

MEDINFO 2007

413 Lygon Street, Brunswick East 3057 Australia

Presenter Name: Stefan Schulz

Country: 1. Germany, 2. Brazil

Qualification(s) :

MD (Doctor in Theoretical Medicine)

Vocational Training in Medical Informatics

Postdoctoral Habilitation degree in Medical Informatics

Position:

Associate Professor

Department/ Organisation :

1. Medical Informatics, Freiburg University Medical Center, Freiburg, Germany
2. Master Program of Health Technology, Catholic University of Paraná, Curitiba, Brazil

Major Achievement(s) :

Research in the fields of Medical Terminology, Biomedical Ontologies, Medical Language Processing, Text Mining, and Document Retrieval.



Medical Thesaurus Anomaly Detection by User Action Monitoring

**Jeferson L. Bitencourt^{a,b}, Pindaro S. Cancian^{a,b},
Edson Pacheco^{a,b}, Percy Nohama^{a,b}, Stefan Schulz^{b,c}**

^a Paraná University of Technology (UTFPR), Curitiba, Brazil

^b Pontifical Catholic University of Paraná, Master Program of Health Technology, Curitiba, Brazil

^c University Medical Center Freiburg, Medical Informatics, Freiburg, Germany



Introduction

Methods

Results

Discussion

Conclusion

Thesaurus

- Controlled Vocabulary for document indexing and retrieval
- Assigns semantic descriptors (concepts) to (quasi-)synonymous terms
- Contains additional semantic relations (e.g. hyperonym / hyponym)
- Examples: MeSH, UMLS, WordNet
- Multilingual thesaurus: contains translations (cross-language synonymy links)

Multilingual Thesaurus Management

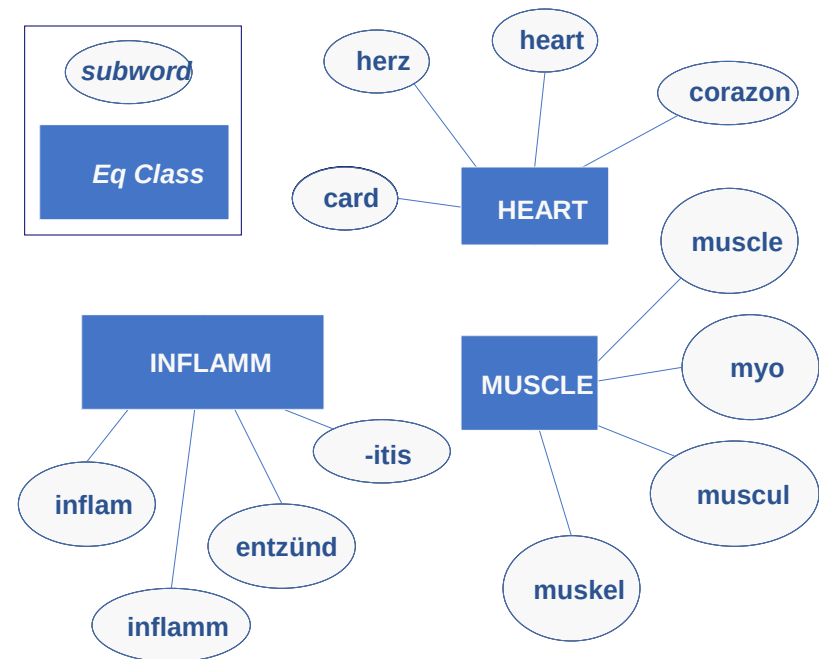
- International team of lexicon curators
- React to new terms and senses
- Decide which terms are synonymous / translations
- Decide which senses of a term have to be accounted for in the domain
- Requires quality assurance measures

Case study: Morphosaurus

- Medical subword thesaurus
- Organizes subwords (meaningful word fragments) in multilingual equivalence classes:
 - #derma = { **derm**, **cutis**, **skin**, **haut**, **kutis**, **pele**, **cutis**, **piel**, ... }
 - #inflamm = { **inflamm**, **-itic**, **-itis**, **phlog**, **entzuend**, **-itis**, **-itisch**, **inflam**, **flog**, **inflam**, **flog**, ... }
- Maintained at two locations:
Freiburg (Germany), Curitiba (Brazil)
- Lexicon curators: frequently changing team of medical students

