Ellipses in an English-German corpus – examining a chameleon concept?

Katrin Menzel, k.menzel@mx.uni-saarland.de

Institute of Applied Linguistics, Translation and Interpreting, Saarbrücken, Germany

Ellipses as a cohesive device - a case study within the DFG project "German-English contrasts in cohesion – Towards an empirically-based comparison (GECCo)"

GECCo project context and objectives: identifying contrasts in the realization of cohesion

- across languages (English vs. German)
- across registers (different text types and communication scenarios along the written-spoken continuum)
- across production types (originals vs. translated texts)

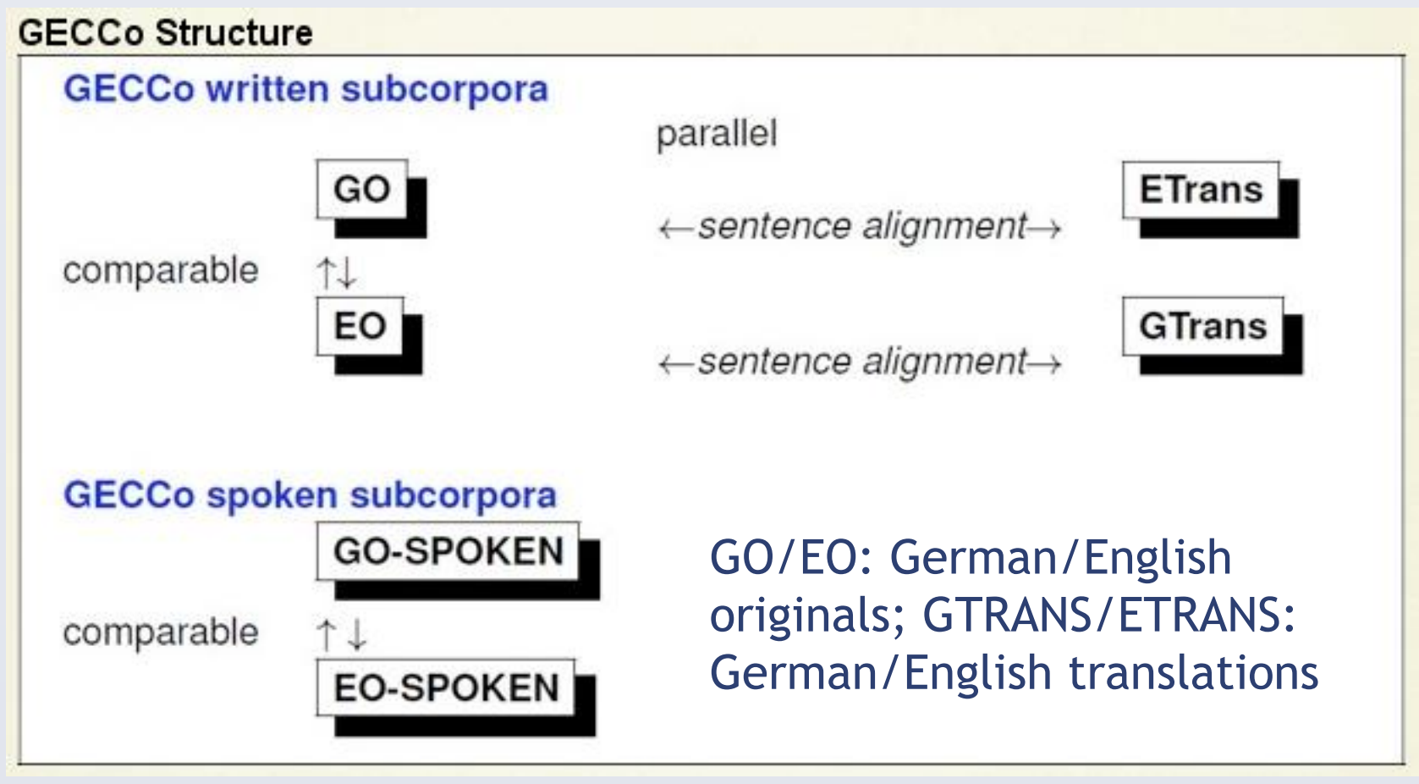
Cohesive devices

lexico-grammatical ties across texts: reference, conjunction, substitution, lexical cohesion, ellipsis (cf. Halliday&Hasan 1976)

