K. Menzel

A CORPUS LINGUISTIC STUDY OF ELLIPSIS AS A COHESIVE DEVICE[1]

Abstract. In this paper, a cross-linguistic comparison is made across various registers / genres with regard to ellipsis as a cohesive device. It involves data extracted from an English-German translation corpus and aims to shed some light on the complex picture of ellipsis.

1. Introduction and definitions

The aim of this paper is to give an overview of some cross-linguistic aspects regarding ellipsis from a theoretical and a corpus linguistic perspective. There is a growing interest in tracing ellipsis automatically. In the past however, ellipsis as a cohesive device, i.e. creating links within a discourse or text, has been studied less extensively than other cohesive phenomena, probably due to its complexity and fuzziness. Furthermore, there are mainly monolingual accounts of ellipsis while corpus linguistic studies on ellipsis are rare and usually focus on specific subcategories (e.g. nominal ellipses after adjectives in English, Günther, 2012[2]). There are no tools yet to efficiently spot and annotate ellipsis automatically and it is not possible to query ellipses as such as they are empty elements. However, they occur in the environment of certain syntactical structures or trigger words. Therefore, systematic query patterns to spot ellipses in electronic corpora can be developed.

A quantitative and qualitative analysis of ellipsis as a cohesive device has to deal with the fact that it is a rather fuzzy concept and there are several definitions from various perspectives. It seems to be a gradual notion with prototypical and less typical or marginal cases and very similar other phenomena. Obviously, the analysis of data and

[1] Part of DFG-project: German-English contrasts in cohesion – towards an empirically-based comparison – GECCo.

[2] *Günther C.* The Elliptical Noun Phrase in English: Structure and Use. New York: Routledge. 2012.

outcome of studies on ellipsis will depend on its definition and subclassification. Ellipsis can be regarded as the omission of obligatory clause or phrase elements which must be recoverable in their precise form from either the immediate context or the surrounding co-text or on the basis of our knowledge of the grammar of English[3], an extreme form of phonological reduction or substitution by zero when something that is structurally necessary is left unsaid[4]. This paper builds on the SFL-based classification of cohesive devices by Halliday & Hasan (1976)[5] who divide ellipsis in English into nominal, verbal and clausal. This subcategorization also fits German relatively well despite some typological differences between the two languages. Nominal ellipsis is ellipsis within the NP, where usually the head noun is omitted and the modifier is upgraded to take its function. Here, omission of the full NP, such as predicative nominals, will also be subsumed under this category for practical reasons, but it might also fall under clausal ellipsis. Verbal ellipsis is ellipsis within the VP (operator, lexical or modal verb). The omission of clauses or clause elements, a variety of different structures, is called clausal ellipsis. Ellipsis with a cohesive function is different from other types of fragments: missing information must be supplied from the surrounding text, usually anaphorically and across sentences or clauses. There is no entire agreement in the literature as to where to draw the exact line to distinguish it from similar phenomena such as ellipses that are not or only marginally cohesive (e.g. lexicalized ellipsis (Ágel, 1991[6]), exophoric / situational ellipsis, right node

[3] *Collins P., Hollo C.* English Grammar – an Introduction. London: Macmillan. 2000. P. 155.

[4] *Winkler S.* Ellipsis and information structure in English and German: The phonological reduction hypothesis. Arbeitspapiere des Sonderforschungsbereichs 340, № 121. 1997.

[5] *Halliday M.A.K., Hasan R.* Cohesion in English. London: Longman. 1976. P.144.

[6] *Ágel V.* Lexikalische Ellipsen. Fragen und Vorschläge // Zeitschrift für germanistische Linguistik 19. 1991. P. 24–48.

raising, clause-internal phenomena) or from substitution (cf. Kunz & Steiner, 2013)[7].

2. Corpus resources and methods

This study is part of a larger DFG-funded research project (German-English contrasts in cohesion – towards an empirically-based comparison – GECCo)[8]. The GECCo corpus provides English and German texts of various registers along the written/spoken continuum. The written part of the corpus consists of English and German original texts that are aligned with their translation. GECCo is tagged for: tokens, lemmas, morpho-syntactic information, parts of speech, chunks and sentence boundaries. Annotation of ellipses and their antecedents is currently done with MMAX2, an open source annotation tool[9]. The corpus can be queried with CQP[10]. It is possible to formulate CQP-based queries to find potential candidates of ellipsis in the corpus.

