ImageCLEFmed

MORPHOSAURUS in ImageCLEFmed 2006:

The effect of subwords on biomedical IR

Philipp Daumke, Jan Paetzold, Kornél Markó

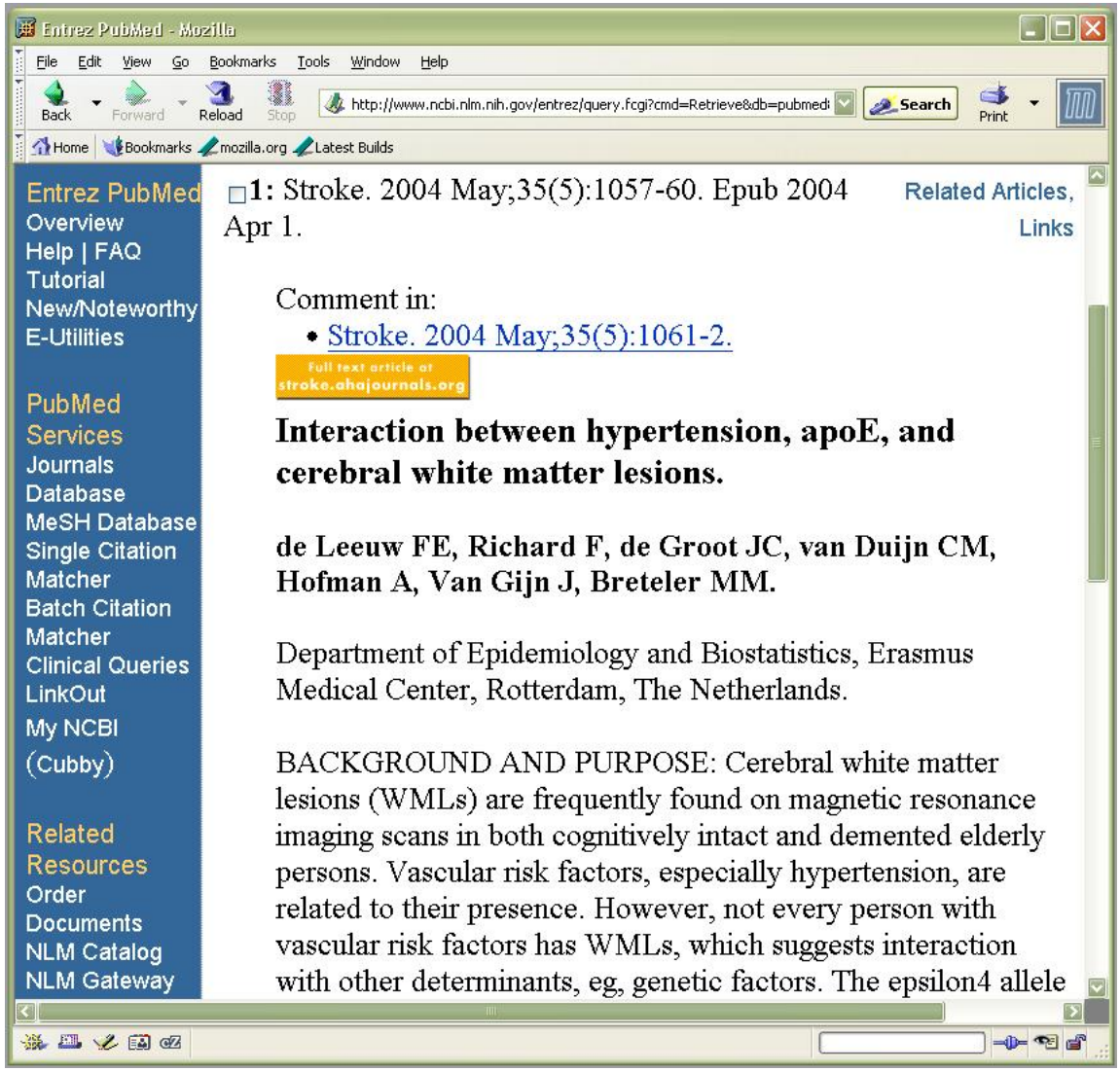
MORPHOSAURUS® MEDICAL SEARCH

UNIVERSITÄTS FREIBURG KLINIKUM

MorphoSaurus

Conventional Search Engines

Korrelation von Hypertonie und Läsion der Weißen Substanz…

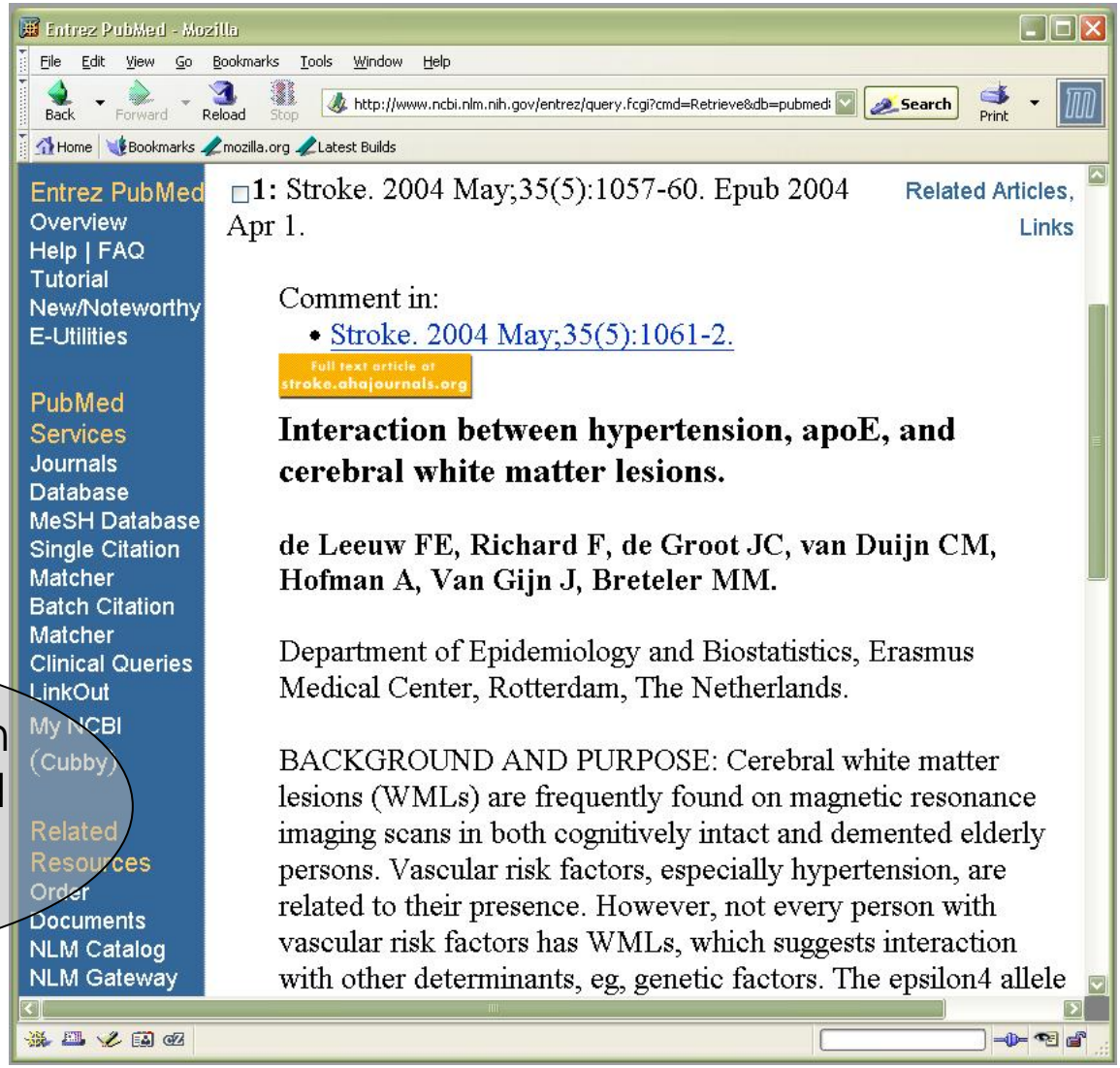