Morphosaurus Structure

- Thesaurus:
~21.000 equivalence classes
- Lexicon entries:
 - English: ~23.000
 - German: ~24.000
 - Portuguese: ~15.000
 - Spanish : ~11.000
 - French: ~ 8.000
 - Swedish: ~10.000



Segmentation:

Myo | kard | itis

Herz | muskel | entzünd | ung

Inflamm | ation of the heart musc

Indexation:

#muscle #heart #inflamm

#heart #muscle #inflamm

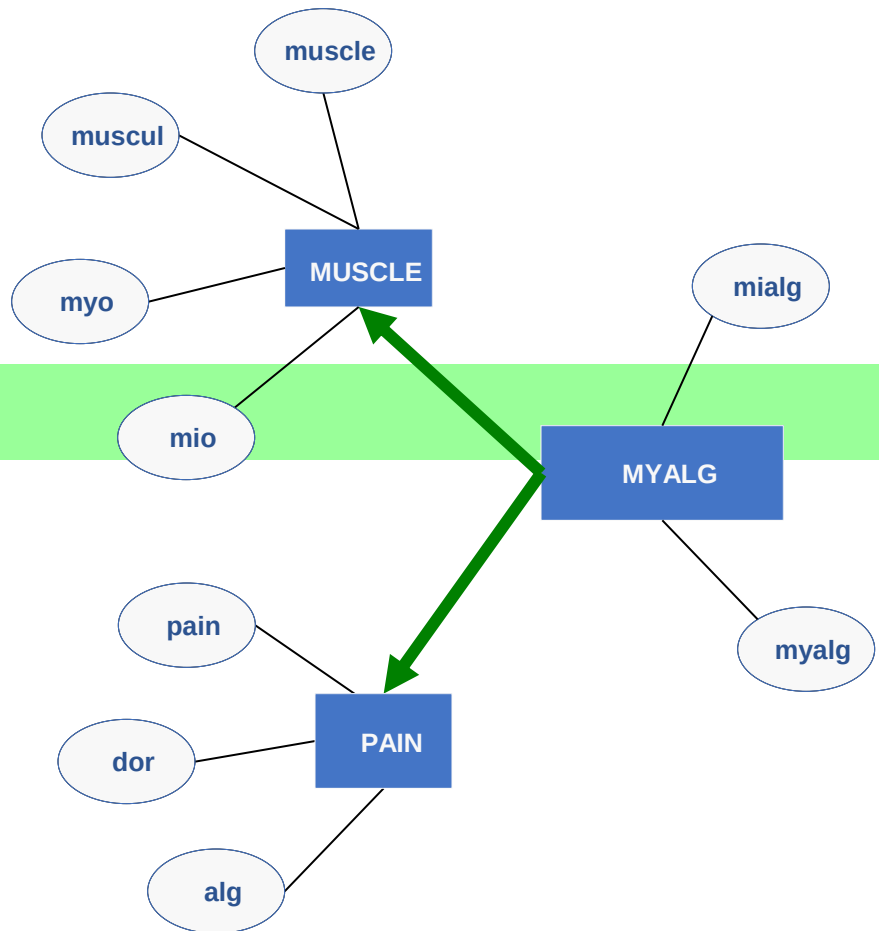
#inflamm #heart #muscle

Morphosemantic Normalization

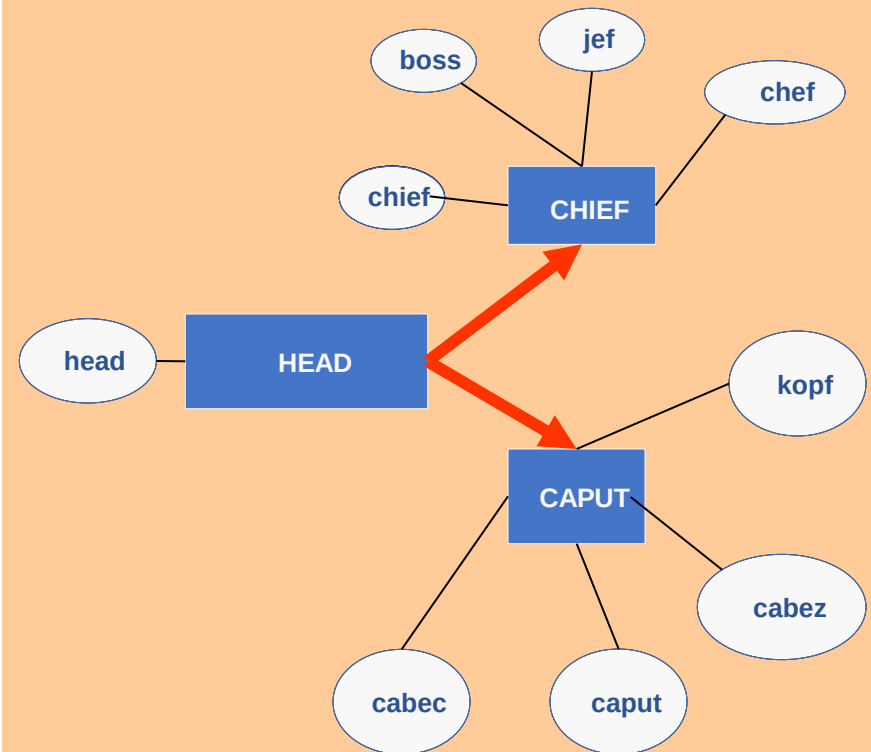
Original Document	Orthographic Normalization	Morphological Segmentation	Semantic Normalization
High TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism.	high tsh values suggest the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.	high tsh value s suggest the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.	#up# tsh #value# #suggest# #diagnost# #primar# #small# #thyre# #suppress# tsh #nivell# #suggest# #up# #thyre# .
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose, ein supprimierter TSH-Spiegel spricht dagegen für eine Schilddrüsenüberfunktion.	erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose, ein supprimierter tsh-spiegel spricht dagegen fuer eine schilddruesenueberfunktion.	