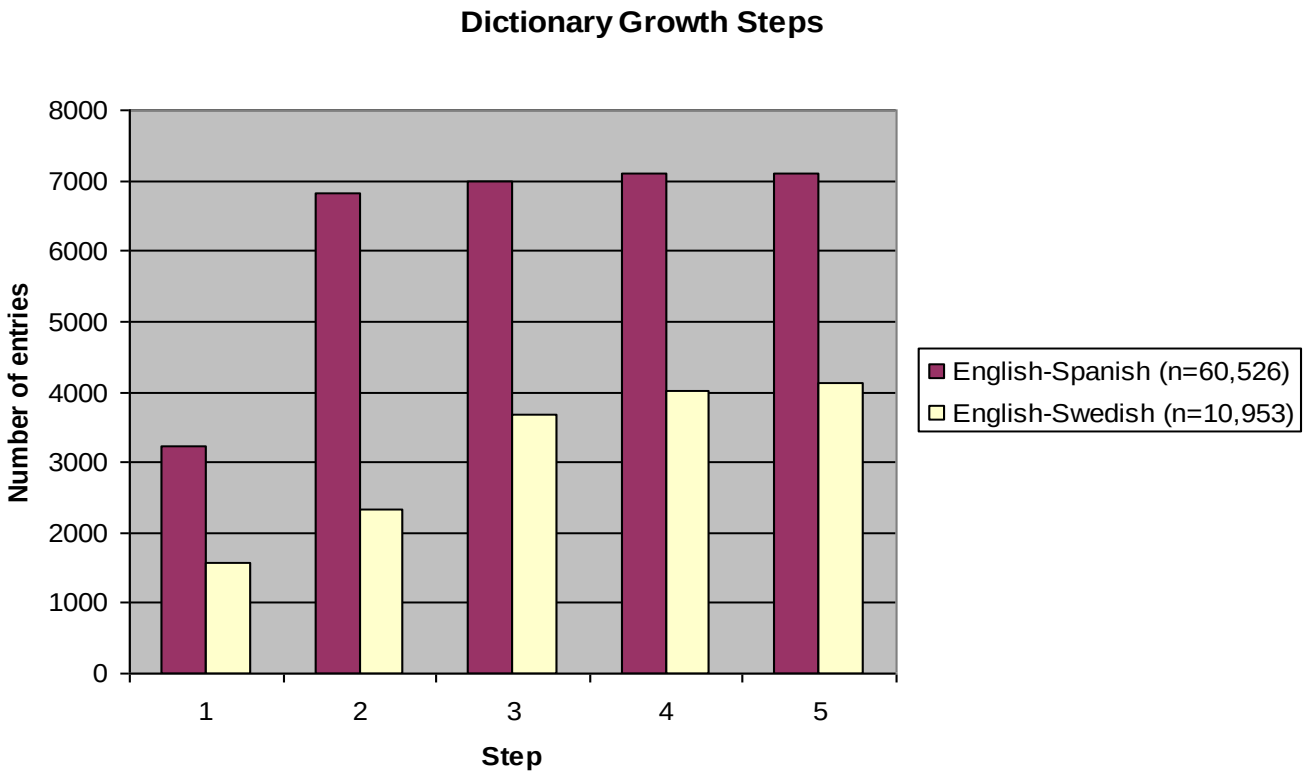
|skin| **#derma**
|abnorm|alities| **#anomal**

„Cirugia de la pared abdominal“:
„Cirugia de piel“:
„Malformacion de la piel“:

|malformation| → **|malform|ation|**
|piel| **#derma**

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

Bootstrapping Results



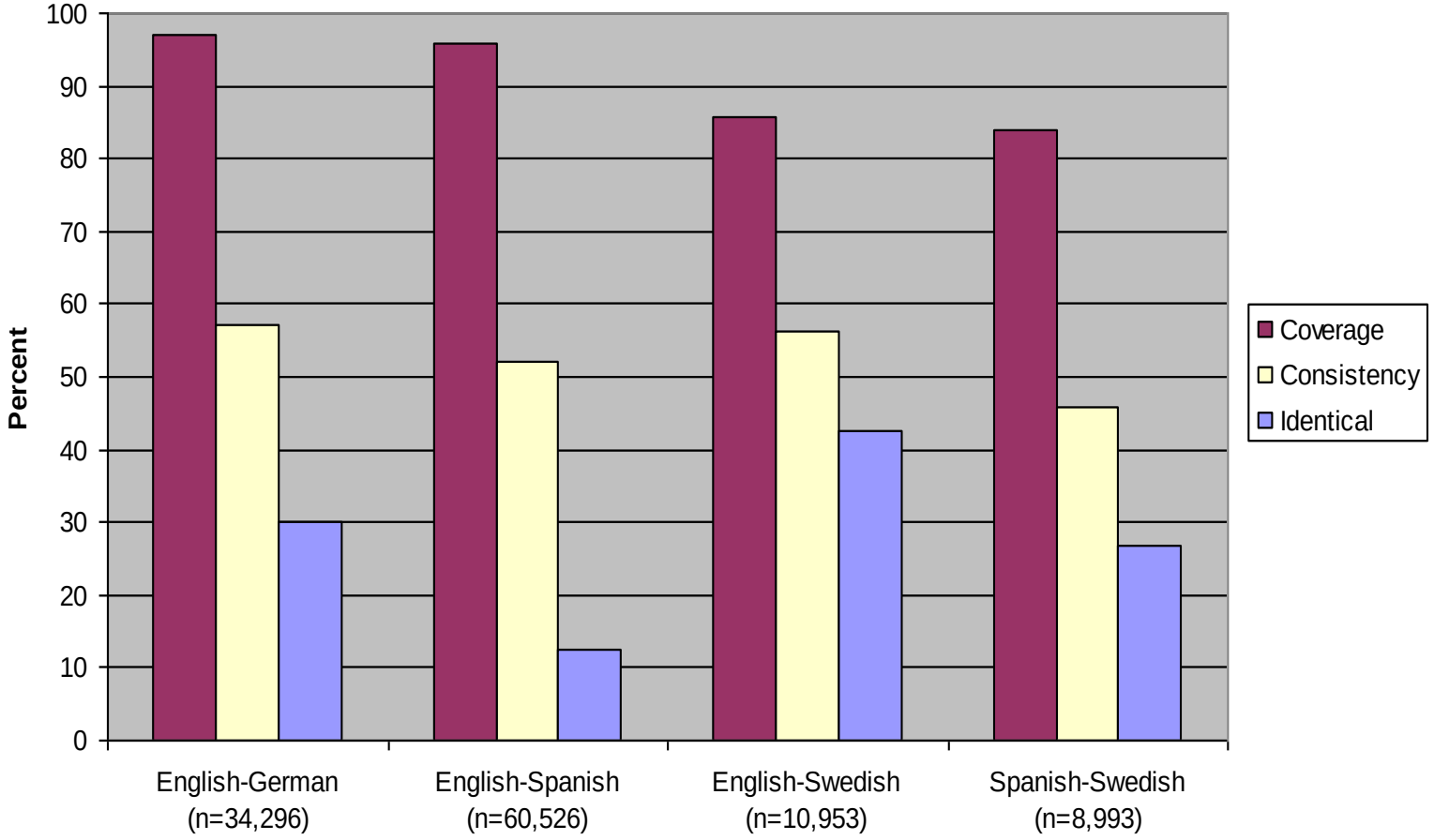
Total: 7,154 Spanish and 4,148 Swedish entries acquired

Evaluation

- Process the English-Spanish and English-Swedish UMLS alignments with the Morphosaurus system
- Additionally process Spanish-Swedish UMLS alignments

- 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-occuring on both sides
    - *N,M*: Number of MIDs occuring on only one side
  - Identical Indexes

Results


100
90
80
70
60
50
40
30
20
10
0
Percent
English-German (n=34,296)
English-Spanish (n=60,526)
English-Swedish (n=10,953)
Spanish-Swedish (n=8,993)
Coverage
Consistency
Identical

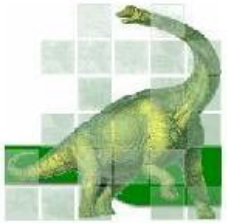
Conclusion

- Cross-Language Document Retrieval based on the matching of search/document terms on a language-independent, interlingual layer.
- Significant amount of English, German, and Portuguese subwords can be mapped to Spanish and Swedish cognates using simple string substitution rules.
- These cognate seed lexicons are further enlarged by subword translations which are *not* cognates by bootstrapping and using parallel corpora.
- Methodology proved to be useful in a standardized CLIR experimental setting
- Generality of the approach:
  - Need of large, aligned corpora
  - Eurodicautum (12 languages, 5M entries), Eurovoc (13 languages), OECD, UNESCO, AGROVOC, etc.