MorphoSaurus

Conventional Search Engines

Korrelation von Hypertonie und Läsion der Weißen Substanz…

"Correlation of high blood pressure and lesion of the white substance"

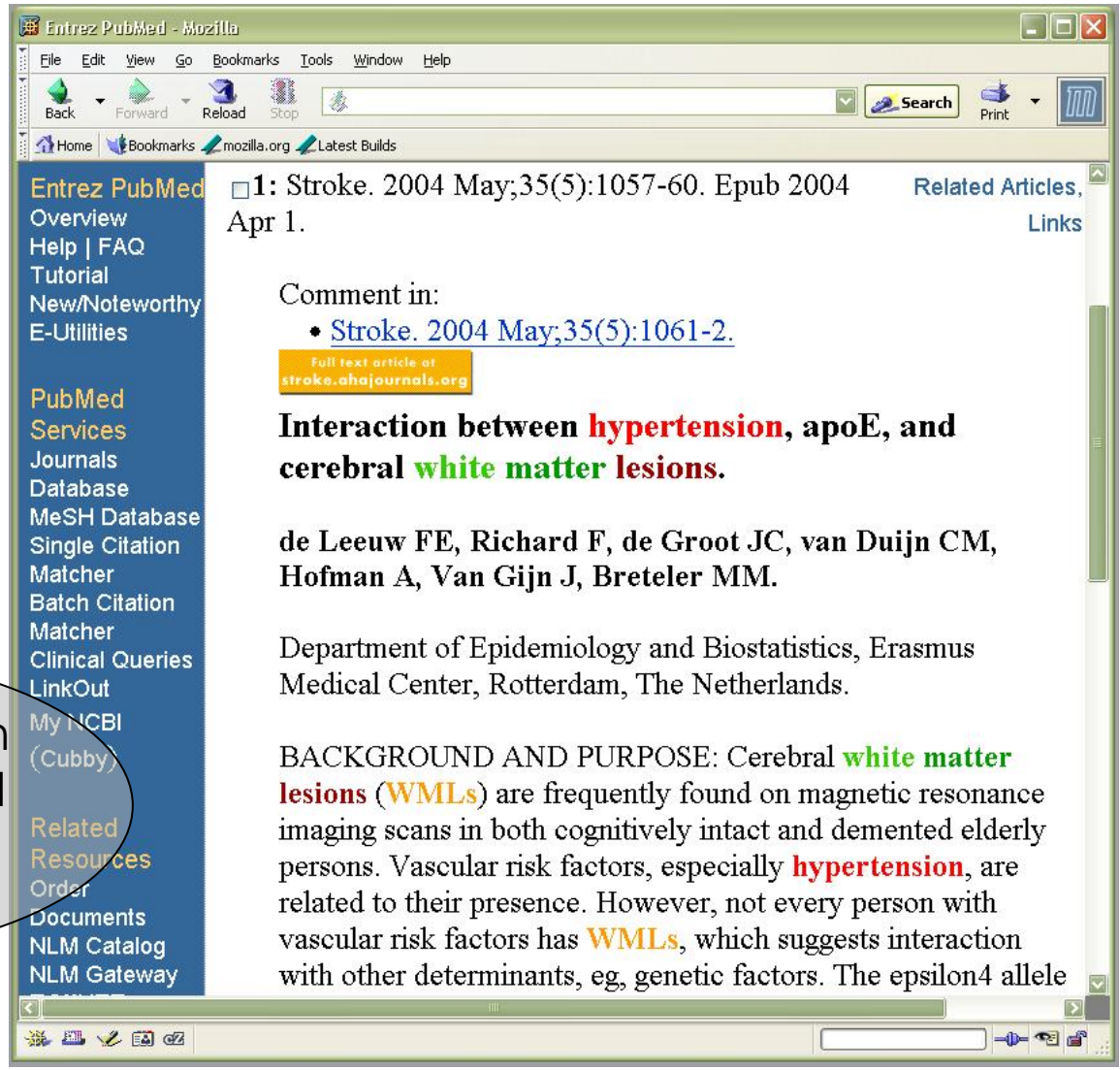


MorphoSaurus

Conventional Search Engines

Korrelation von Hypertonie und Läsion der Weißen Substanz…

"Correlation of high blood pressure and lesion of the white substance"