CORPUS RESOURCES

GECCo corpus: multilevel-annotated bilingual corpus (ca. 1.44 m. tokens) - fictional texts, political essays, instruction manuals, popular science texts, letters to shareholders, prepared speeches, tourism leaflets, corporate websites, academic lectures, interviews (+ recently compiled registers: talk shows, internet forums, medical consultation, sermons)

- written registers: sentence-aligned parallel corpus
- spoken registers: comparable corpus



GO/EO: German/English originals; GTRANS/ETRANS: German/English translations

subcorpora	registers
written	imported from CroCo*
EO 🇬🇧	FICTION, ESSAY
GO 🇩🇪	INSTR, POPSCI
Etrans 🇬🇧→ 🇩🇪	SHARE, SPEECH
GTrans 🇩🇪→ 🇬🇧	TOU, WEB
spoken	collected at FR4.6, UdS**
EO-SPOKEN 🇬🇧	INTERVIEW, ACADEMIC
GO-SPOKEN 🇩🇪	*FORUM, TALKSHOW*

* cf. Hansen-Schirra et al. 2013 / http://www.gecco.uni-saarland.de

GECCo annotation levels

- word: word, lemma, POS
- chunk: sentences, syntactic chunks, clauses
- extralinguistic: register analysis, speaker information
- cohesion: e.g. semi-automatic annotation of co-reference, conjunction, substitution (cf. Lapshinova-Koltunski & Kunz 2014)

annotated corpus is available in XML format, can be queried with CQP (Evert, 2005), additional CQPweb version (https://fedora.clarin-d.uni-saarland.de/cqpweb/)

ELLIPSIS AS A CASE STUDY – METHODOLOGY

i) Developing systematic and fine-grained conceptualization of ellipsis as a cohesive device aiming at cross-linguistic applicability of annotation scheme categories (Menzel 2014 a/b/c)

- ellipsis has been described as a **'chameleon concept'** as linguists have produced rather heterogeneous definitions so that various constructions and discourse phenomena have been subsumed under this category
- **main categories of (potentially cohesive) ellipsis relevant for this study:**

Nominal ellipsis: omission of specific element of noun phrase (head noun) - e.g. *"There are many reasons why Britain is good for Europe. Let me choose just four [ ]."*

Verbal ellipsis: ellipsis within the verbal group (modal/auxiliary/operator or lexical verb, often accompanied by the omission of related elements such as objects) - e.g. *"A little town that is often missed by travellers but shouldn't be [ ]."*

Clausal ellipsis: omission of a part of a clause (broadest subcategory) - e.g. *"Has he brought you presents? What kind of presents [ ]?"*

Co-occurrence of nominal+verbal/clausal - e.g. *"How many slices do you want?" - "[ ] Two [ ]."*

- cohesive ellipses refer endophorically to textual antecedents (ideally not in the same clause so that a textual link between different clauses or sentences is created)