www.morphosaurus.net

Morphosaurus Search

Search Engine



Search: darmkrebsrisiko (English/German/Portuguese)

Language: All Information Level: Any MorphoSaurus: Use

Search

Search results for 'darmkrebsrisiko' : **Best 20 Documents of 2095 matches (298 msecs)**

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%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

Evaluation

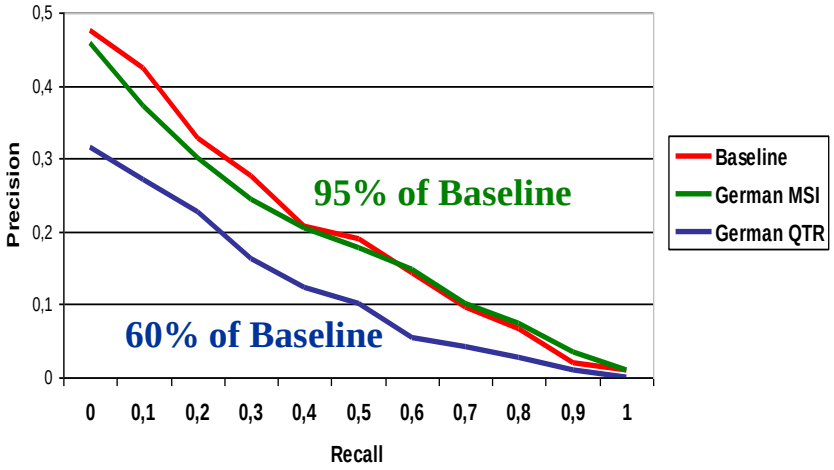
- OHSUMED-Corpus (Hersh et al., 1994)
  - Subset of MEDLINE
  - ~233,000 English documents
  - 106 English user queries, additionally translated to German, Portuguese, Spanish and Swedish by medical experts
  - query-document pairs have been manually judged for relevance
- Search Engine: Lucene
  - http://lucene.apache.org/

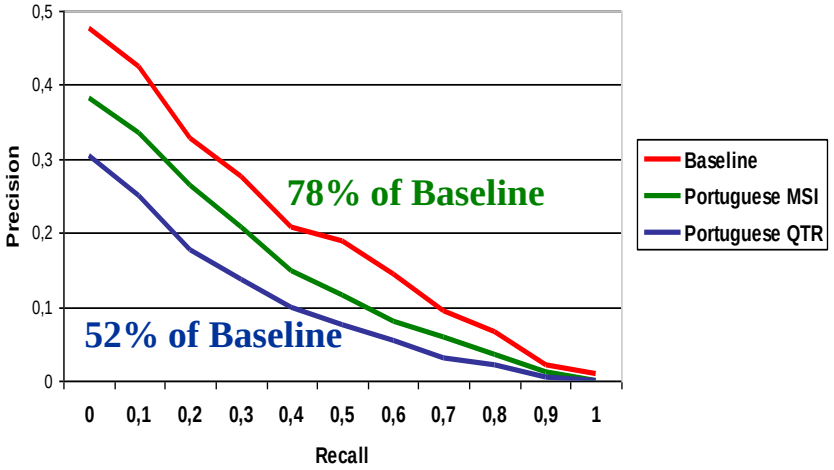
Evaluation

- 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

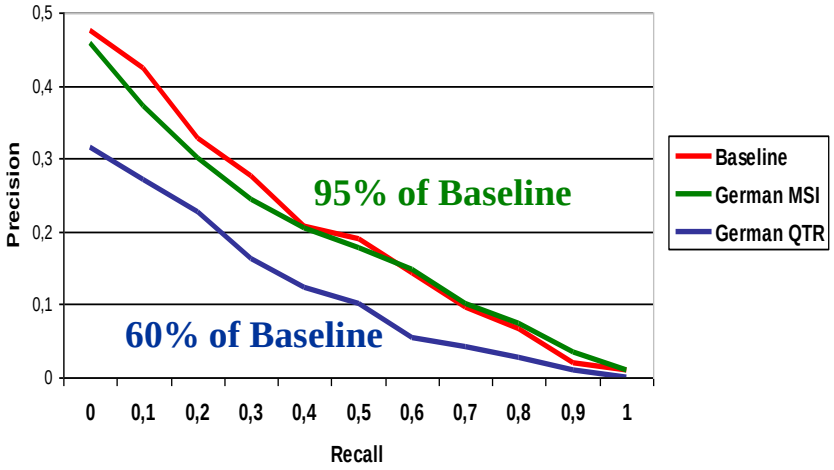
German (n = 22,385) Top 200 **Portuguese (n = 14,862)**



