CROCO

LINGUISTIC PROPERTIES OF TRANSLATIONS
A CORPUS-BASED INVESTIGATION FOR THE LANGUAGE PAIR ENGLISH-GERMAN

# Querying multi-layer Annotation and Alignment in Translation Corpora

Mihaela Vela*, Stella Neumann*
& Silvia Hansen-Schirra°

*Saarland University, Saarbrücken

°Johannes Gutenberg University, Mainz

Project No. STE 840/5-1 sponsored by Deutsche Forschungsgemeinschaft DFG

# Overview

- Motivation
- The CroCo corpus
- Application-oriented queries
- Research-oriented queries
- Conclusion and Outlook

# Motivation

## Corpora in translation training and practice

- Terminology look-up (Pearson 2000, Maia 2003)
- Collocations (Teubert 2001, Barlow 2000)
- Idiomatic language use (Johansson & Hofland 2000, Vintar & Hansen 2005)
- Register-/typology-specific patterns (Pearson 2003, Bowker 1999)
  - → Mainly working with raw data thus restricted to features that are accessible to string-based queries
  - → More abstract features (e.g. grammatical functions) require annotation

# Motivation

## Register in translation studies

- Register analysis (Halliday & Hasan 1989)
- Operationalizations for register analysis in translation (Steiner (1997, 1998, 2004a,b)
  - → Not ready for quantitative exploitation of corpora
- Studies of individual registers (e.g. popular scientific writing (Teich 2003), travel guide books (Neumann 2003)) in translation
- Cross-lingual register variation (Biber 1995)
  - → Without reference to translation
  - → Features specifically selected for variation of speech and writing, thus not a comprehensive feature catalogue
  - → Theoretical framework rather vague