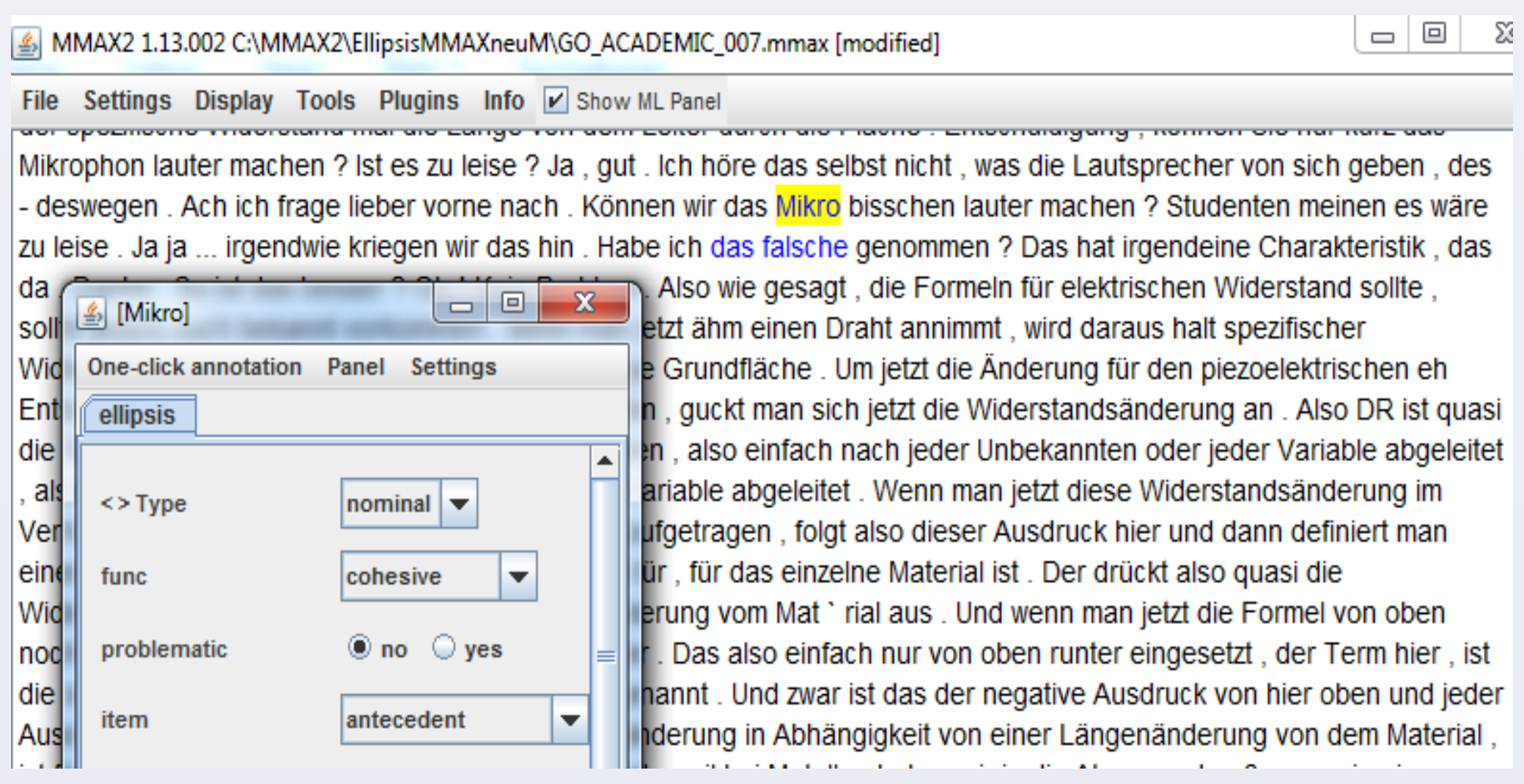
ii) Annotation of endophoric, text-forming ellipses (remnants) and their antecedents with MMAX2 based on annotation guidelines

- ellipsis remnants and their antecedents are annotated with open-source tool MMAX2: nominal ellipsis, verbal ellipsis, clausal ellipsis, nominal+verbal/clausal (mixed)

within these categories, it can be distinguished between:

- non-cohesive (e.g. exophoric, situational)
- cohesive (cross-clausal reference to antecedent)
- clause-internal

- problematic cases that might require further discussion, e.g. borderline cases or ambiguous structures can be treated separately -> annotated as "problematic"


MMAX2 1.13.002 C:\MMAX2\EllipsisMMAXneuM\GO_ACADEMIC_007.mmax [modified]