er hoeh te tsh - wert e erlaub en die diagnosis einer primaer en hypo thyre ose, ein supprim iert er tsh - spiegel spricht dagegen fuer eine schilddrues en ueber funktion.	#up# tsh - #value# #permit# #diagnost# #primar# #small# #thyre# , #suppress# tsh - {#mirror# #nivell#} #speak# #thyre# #up# #function# .
A presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo.	a presenca de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario, enquanto níveis suprimidos de tsh sugerem hipertireoidismo.	a presenc a de valor es elevad os de tsh suger e o diagnost ico de hipo tireoid ismo primari o, enquanto niveis suprimid os de tsh suger em hiper tireoid ismo.	#actual# #value# #up# tsh #suggest# #diagnost# #small# #thyre# #primar# , #nivell# #suppress# tsh #suggest# #up# #thyre# .

Morphosaurus: 2 Semantic Relations

Composition: *Has_word_part*



Specialization: *Has_sense*

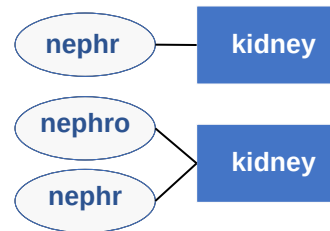


Morphosaurus Building Pragmatics

Morphosaurus Building Pragmatics

- Properly delimit subword entries so that they are correctly extracted from complex words:

nephrotomy -> *nephr* | *oto* | *my*
nephrotomy -> *nephro* | *tomy*

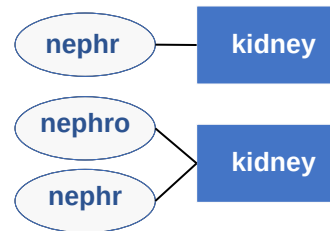


OR

Morphosaurus Building Pragmatics

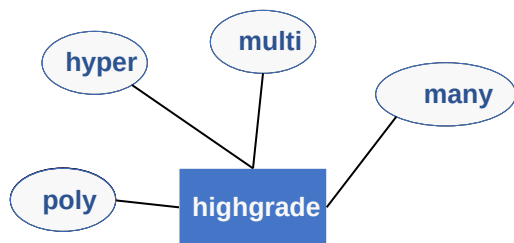
- Properly delimit subword entries so that they are correctly extracted from complex words:

nephrotomy -> *nephr* | *oto* | *my*
nephrotomy -> *nephro* | *tomy*

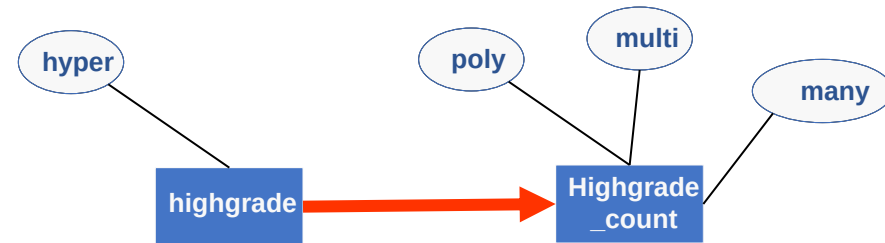


OR

- Create consensus about the scope of synonymy classes, especially with regard to highly ambiguous words



OR



Morphosaurus Quality Assurance

- Content quality: Identify content errors in the thesaurus content (see *Andrade* et al., MEDINFO 2007)
- Process quality: Detect and prevent user action anomalies
- User action anomalies: actions that consume effort without any positive impact :
uncoordinated edit / update / delete “do undo” transactions done by different lexicographers

Introduction

Methods

Results

Discussion

Conclusion

Identification of Editing Anomalies

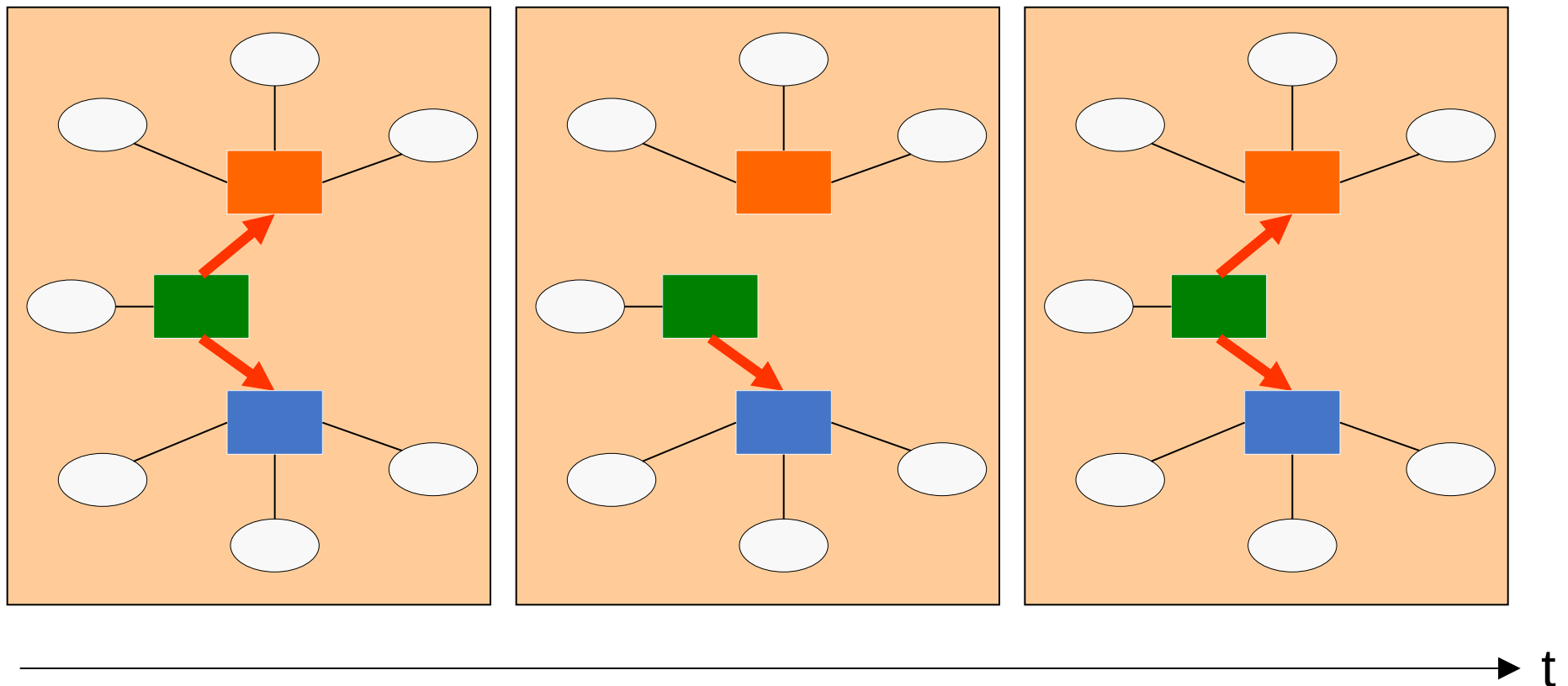
- Analysis of data logs patterns:
86 thesaurus backups covering 9 months
- Assessing relevance of anomaly patterns by comparing the thesaurus descriptors affected with those debated in a Morphosaurus editor online forum