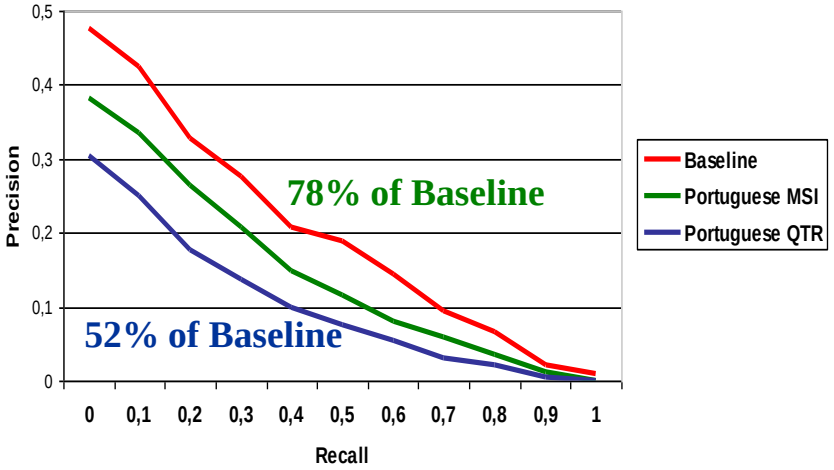


Evaluation Results

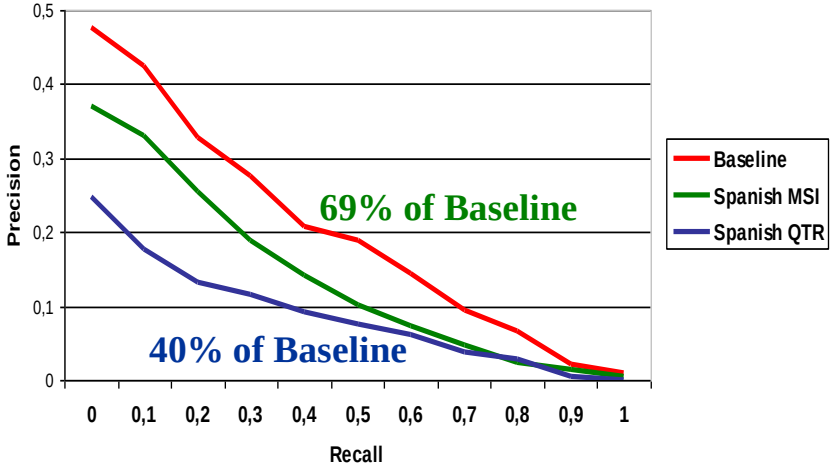
German (n = 22,385) — Top 200



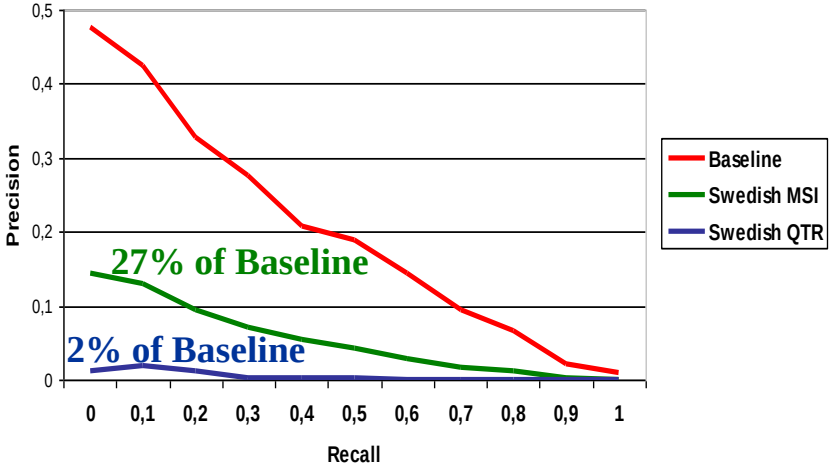
Portuguese (n = 14,862)



Spanish (n = 7,154) — Top 200



Swedish (n = 4,148)



Semantic Mapping

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**