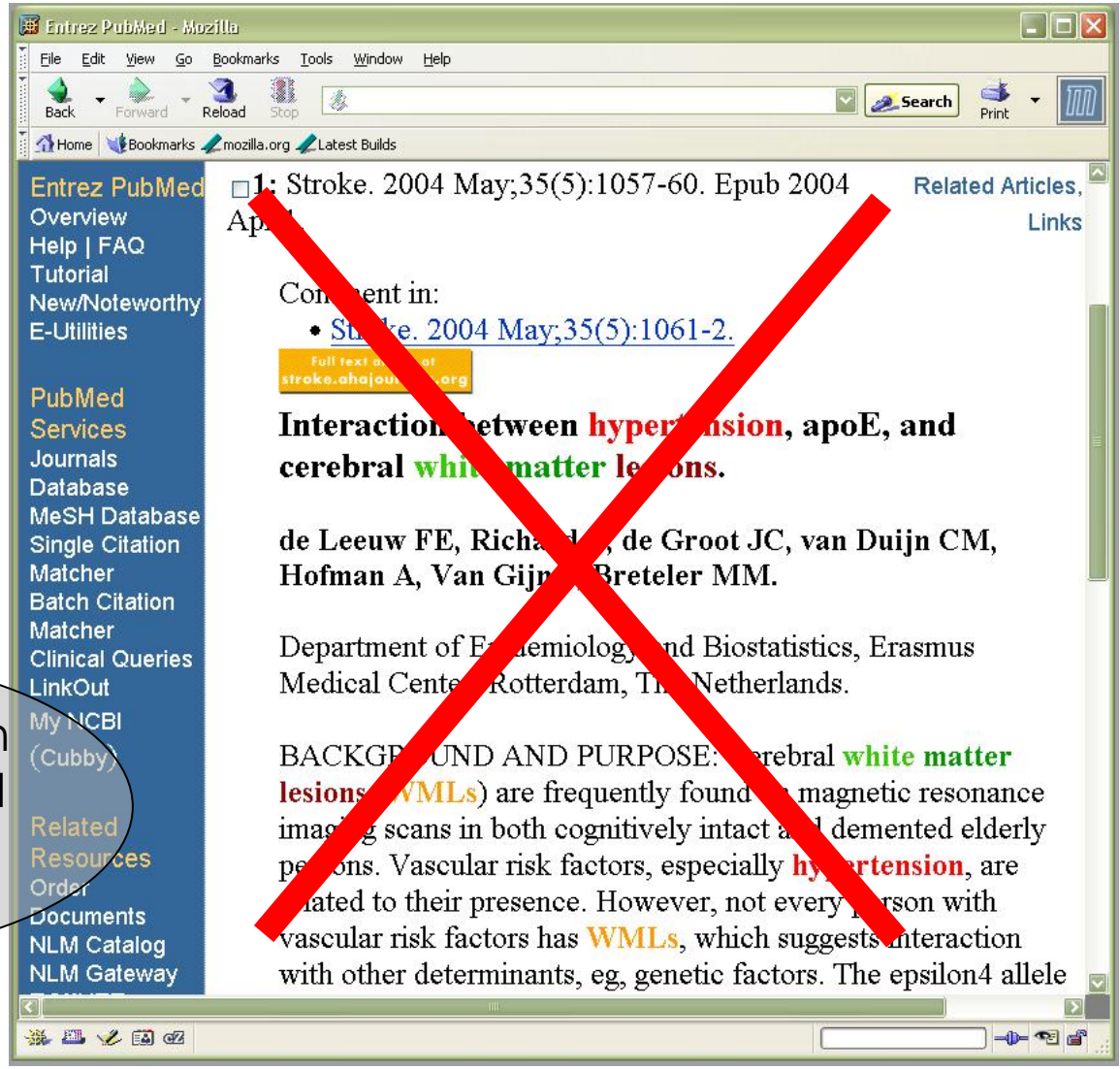


MorphoSaurus

Conventional Search Engines

Korrelation von Hypertonie und Läsion der Weißen Substanz…

"Correlation of high blood pressure and lesion of the **white** substance"



MorphoSaurus

Linguistic Phenomena

- **Inflection**: leukocytes, appendix – appendices
- **Derivation**: leukocyte → leukocytic
- **Composition**: *para|sympath|ectomy*,
  Magen|schleim|haut|entzünd|ung
- **Acronyms, Abbreviations**: AIDS(en,de)↔SIDA(fr), SARS, OECD
- **Proper Names**: Aspirin ↔ ASS, ...
- **Orthographic Variants**:
  - Kolonkarzinom, Coloncarzinom, Coloncarcinom
  - Oesophagus, Esophagus
- (Multi Word) **Synonyms**:
  - Hypertension ↔ High Blood Pressure
- **Homonyms/Polysemes**: Cold → Common Cold, Low Temperature, Chronic Obstructive Lung Disease

MorphoSaurus

Domain specific challenges

Biomedical language is characterized by:

- Diverse user groups → **diverse documents**
  - **Telegraphic** style in Electronic Patient Records („58 yo patient c/o pain, o/e NAD“) vs.

    flowery language of pathological reports („size of a ripe peach“)
  - **Laymen** terminology in health portals („inflammation of the gall bladder“) vs.

    Professional terminology in scientific papers („cholecystitis“)
- **Neoclassical compounds** (mix of latin and greek roots)

  („esophagogastroduodenoscopy“)
- **Medical Neologisms / Collocations / Phraseologisms**

  („cardiac muscle bridges“)

MorphoSaurus

MORPHOSAURUS Solutions

Subwords as smallest units of meaning, whose meaning cannot be derived from smaller meaningful parts:

- Stems: *stomach, gastr, diaphys, vitamin c*
- Prefixes: *anti-, bi-, hyper-*
- Suffixes: *-ary, -ion, -itis*
- Infixes: *-o-, -s-*

Equivalence classes contain synonymous subwords and their translations in a thesaurus:

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

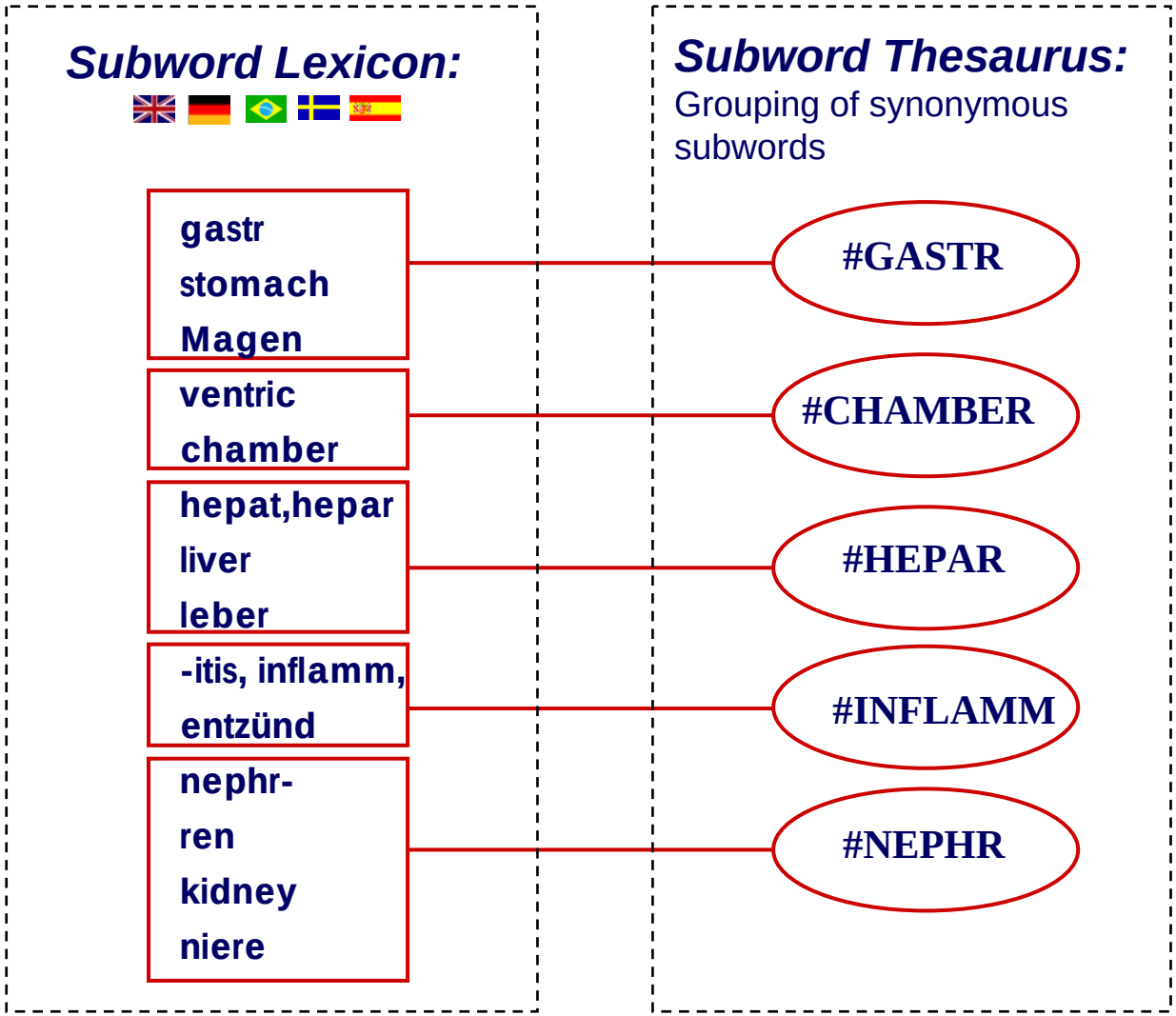
#inflamm = { inflam, -itic, -itis, entzuend, -itis, -itisch, inflam, flog, inflam, flog, -iolitis, ... }

Two lexical relations between equivalence classes:

- **expandsTo**: #hypertens → #high #blood #pressure
- **senseOf**: #head → {#chief;#caput}

MorphoSaurus

MORPHOSAURUS Lexical Resources



Lexicon statistics

- six languages
- 100,000 subwords
  25,000 (en,de), 15,000 (pt)…
- 20,000 eqClasses
- 500 exandsTo relations
- 1500 senseOf relations

MORPHOSAURUS

MORPHOSAURUS Resources

- **Subword-Lexicon:**
  - Organizes subwords in several languages

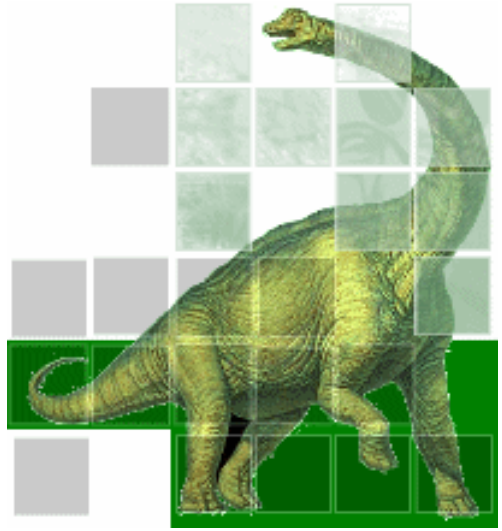
    (English, German, Portuguese, Spanish, French, Swedish)

- **Subword-Thesaurus:**
  - Contains equivalence classes, groups synonymous subwords (within and between languages)

- **Subword-Segmenter:**
  - Extraction of Subwords and Assignment of *Equivalence Classes*

MORPHOSAURUS Ressources

- **Subword-Lexicon:**
  - Organizes subwords in several languages
    (English, German, Portuguese, Spanish, French, Swedish)
- **Subword-Thesaurus:**
  - Contains equivalence classes, groups synonymous subwords (within and between languages)
- **Subword-Segmenter:**
  - Extraction of Subwords and Assignment of *Equivalence Classes*



MORPHOSAURUS
(www.morphosaurus.net)

MorphoSaurus

Example

High TSH values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...

Original

MorphoSaurus

Example

Original		Orthografic Normalisation
High TSH values suggest the diagnosis of primary hypo-thyroidism ...	**Orthografic Normalisation** →	high tsh values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...	***Orthografic Rules***	erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...

MorphoSaurus

Example

Original
High TSH values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...

Orthografic Normalisation

Orthografic Rules

high tsh values suggest the diagnosis of primary hypo-thyroidism ...
erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...

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 **schilddruese** n **unter funktion**

MorphoSaurus

Example

Original	Orthografic Normalisation	Orthografic Rules	Segmenter Subword Lexicon
High TSH values suggest the diagnosis of primary hypo-thyroidism ...			high tsh values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...			erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...