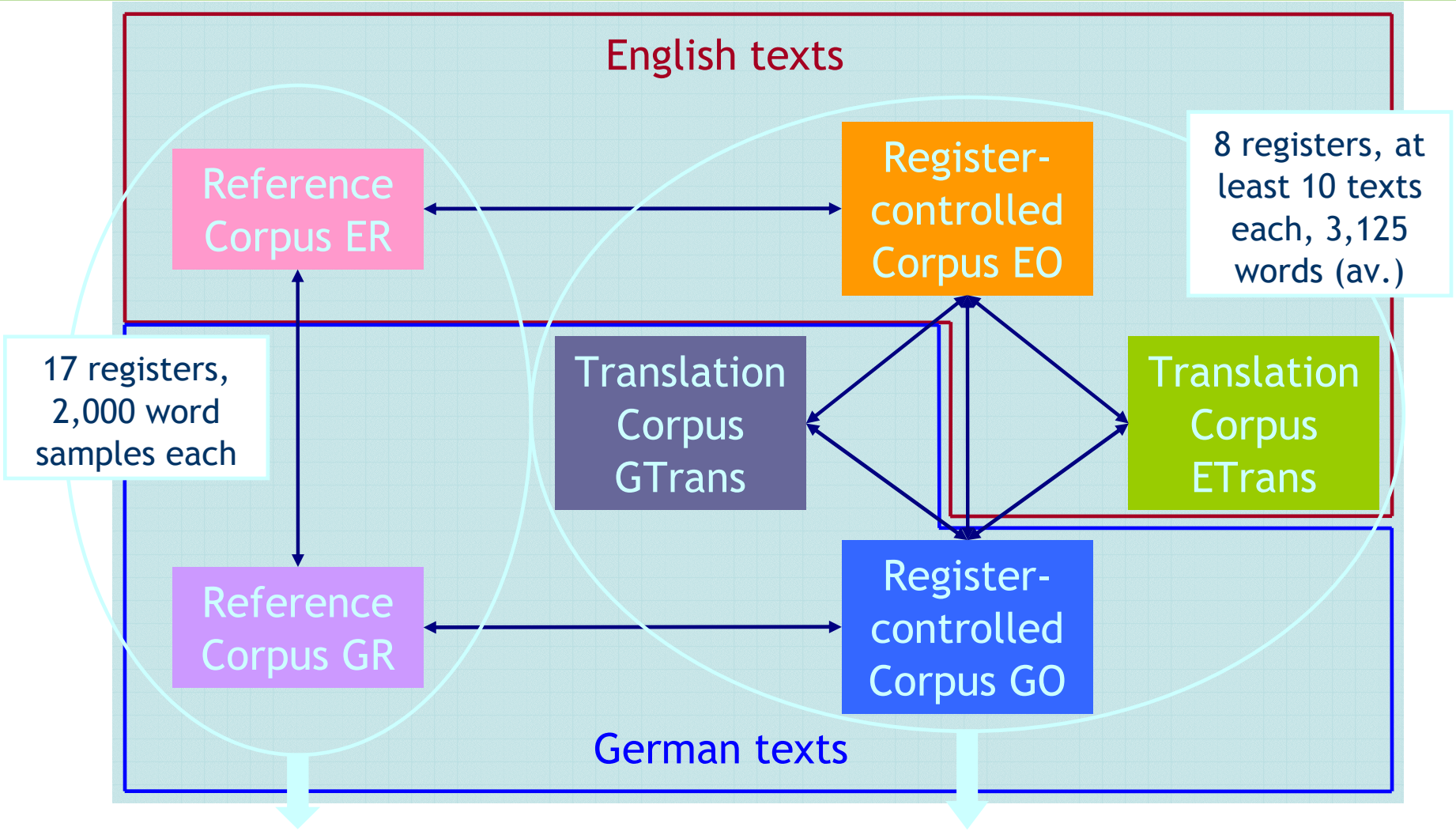
# The CroCo Corpus
## Overview 1

- Analysis of **corpora**, not examples
  → measurable linguistic evidence to establish properties
- Analysis of different translation relevant **registers**
  → register-specific distinctions available
- **Aligned** source and target texts
  → to distinguish between comparable and parallel texts
- **Annotation** of the corpora
  → to analyse lexico-grammatical and cohesive indicators

Basic Principle of the study
theory-neutral design, but theory-driven deduction of indicators

