



Norman Kummer, Joachim Wagner

Phrase processing for detecting collocations with KoKS*

- ***Korpusbasierte Kollokationssuche**
(corpus based search for collocations)

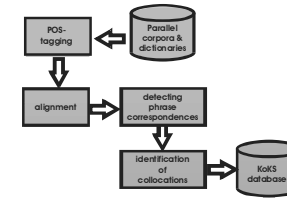
University of Osnabrück (Germany): KoKS-Project

contents

- **detection of phrases**
- **identifications of collocations**
- **evaluation (results)**

University of Osnabrück (Germany): KoKS-Project

system overview



University of Osnabrück (Germany): KoKS-Project

used bilingual corpora

- **DE-News**
 - radio news broadcast
 - translated by volunteers
- **EU-publications**
 - press releases
 - political documents
 - contracts
- **the four Harry Potter books**

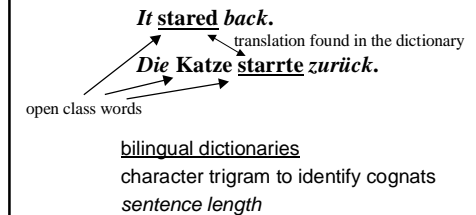
University of Osnabrück (Germany): KoKS-Project

alignment of sentences ^{1/2}

- **distance measure**
 - bilingual dictionaries
 - character trigram to identify cognats
 - sentence length

University of Osnabrück (Germany): KoKS-Project

alignment of sentences ^{2/2}

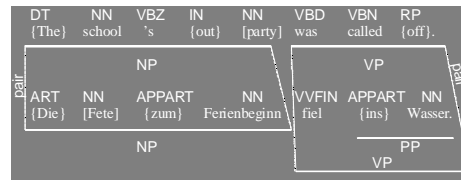


University of Osnabrück (Germany): KoKS-Project

detecting phrase correspondences 3/5

- **POS tags sequences**
 - extracted from chunk-parsed monolingual corpora
 - distinguished by syntactic category
- **pair matching phrases**
- **example:**

detecting phrase correspondences 4/5



detecting phrase correspondences 5/5

- **multiple NPs**
- **identify non-literal-phrases**
- **no word alignment is used**
- **all combinations are considered**
- **a predefined number of references is required**

collocativity measure

- **Breidt's definition of collocations**
 - compositional semantics
- **translation as semantics**
- **distance measure used in sentence alignment**

results

- **detecting phrase correspondences**
- **collocativity measure**

results (phrase detection) 1/3

- **so far, we processed**
 - all sentences with at most 19 words
 - approx. 70,000 sentence pairs
- **next table shows examples**
 - ordered by frequency (*f*)

results (phrase detection) ^{2/3}

| rank | f | German | English | correspondence |
|------|----|--------------------|-----------------|----------------|
| 22 | 30 | Professor | Dumbledore | bad |
| 23 | 30 | die Tür (the door) | Harry | bad |
| 24 | 29 | Professor | Professor Lupin | near |
| 25 | 29 | Schloss | the castle | good |
| ⋮ | ⋮ | | | |
| 33 | 25 | zu Harry | to Harry | good |
| 34 | 24 | will | do n't want | near |
| 35 | 24 | schien | seemed to be | good |
| 36 | 24 | ist | do n't know | bad |
| 37 | 24 | sagte (said) | 've got | bad |
| 38 | 23 | Dementoren | the dementors | good |
| 39 | 22 | Kammer | the Chamber | good |

University of Osnabrück (Germany): LoKS-Project

results (phrase detection) ^{3/3}

- candidate set with $f > 6$
 - does not contain any collocations according to Breidt (human annotators)
 - a lot of compositional compounds
 - only a few non-compositional translations
- useless to apply collocativity measure

University of Osnabrück (Germany): LoKS-Project

results (collocativity measure) ^{1/6}

- manually aligned phrase pairs
 - 250 phrase pairs
 - 83 with non-compositional translation
 - 45 with non-compositional semantics (Breidt's definition of collocation)
 - agreement of two annotators
 - 31 unresolved disagreements

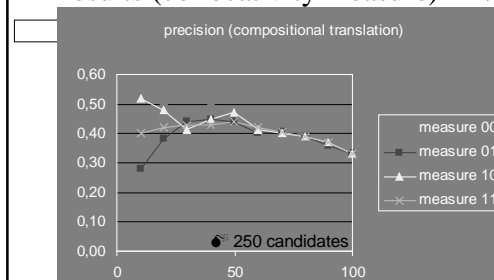
University of Osnabrück (Germany): LoKS-Project

results (collocativity measure) ^{2/6}

| variant | ignores words with high f | uses length of phrases |
|---------|---------------------------|------------------------|
| 00 | no | only if very different |
| 01 | no | always |
| 10 | yes | only if very different |
| 11 | yes | always |