Identification of Editing Anomalies

- Analysis of data logs patterns:
86 thesaurus backups covering 9 months
- Assessing relevance of anomaly patterns by comparing the thesaurus descriptors affected with those debated in a Morphosaurus editor online forum

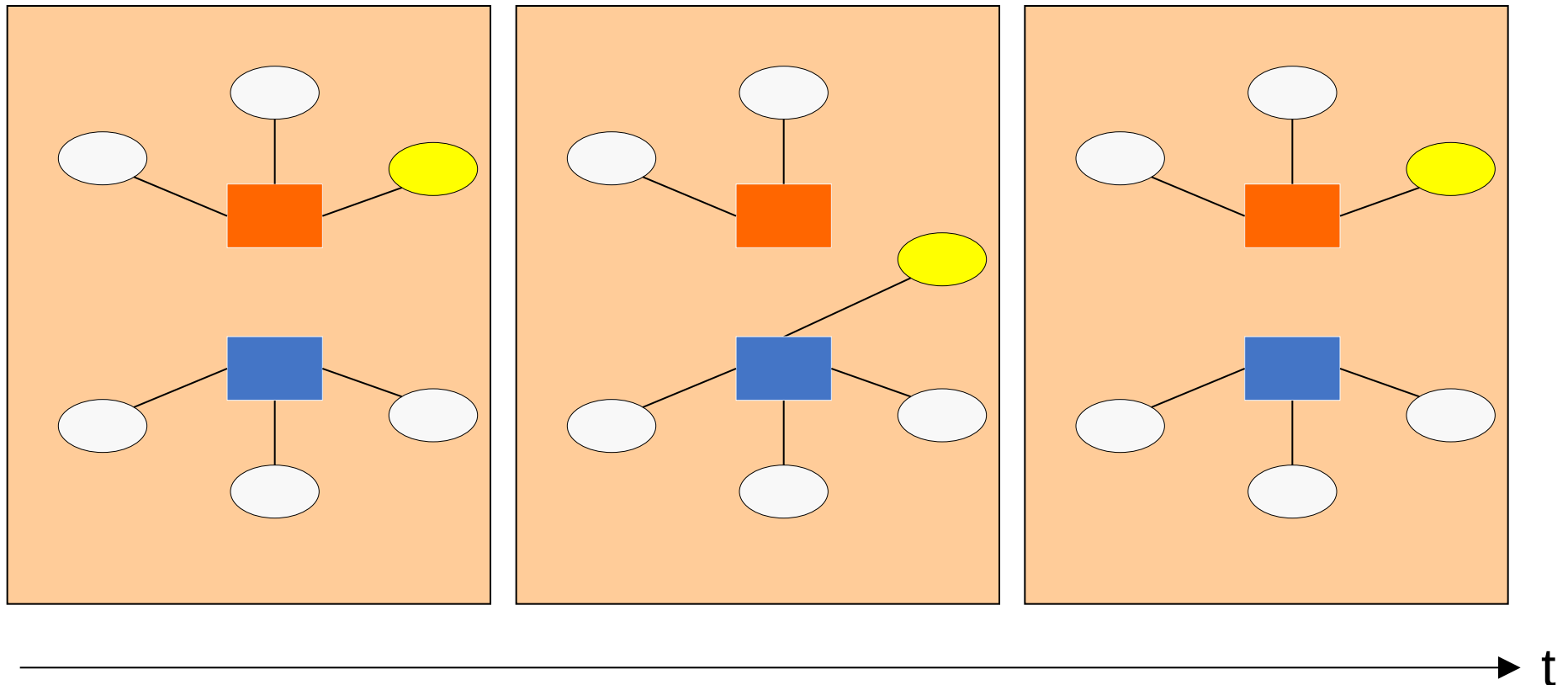
Anomalies: Typology

1. Relationship anomaly



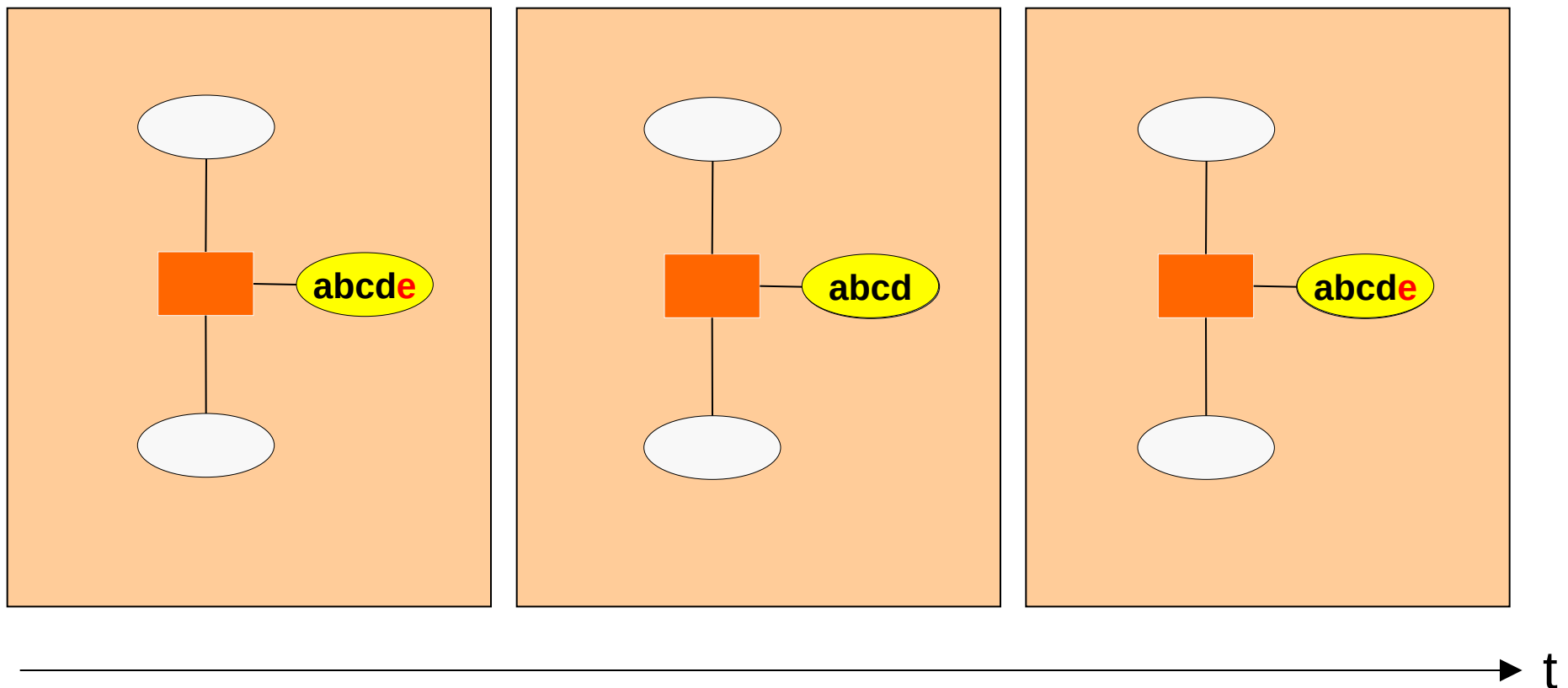
Anomalies: Typology

2. Type Anomaly



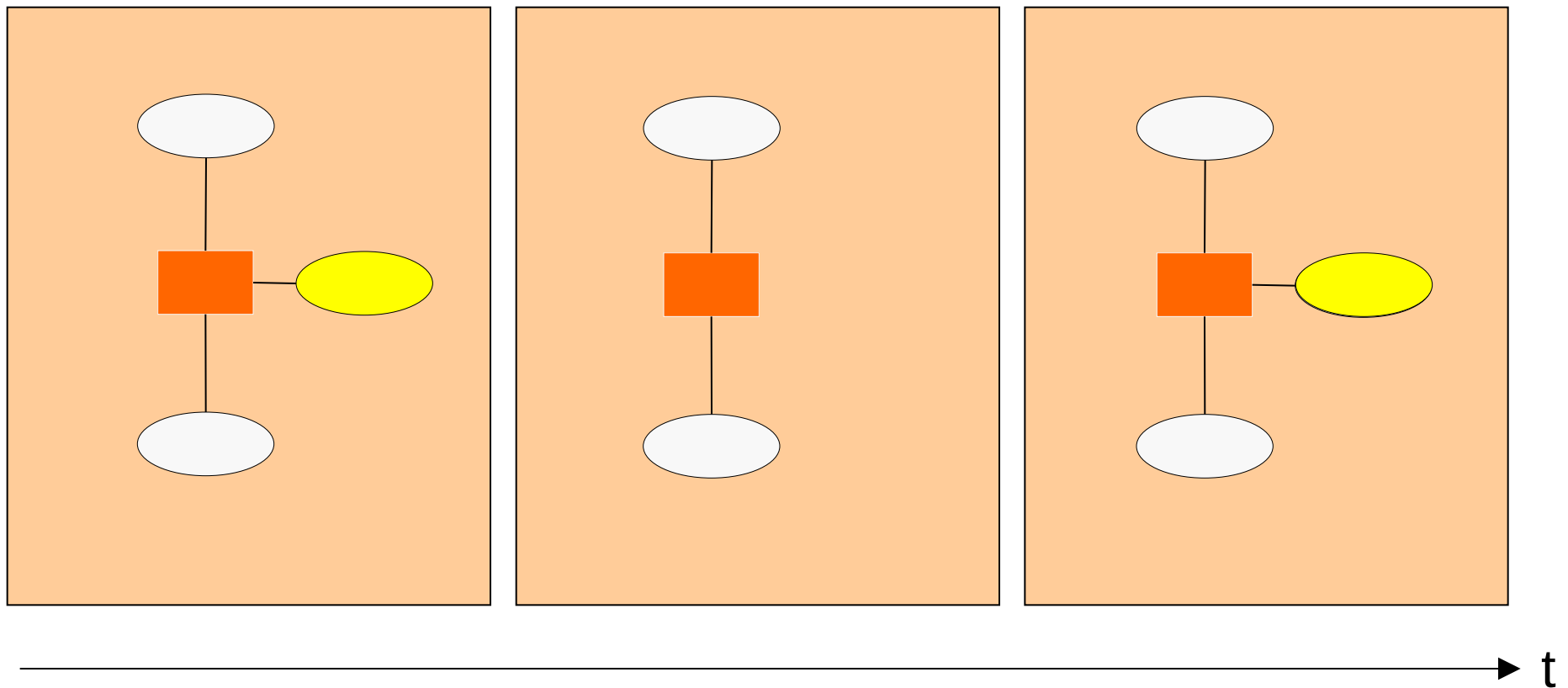
Anomalies: Typology

3. Delimitation Anomaly



Anomalies: Typology

4. Permanence anomaly



Identification of Editing Anomalies

- Analysis of data logs patterns:
86 thesaurus backups covering 9 months
- Assessing relevance of anomaly patterns by comparing the thesaurus descriptors affected with those debated in a Morphosaurus editor online forum