# The CroCo Corpus

## Overview 2 (cf. Neumann & Hansen-Schirra 2005)



# The CroCo Corpus

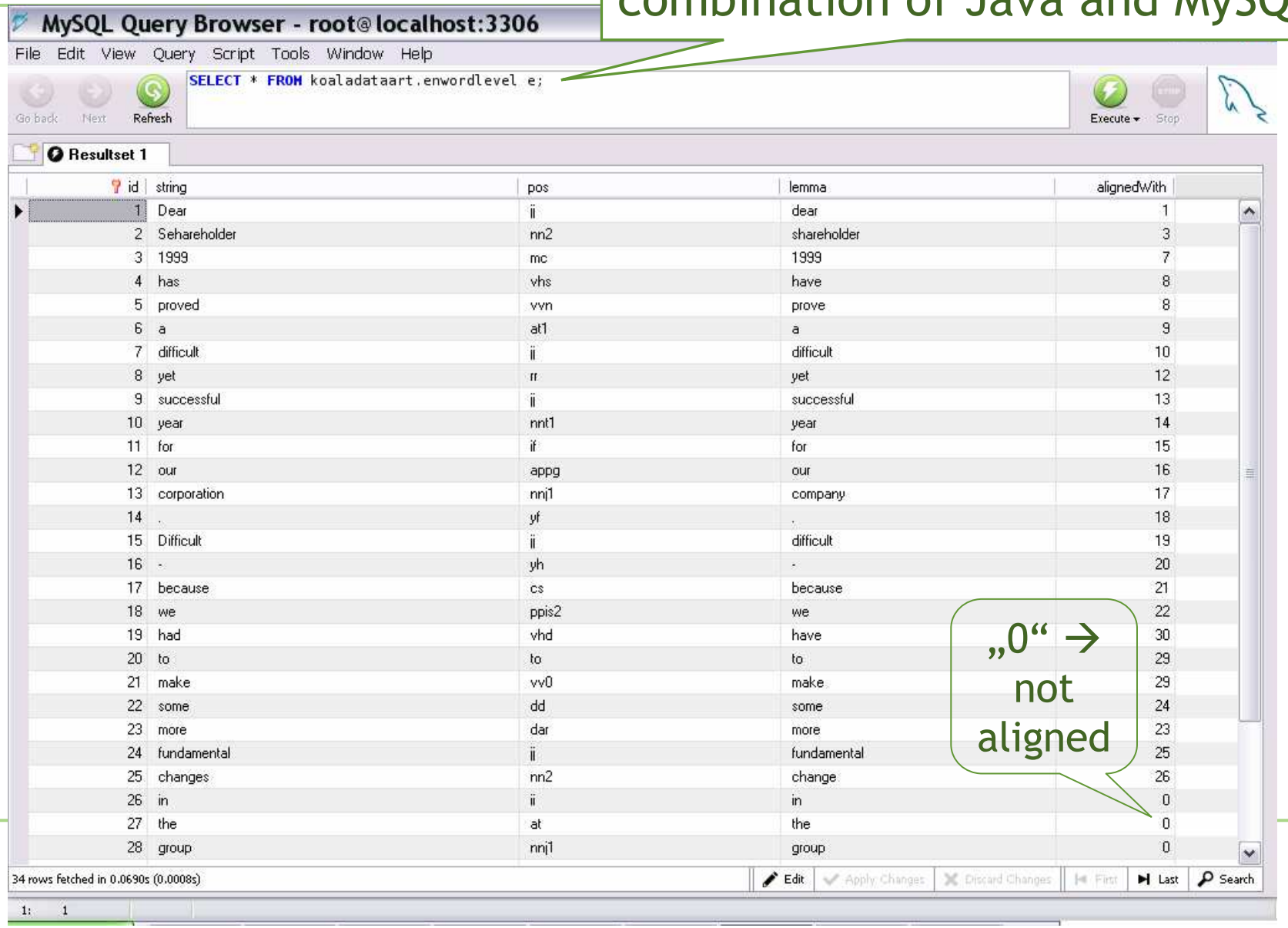
## Multi-layer annotation and alignment

- Annotation
  - Metainformation
  - Tokenization, PoS-tagging, morphology
  - Phrase structure, grammatical functions
- Alignment
  - Word, chunk, clause and sentence level
- Representation
  - XML multi-layer stand-off
  - Connection between the files via Xlink/Xpointer and xml:base attributes
  - Conversion into a MySQL database

# The CroCo Corpus

## Conversion into a database

Complex queries written in a combination of Java and MySQL




| id | string | pos | lemma | alignedWith |
|---|---|---|---|---|
| 1 | Dear | jj | dear | 1 |
| 2 | Sehareholder | nn2 | shareholder | 3 |
| 3 | 1999 | mc | 1999 | 7 |
| 4 | has | vhs | have | 8 |
| 5 | proved | vvn | prove | 8 |
| 6 | a | at1 | a | 9 |
| 7 | difficult | jj | difficult | 10 |
| 8 | yet | rr | yet | 12 |
| 9 | successful | jj | successful | 13 |
| 10 | year | nnt1 | year | 14 |
| 11 | for | if | for | 15 |
| 12 | our | appg | our | 16 |
| 13 | corporation | nnj1 | company | 17 |
| 14 | . | yf | . | 18 |
| 15 | Difficult | jj | difficult | 19 |
| 16 | - | yh | - | 20 |
| 17 | because | cs | because | 21 |
| 18 | we | ppis2 | we | 22 |
| 19 | had | vhd | have | 30 |
| 20 | to | to | to | 29 |
| 21 | make | vv0 | make | 29 |
| 22 | some | dd | some | 24 |
| 23 | more | dar | more | 23 |
| 24 | fundamental | jj | fundamental | 25 |
| 25 | changes | nn2 | change | 26 |
| 26 | in | ii | in | 0 |
| 27 | the | at | the | 0 |
| 28 | group | nnj1 | group | 0 |

„0“ → not aligned

# Application-oriented queries
## Overview

Detect existing solutions for grammatical translation problems on the basis of language typological differences (cf. Hawkins 1986)

- Raising constructions
  English accommodates more raising than German
- Cleft constructions
  Available in both languages but more frequent in English, because German has other options for focussing elements
- Substitutions
  Very restricted in German
- Deletions
  English more amenable to deletions than German

# Application-oriented queries
## Raising constructions

In EN source text: grammatical function=”finite verb” (FOLLOWED BY grammatical function=”direct object” (REALISED THROUGH phrasal category=”clause”))

We continue to benefit from the strong natural gas market in North America. --- Wir profitieren weiterhin von einem starken Erdgasmarkt in Nordamerika.