Interlingua	Semantic Normalisation	Subword Thesaurus	Segmenter Subword Lexicon
#up tsh #value #suggest #diagnost #primar #hypo #thyre			**high tsh value** s **suggest** the **diagnos** is of **primar** y **hypo thyroid** ism
#up tsh #value #permit #diagnost #primar #thyre #hypo #function			er **hoeh** te **tsh wert** e **erlaub** en die **diagnos** e einer **primaer** en **schilddrues** en **unter funktion**

MorphoSaurus

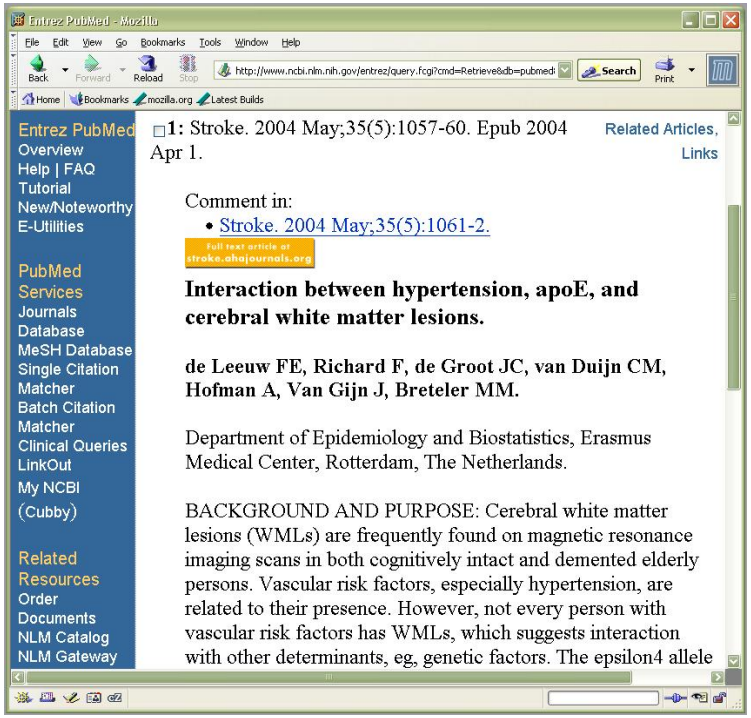
Example

Original	Orthografic Normalisation	Orthografic Rules	Segmenter Subword-Lexicon
High TSH values suggest the diagnosis of primary hypo-thyroidism ...			high tsh values suggest the diagnosis of primary hypo-thyroidism ...
Erhöhte TSH-Werte erlauben die Diagnose einer primären Schilddrüsenunterfunktion ...			erhoehte tsh werte erlauben die diagnose einer primaeren schilddruesenunterfunktion ...

Interlingua	Semantic Normalisation	Subword-Thesaurus	
#up tsh #value #suggest **#diagnost #primar #hypo #thyre**			**high tsh value** s **suggest** the **diagnos** is of **primar** y **hypo thyroid** ism
#up tsh #value #permit **#diagnost #primar #thyre #hypo** #function			er **hoeh** te **tsh wert** e **erlaub** en die **diagnos** e einer **primaer** en **schilddruese** n **unter funktion**