Example of Morphosaurus forum entry

MIDcompare eng-ger murmuriikrpia 002530 [Inbox 100](#) [Eng Ger 100](#) [Eng Por](#)

★ **Michael Schultheiss** to morphosaurus [show details](#) 8/22/05 [Reply](#)

MIDcompare eng-ger-doc list

1. Current status in list:
|murmuriikrpia |002530 |221 |0 |0,0038|1,0000|0,6679|
2. Current status in thesaurus (lexicon)
Eq Class 2530 for indexing (all entries are stems)
"murmur" (ger)
"murmur" (eng)
"murmur" (por)
"_murmull" (span)
"_soplo" (span)
3. Problem description
Kind of problem: language specific problem. The english "murmur" is frequently used for an abnormal heart sound. The german "murmur" might exist, but is very, very rare
4. Solution:
I added the german lexemes "murmeln" and "raun" to Eq class 2530. They are not heart-specific auscultation terms like the english "murmur", but important german equivalents.
5. Documentation in Comment field of Eq class: ---
6. Neighborhood:

EqClass spotted by corpus based content quality analysis, cf. Andrade et al., MEDINFO 2007

Introduction

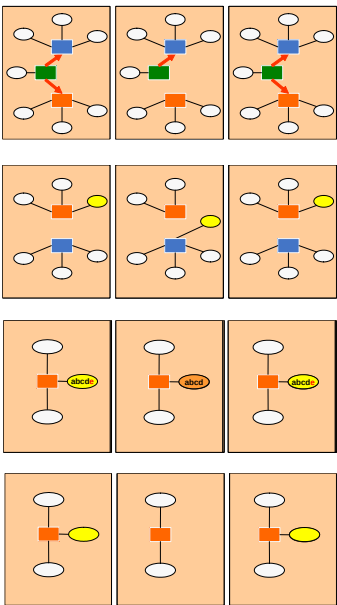
Methods

Results

Discussion

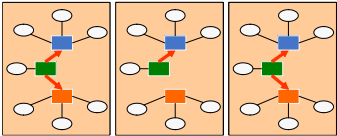
Conclusion

Results



Anomaly Type	Occurrences	Discussed in Forum
Relationship anomaly	76	28
Type anomaly	18	18
Delimitation anomaly	0	0
Permanence anomaly	5	4

Relationship anomalies: multiple changes



Number of do-undo actions	Occurrences	Discussed in Forum
2	4	1
4	16	6
5	2	2
6	2	0
7	23	10

Problems found by Log Analysis

Problem Type	Occurrences	Found by Log Analysis
An expected relation relating ambiguous or expandible semantic indentifiers (<i>has_sense</i> type or <i>has_word_part</i> type)	86	24
Entries assigned to one semantic indentifier did not cover all languages.	80	6
The same sense is represented by two unrelated semantic indentifier.	70	8
Lexicon entries assigned to one semantic indentifier diverge in meaning.	11	1
Language specific entry do not translate to other languages.	11	0
Orthographic errors.	16	1
Similar senses are represented by two unrelated semantic indentifiers, one of them of the type “excluded from indexing”.	31	0
Errors caused by incorrect subword delimitation	16	0
Errors caused by incorrect functioning of the segmentation engine.	4	0

Introduction

Methods

Results

Discussion

Conclusion

Discussion of Results

- Assignment of semantic relations: main cause of do-undo anomalies (up to seven do-undos)
- Nearly half of editing anomalies concern semantic indentifiers also identified as problematic by corpus analysis
- Problems discussed in forum exceeds those identifiable by log analysis
- Surprising: no anomaly of string delimitation found

Introduction

Methods

Results

Discussion

Conclusion

Anomaly detection

- Detects waste of resources by “do - undo” actions in thesaurus management
- Helps create consensus in borderline decisions
- Useful to discover common anomalies
- To be complemented by other techniques
- Higher process effectiveness by integration of quality assessment routines in the thesaurus management tools: User alert at runtime

Anomaly detection at runtime