File Settings Display Tools Plugins Info ☑ Show ML Panel

Mikrophon lauter machen ? Ist es zu leise ? Ja , gut . Ich höre das selbst nicht , was die Lautsprecher von sich geben , des
- deswegen . Ach ich frage lieber vorne nach . Können wir das Mikro bisschen lauter machen ? Studenten meinen es wäre
zu leise . Ja ja ... irgendwie kriegen wir das hin . Habe ich das falsche genommen ? Das hat irgendeine Charakteristik , das
da . Also wie gesagt , die Formeln für elektrischen Widerstand sollte ,
soll etzt ähm einen Draht annimmt , wird daraus halt spezifischer
Wid e Grundfläche . Um jetzt die Änderung für den piezoelektrischen eh
Ent n , guckt man sich jetzt die Widerstandsänderung an . Also DR ist quasi
die en , also einfach nach jeder Unbekannten oder jeder Variable abgeleitet
, als ariable abgeleitet . Wenn man jetzt diese Widerstandsänderung im
Ver ufgetragen , folgt also dieser Ausdruck hier und dann definiert man
eine ür , für das einzelne Material ist . Der drückt also quasi die
Wid erung vom Mat ` rial aus . Und wenn man jetzt die Formel von oben
noc r . Das also einfach nur von oben runter eingesetzt , der Term hier , ist
die nannt . Und zwar ist das der negative Ausdruck von hier oben und jeder
Aus nderung in Abhängigkeit von einer Längenänderung von dem Material ,

[Mikro]

One-click annotation Panel Settings

ellipsis

<> Type: nominal

func: cohesive

problematic: ● no ○ yes

item: antecedent

- some additional annotation categories cover other types of fragments, non-clausal units and omission phenomena that might look similar to ellipsis (e.g. anacoluthon, headlines…) but actually need different analysis
- manual annotation currently more accurate than automatic methods (numerous theoretical omission possibilities in different syntactic environments and POS tagging sometimes wrong in ellipsis environments due to deficient/non-standard syntax)
- annotations can be used to identify typical syntactic patterns of ellipsis contexts to improve (**semi-)automatic annotation methods** for certain types (e.g. nominal e. after adjectives, clausal e. in question-answer pairs)

iii) Data extraction and interpretation

some frequency distributions:

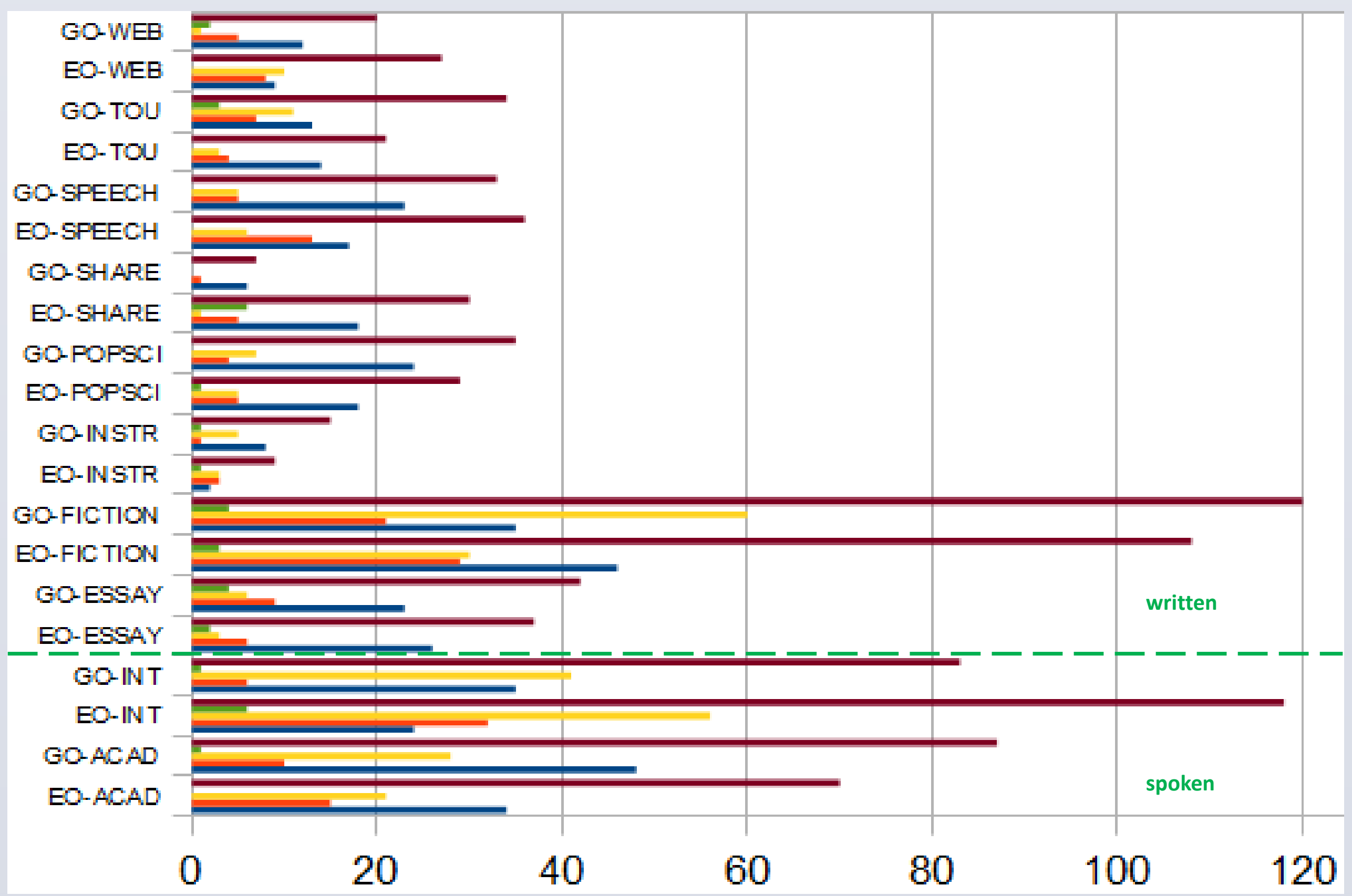


Fig. 1: Cohesive ellipses in English & German original texts (written: websites, tourism leaflets, prepared speeches, letters to shareholders, popular science texts, instruction manuals, political essays, fictional texts; spoken: interviews, academic lectures - abs. values, comparable size of registers: ca. 30.000 tokens & ca. 10 texts each)

Legend Fig. 1 & 3:

- ■ nominal
- ■ verbal
- ■ clausal
- ■ mixed
- ■ total —

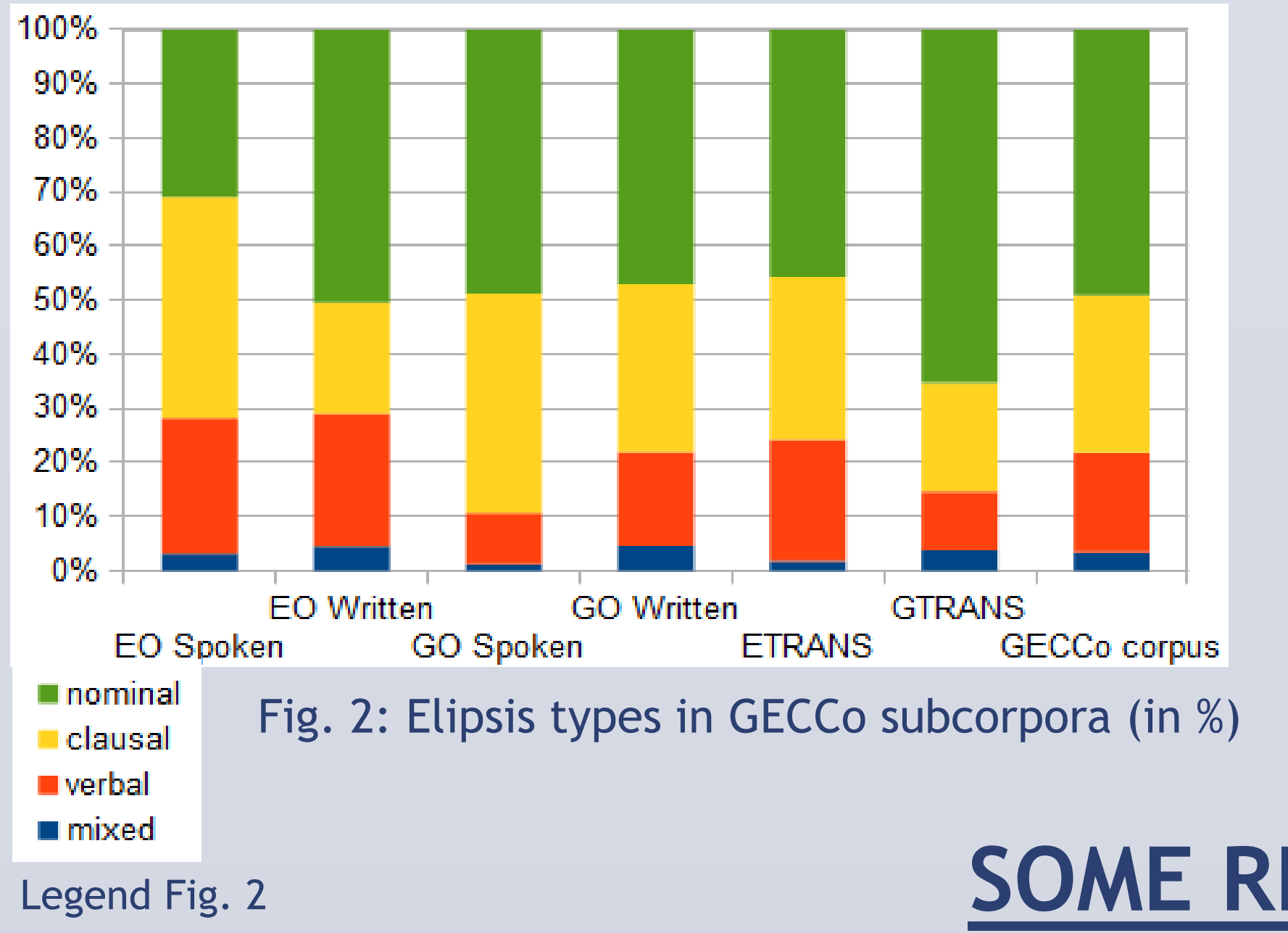


Fig. 2: Elipsis types in GECCo subcorpora (in %)