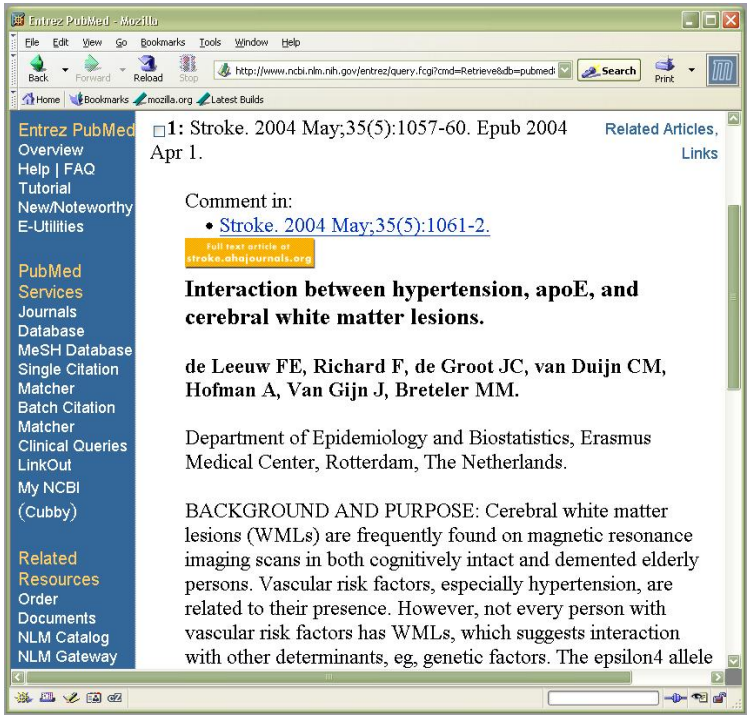
MORPHOSAURUS

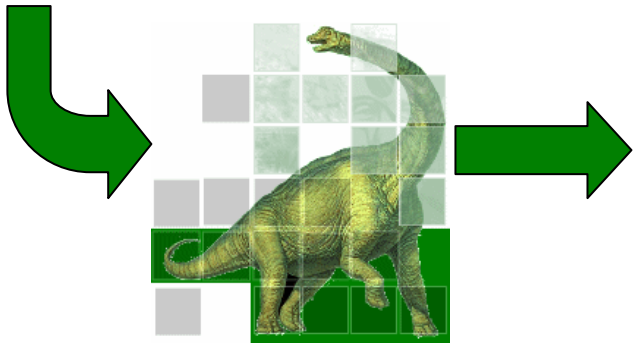
MORPHOSAURUS Search

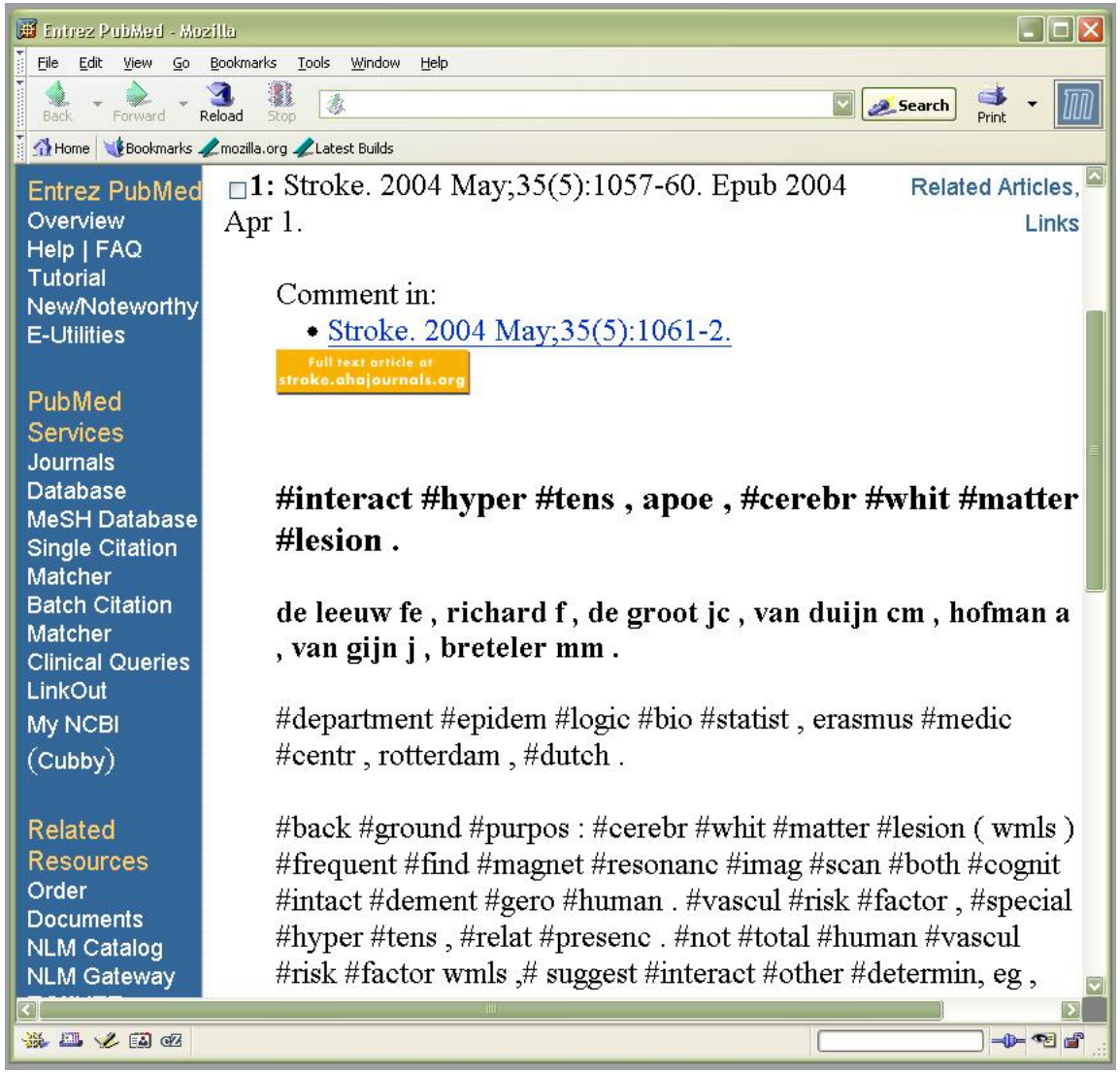


MorphoSaurus

MORPHOSAURUS Search



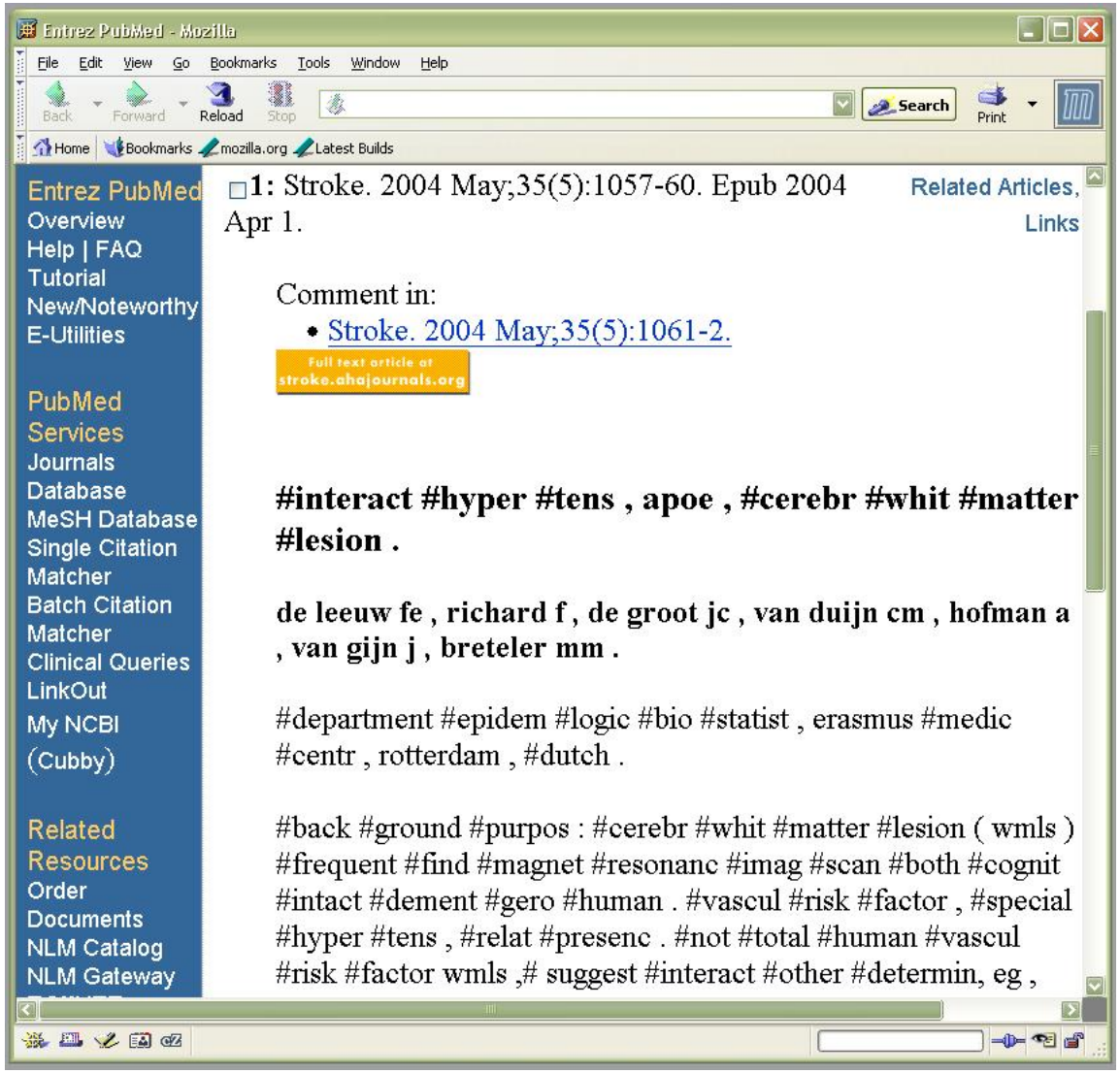




MorphoSaurus

MORPHOSAURUS Search

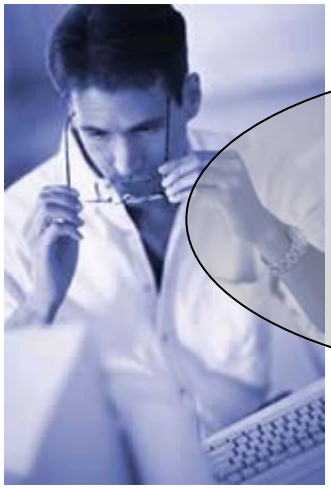