We defined the minivan, and will continue to do so. --- Wir haben den Minivan erfunden und wir werden auch künftig neue Marktsegmente definieren.

Finite verb translated as time adverbial

… and attracting the best talent possible as we continue to grow our business. --- … und werben zur Erweiterung unseres Geschäftes die besten Talente an, die wir nur finden können.

Nominaliza-tion

# Application-oriented queries
## Cleft constructions

In EN source text: word=”it” FOLLOWED BY lemma=”be” (FOLLOWED BY gram-matical function=”complement” (INCLUDING part-of-speech=”relative pronoun”))

It is this ownership that we truly believe helped our employees to drive toward success, despite the challenges of this year. --- Mit dieser Beteiligung am Unternehmen im Rücken haben unsere Mitarbeiter nach unserer Überzeugung maßgeblich zum Erfolg des Unternehmens trotz der großen Herausforderungen dieses Jahres beigetragen.

Fronted adverbial in the form of a PP

# Application-oriented queries
## Substitutions and deletions

In DE target text: phrasal category=”prepositional phrase/noun phrase” NOT INCLUDING part-of-speech=”noun”

After the interviews, I told our employees that I wanted Baker Hughes to improve from being a good company to become a great one. --- Nach den Gesprächen sagte ich den Mitarbeitern, dass ich Baker Hughes von einer guten Firma zu einer erstklassigen machen wolle.

Deletion replaces substitution

In DE target text: phrasal category=”sentence” INCLUDING 2 * grammatical function=”finite verb” AND 1* grammatical function=”subject”

We want to thank shareholders for your confidence, and we will continue to do everything possible to reward that confidence. --- Wir möchten den Aktionären für das uns entgegengebrachte Vertrauen danken und werden weiterhin alles Erdenkliche tun, dieses Vertrauen zu belohnen.

Repeated subject deleted

# Research-oriented queries

## Overview

### Register variables (Halliday & Hasan 1989)

- referential meaning → *Field of discourse*
- pragmatic aspects → *Tenor of discourse*
- textual means → *Mode of discourse*

### Operationalization necessary

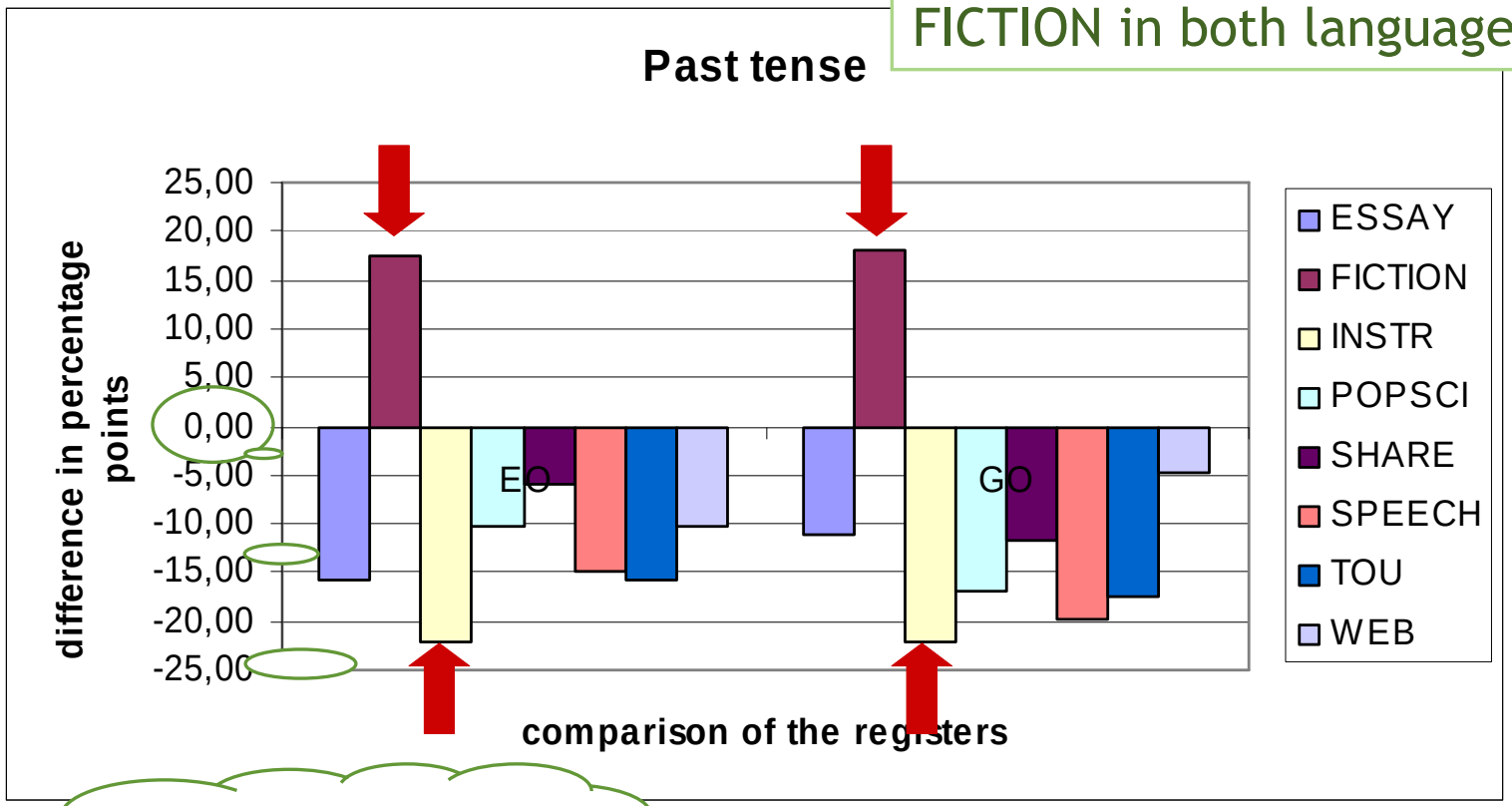
tenor of discourse › social hierarchy › level of expertise › grammar typical of languages for specific purposes (LSP)