Legend Fig. 2

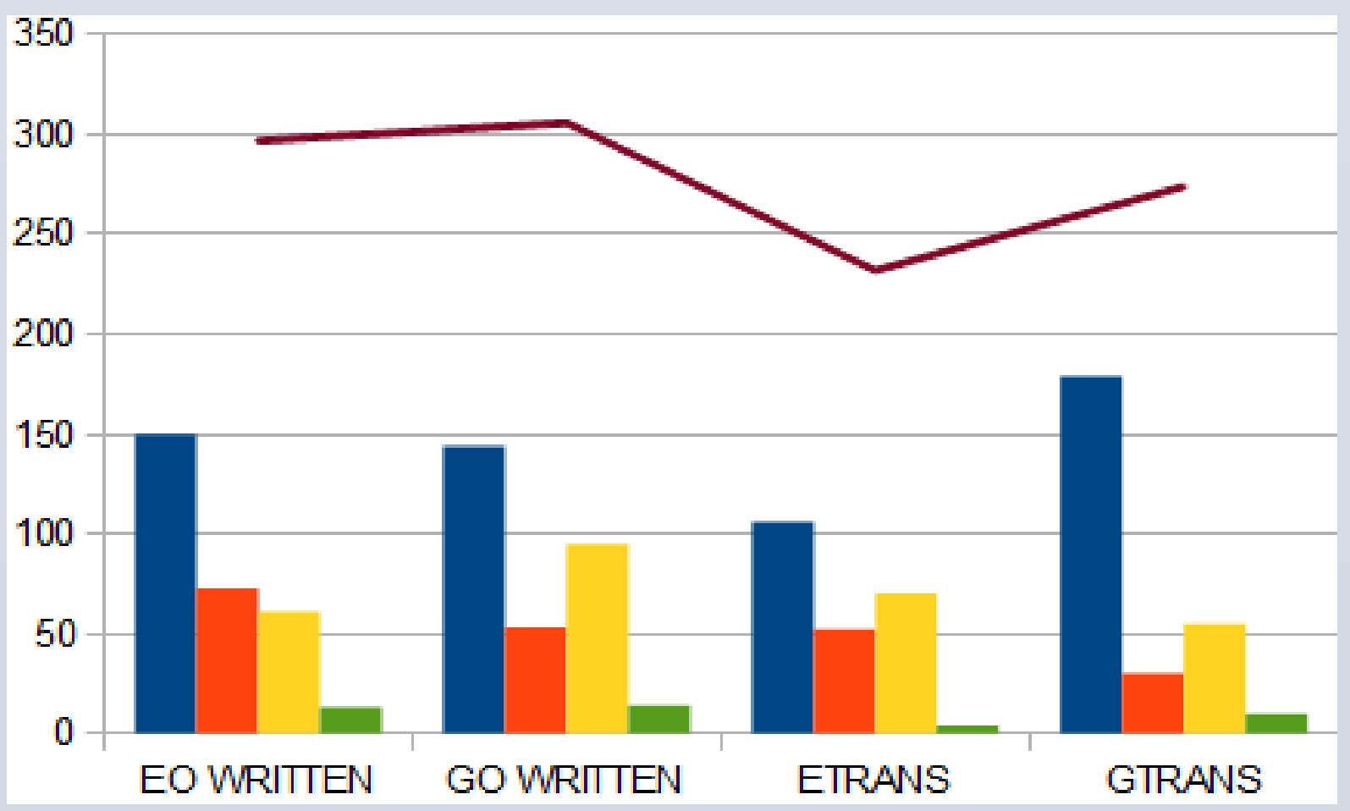


Fig. 3: Originals vs. translations (abs. values)

SOME RESULTS

- differences between certain registers and between written and spoken language greater than those between English and German in general
- ellipsis frequency lower in translations compared to that in originals of the same language

REFERENCES

- Evert, S. 2005. The CQP Query Language Tutorial. IMS, Stuttgart University.
- Halliday, M.A.K. & R. Hasan. 1976. Cohesion in English. London/NY: Longman.
- Hansen-Schirra, S., S. Neumann & E. Steiner. 2012. Cross-linguistic Corpora for the Study of Translations. Insights from the language pair English-German. Series Text, Translation, Computational Processing. Berlin/NY: Mouton de Gruyter.
- Menzel, K. 2014a. Ellipsen als Stil- und Kohäsionsmittel in deutschen & englischen politischen Reden. In: Leuschner, T. & M. Koliopoulou (eds.). Germanist. Mitteilungen. Zeitschr. für Deutsche Sprache, Literatur & Kultur, 40.1: "Deutsch kontrastiv: 31-50.
-------- 2014b. Project Working Report on the Conceptualization and Annotation of Ellipses in English and German Corpus Texts (GECCo-Corpus) with MMAX2. -------- 2014c. Project Working Report: Ellipsis in English and German – Systemic Contrasts.
- Müller, C. & M. Strube. 2006. Multi-Level Annotation of Linguistic Data with MMAX2. In: Braun S., K. Kohn & J. Mukherjee (eds.): Corpus Technology & Language Pedagogy. New Resources, New Tools, New Methods Frankfurt: Peter Lang: 197-214.