Korrelation von Hypertonie und Läsion der Weißen Substanz…



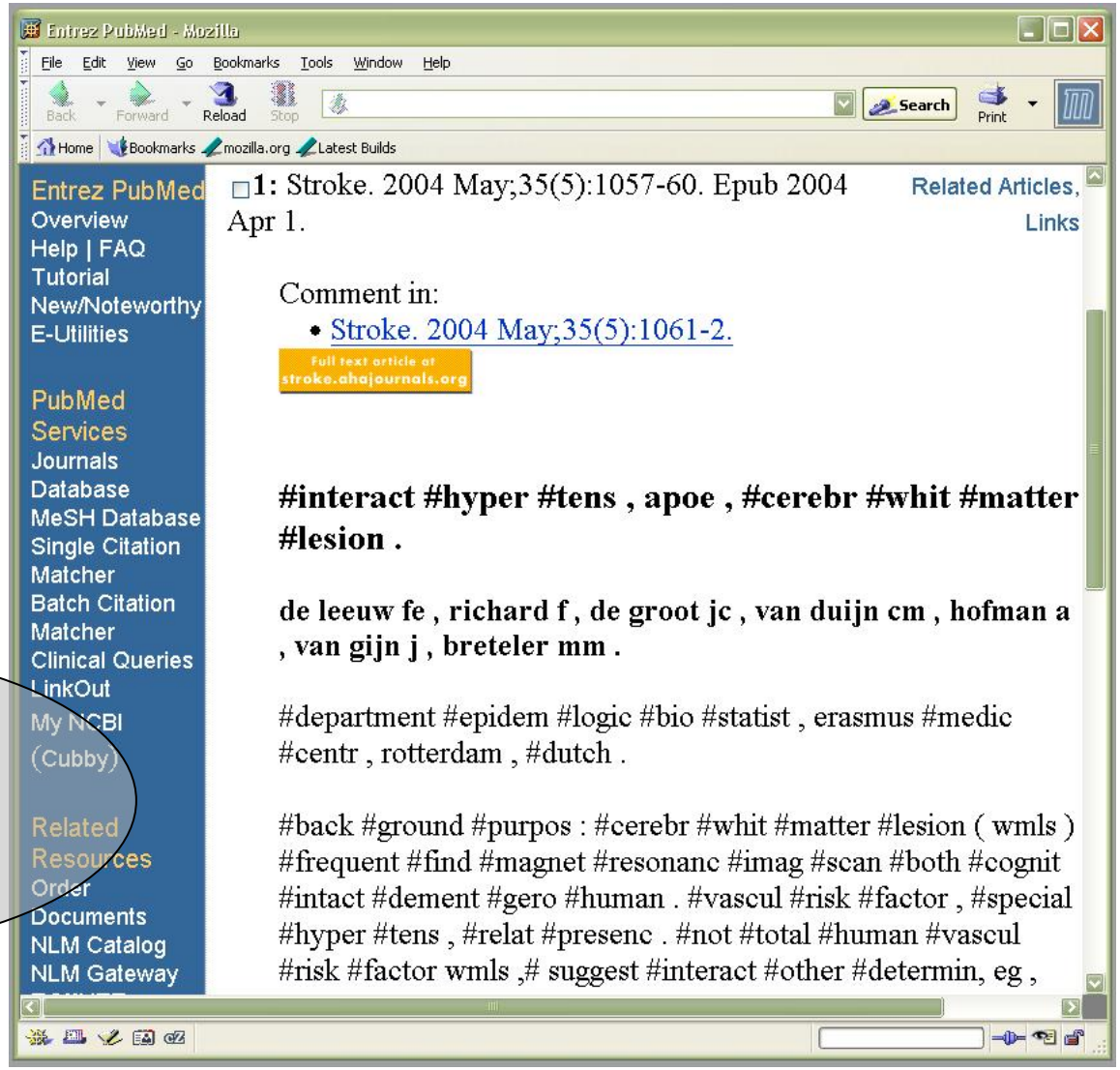
MorphoSaurus

MORPHOSAURUS Search

Korrelation von Hypertonie und Läsion der Weißen Substanz…



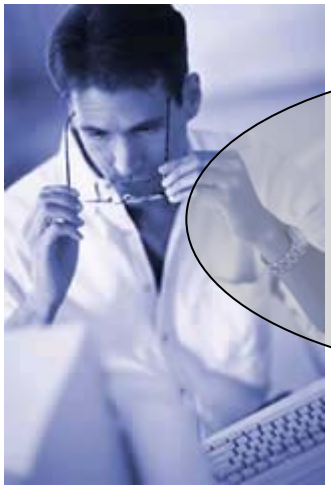
#correl #hyper #tens #lesion #whit #matter