### Result: comprehensive text analysis

# Research-oriented queries

## Field of discourse > goal orientation > past tense

→ Positive feature of FICTION in both languages




Comparable value of the reference corpora

→ Negative feature of INSTR in both languages

# Research-oriented queries

## Tenor of discourse > social hierarchy > LSP grammar

| SHARE | GO | Etrans | Diff. | EO | Gtrans | Diff. |
|---|---|---|---|---|---|---|
| no. of sentences | 1,734 | 1,738 | 4 | 1,489 | 1,467 | -22 |
| no. of clauses | 2,931 | 3,797 | 866 | 3,649 | 3,097 | -552 |
| no. of chunks | 9,353 | 8,602 | -751 | 7,251 | 8,400 | 1,149 |
| no. of words | 35,223 | 39,493 | 4,270 | 35,814 | 36,370 | 556 |
| chunks per sentence (av.) | 5.39 | 4.95 | -0.44 | 4.87 | 5.73 | 0.86 |
| chunks per clause (av.) | 3.19 | 2.27 | -0.93 | 1.99 | 2.71 | 0.73 |
| clauses per sentence (av.) | 1.69 | 2.18 | 0.49 | 2.45 | 2.11 | -0.34 |
| words per sentence (av.) | 20.31 | 22.72 | 2.41 | 24.05 | 24.79 | 0.74 |
| words per clause (av.) | 12.02 | 10.40 | -1.62 | 9.81 | 11.74 | 1.93 |
| words per chunk (av.) | 3.77 | 4.59 | 0.83 | 4.94 | 4.33 | -0.61 |
| sentences per text (av.) | 157.64 | 158.00 | 0.36 | 114.54 | 112.85 | -1.69 |
| clauses per text (av.) | 266.45 | 345.18 | 78.73 | 280.69 | 238.23 | -42.46 |
| chunks per text (av.) | 850.27 | 782.00 | -68.27 | 557.77 | 646.15 | 88.38 |