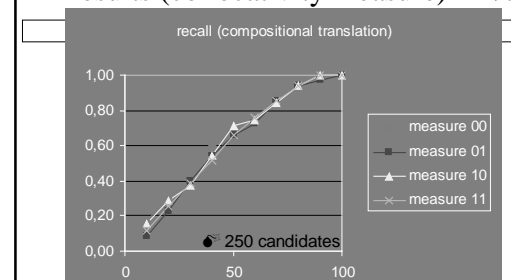
University of Osnabrück (Germany): LoKS-Project

results (collocativity measure) ^{3/6}

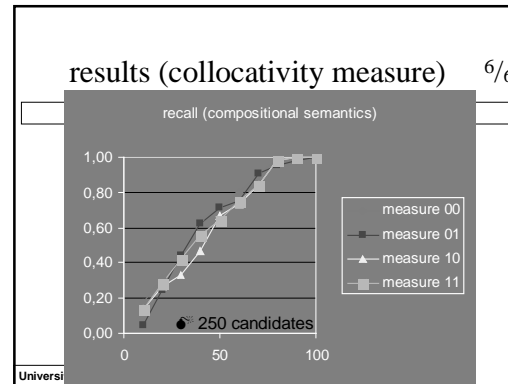
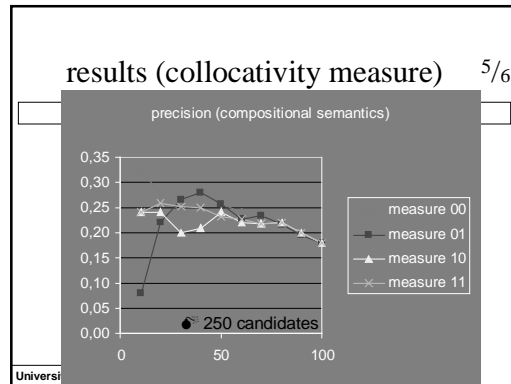


University of Osnabrück (Germany): LoKS-Project

results (collocativity measure) ^{4/6}



University of Osnabrück (Germany): LoKS-Project



- ### outlook 1/2
- **improve phrase correspondences**
 - use proper chunking to find phrases
 - use word alignment
 - **weight phrase pairs according to their correspondence probability**
 - **replace simple counts with advanced statistics (associations measure)**
 - **exploit substring relations among phrases**
- University of Osnabrück (Germany): KoS-Project

- ### outlook 2/2
- **improve collocativity measure**
 - decompose composita
 - find translation equivalences across word classes
 - better combine the different parts
- University of Osnabrück (Germany): KoS-Project

discussion / questions / contact

- **Norman Kummer, norman@VauDePe.de**
- **Joachim Wagner, jowagner@uos.de**

University of Osnabrück
 Institute of Cognitive Science
 49078 Osnabrück
 Germany

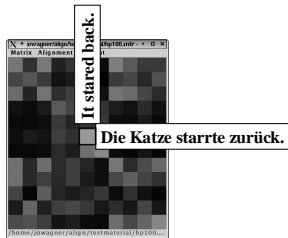
Link:

<http://www.cl-ki.uos.de/~koks/>

University of Osnabrück (Germany): KoS-Project

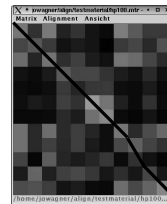


alignment of sentences (extra 1)



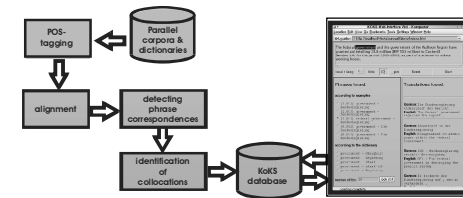
University of Osnabrück (Germany): iKoKS-Project

alignment of sentences (extra 2)



University of Osnabrück (Germany): iKoKS-Project

system overview



University of Osnabrück (Germany): iKoKS-Project

application

1/2

- CALL-context
- provides help to L2 learner in text understanding
- web based interface

University of Osnabrück (Germany): iKoKS-Project

application

2/2

- current KoKS demo application (screen-shot)

He is the **only** one.

local 0 lang: hits: 14

Phrases found:

according to examples

1 (1.000) only one - Einsige
 * (0.992) only one - der Einsige

according to the dictionary

only - lediglich
 only - nur
 only - nur bloss
 only - einzig
 only - nur
 only - nur
 only - nur
 only - einzig
 only - nur
 only - nur
 receive-only - nur zum Empfang
 only - einzig

number of hits: 10

Translations found:

German: # He ist er nicht der Einsige!
 English: # He is not the only one!

German: Hgqid war der Einsige, der die 98 Briefe schickte.
 English: Hgqid was the only one who ever sent him letters.

German: # Aber ich bin nicht der Einsige, der davon weiß.
 English: # But I am not the only one who knows.

German: Und ich war nicht der Einsige, den sie ohne Prozess sofort den Demontoren ausgeliefert haben.
 English: And I was not the only one who was handed straight to the demonters without trial.

University of Osnabrück (Germany): iKoKS-Project

other possible applications

- intelligent lexicon lookup (iKoKS)
- translation memory in CAT (computer assisted translation)
- full text search based on the lemmas

University of Osnabrück (Germany): iKoKS-Project