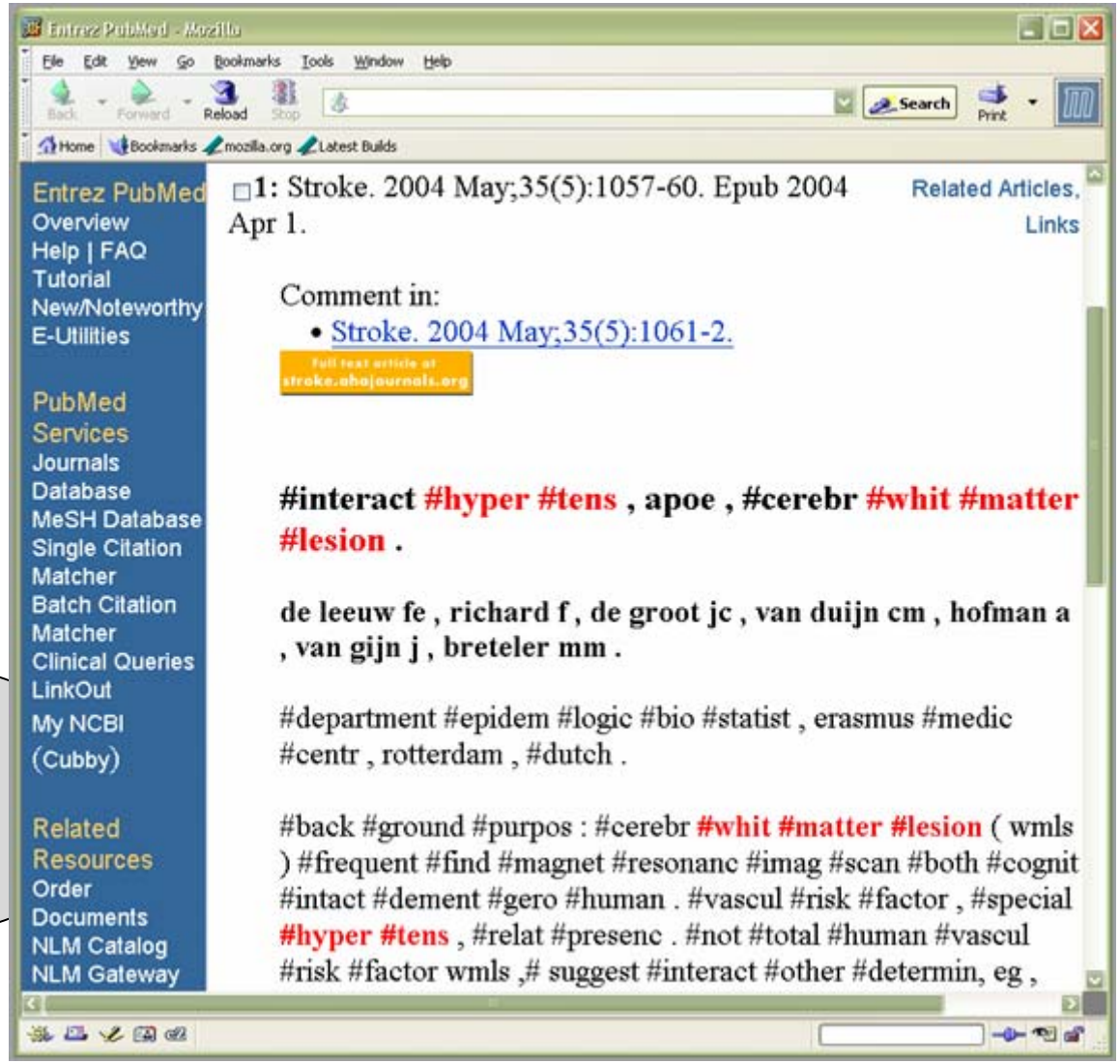
MorphoSaurus

MORPHOSAURUS Search

Korrelation von Hypertonie und Läsion der Weißen Substanz…



#correl #hyper #tens #lesion #whit #matter



MorphoSaurus

Evaluation

- **ImageCLEFmed 2006 Corpus**
  - 4 Subsets (CasImage, MIR, PEIR, PathoPic)
  - 40,708 English, 7805 German, 1899 French documents
  - 30 Queries in all three languages
- Search Engine: **Apache Lucene**
- Image Retrieval Tool: **GIFT (the GNU Image-Finding Tool)**
- **Text Only Scenarios**

Mono-Lingual Test Runs (English Subset)
- Original (Baseline)
- Stemming (Porter)
- Subwords
- Original + Stemming
- Original + Subwords

Cross-Lingual Test Runs
- Original (En-En, De-De, Fr-Fr)
- Subwords (En – All)
- Subwords (De – All)
- Subwords (Fr – All)

- **Text + Image Scenarios:** combining above scenarios with GIFT results

MORPHOSAURUS

Monolingual Results

Text Only

Scenario	Original (Baseline)	Stemming (Porter)		Subwords		Orginal + Stemming		Original + Subwords	
map	**0.1625**	**0.1482**	**91%**	**0.1297**	**80%**	0.1732	106%	**0.1792**	**110%**
top2avg	0.3778	0.3669	97%	0.3175	84%	0.4146	110%	**0.4525**	**120%**
P5	**0.4867**	**0.4667**	**96%**	**0.4867**	**100%**	0.5133	106%	**0.5933**	**122%**
P20	0.3600	0.4000	111%	0.3550	99%	0.4017	112%	**0.4500**	**125%**

Findings:

- neither Stemming nor Subword Search alone performed as good as Original Search regarding MAP
- Combination of Original and Subword Search increases precision values up to 25%

MORPHOSAURUS

Multilingual Results

Text Only

Scenario	Orig-All-All	Orig-En-En		Subwords-En-All		Subwords-De-All		Subwords-Fr-All	
map	**0.1068**	**0.1625**	**152%**	**0.1366**	**128%**	**0.1439**	**135%**	**0.0734**	**69%**
top2avg	0.2881	0.3778	131%	**0.3970**	**138%**	**0.3751**	**130%**	0.2028	70%
P5	0.3200	0.4867	152%	**0.4200**	**131%**	**0.4000**	**125%**	0.3103	97%
P20	0.2600	0.3600	138%	**0.3467**	**133%**	**0.3383**	**130%**	0.2293	88%

Findings:

- Baseline Search on the whole document collection performs much worse than on the English Subset
- Subword Search in English and German performs clearly better than the baseline (original)
- German Subword Search performs (almost) as good as the English Subword Search

Text + Image

- Image results provided by GIFT (the GNU Image-Finding Tool)
- Several merging algorithm were tested → All performed worse than *Text Only* results

MORPHOSAURUS

Conclusion/Outlook

- Mono- and Cross-Language Text Retrieval based on a sophisticated **interlingual layer**
- Can be combined with **any search engine**
- Mono-Lingual: **Combination of original and subword search remarkably boosts performance (up to 25%)**
- Cross-Lingual: **Subword approach clearly outperforms baseline, German as good as English**
- **Mean map values, good top-5(20)-precision values**
- A capable combination with Image Retrieval Tools is due

MorphoSaurus

www.morphosaurus.net