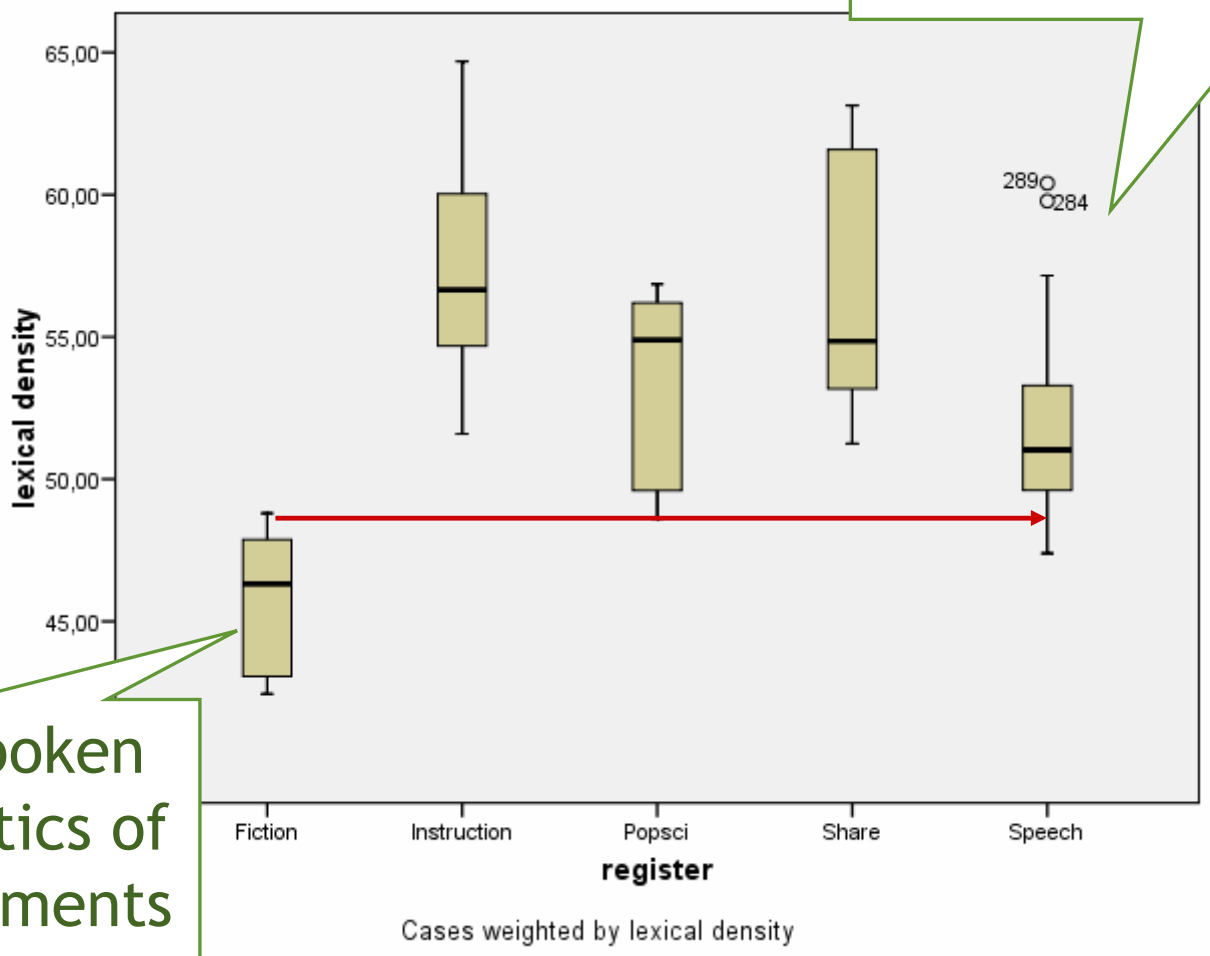
In comparison with other registers specialized registers should contain fewer clauses per sentence and more words per chunk

# Research-oriented queries

## Mode of discourse > medium > lexical density

Assumption:

Low lexical density points to spoken registers




„WRITTEN-to-be-spoken“

Reflects spoken characteristics of dialogic elements

# Conclusion and outlook

- Just a few examples out of the many possible queries
- Wealth of information available on the basis of linguistic enrichment of corpora
  - Standard queries for translation problems due to contrastive differences
  - Theory-based register profiles available combining top-down and bottom-up methodology
- Future work
  - Finish up annotation and alignment
  - Add semantic annotation
  - Create a query interface

# CroCo Project Web Site

http://fr46.uni-saarland.de/croco/