[7] *Kunz K., Steiner E.* Cohesive substitution in English and German: a contrastive and corpus-based perspective. In: Aijmer, K. & Altenberg, B. eds. Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson. Amsterdam, 2013. P. 201–231.

[8] *Kunz K. & Lapshinova-Koltunski E.* Tools to Analyse German-English Contrasts in Cohesion // Proceedings of GSCL-2011, Hamburg, Germany. 2011; *Lapshinova-Koltunski E., Kunz K., Amoia M.* Compiling a Multilingual Spoken Corpus // Proceedings of GSCP-2012, Belo Horizonte, Brazil. Forthcoming; *Neumann S., Hansen-Schirra S.* The CroCo Project. Cross-linguistic corpora for the investigation of explicitation in translations // Proceedings from the Corpus Linguistics Conference Series (PCLC), 2005. Vol. 1 № 1. cf. **http://134.96.85.104/gecco/GECCo/Korpus.html** for GECCo corpus and structure

[9] *Müller C., Strube M.* Multi-Level Annotation of Linguistic Data with MMAX2. In: S. Braun, K. Kohn, J. Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt, 2006. P. 197–214.

[10] *Evert S.* The CQP Query Language Tutorial. IMS, Universität Stuttgart. 2005.

Table 1. Samples for CQP queries to find nominal ellipsis.

Pattern nominal ellipsis	**CQP query design**
1. nominal ellipsis after article / determiner / numeral / quantifier / possessive marker (+optional adjective)	e.g. in German subcorpora (Stuttgart-Tübingen-TagSet STTS) [pos='adja'][pos='vafin']; (adjective + finite verb); [pos='art'][pos='adja'][pos!='nn|ne']; (article + adjective, not followed by noun/proper noun) in English subcorpora (Penn Treebank tagset): [pos='jj'][pos='vv.*']; (adj. + verb)
2. possessive marker '*s* not followed by noun	[word='s'][pos!='nn|ne']

Queries to find nominal ellipsis are based on typical structures that might trigger them: e.g. articles / determiners / adjectives / numerals not followed by nouns, possessive markers not followed by nouns (cf. Table 1), subject ellipsis in coordination etc. Query patterns to find VPE with CQP may involve verbless clauses, lexical verb ellipsis in comparative constructions (*He can run faster than Jane can* [ ].), before conditional clauses (e.g. *I can* [ ] *if I want*.), etc. Clausal ellipsis can be found by querying adjacency pairs, clauses consisting of one or very few constituents or, in the case of sluicing, with queries of wh-words at the end of a clause. The queries can be adapted to fit German word order patterns. Some structures do not exist in German (operator ellipsis, substitution with do, inclusive imperative followed by verbal ellipsis etc.). The list of potential candidates for ellipsis found by corpus queries had to be disambiguated manually, particularly to exclude other types of ellipsis and fragments. Additionally, some entire registers were looked through manually to find all cases of cohesive ellipsis. This served as a comparison with ellipses found by CQP queries.

3. Preliminary Results

The CQP queries worked best for all subtypes of nominal ellipsis in both languages. CQP queries did not lead to a sufficient number of hits yet with regard to verbal ellipsis, and manual search confirmed that our corpus does not provide a high number of VPE in general. There was a high recall for nominal ellipsis, but unfortunately precision was still relatively low. Some subtypes of verbal and clausal ellipsis were easy to query with CQP but they were rare (e.g. sluicing) others were more difficult to spot. In general, verbal ellipsis, particularly in English, has a rather sophisticated system of possible subtypes, however some patterns were not found very frequently in our corpus (e.g. pseudogapping).

Nominal ellipsis occurs mainly in certain text types (e.g. texts with many adjectives and nominal style or limited space for printing) but also in the context of certain topics (involving numerals, comparisons, contrasts…). Written discourse in general offers more possibilities for nominal ellipsis due to structural complexity, lexical density, nominalization and longer noun groups. Nominal ellipsis is generally more frequent in German because English can use substitution with ‘one’ instead and avoids nominal ellipsis when it could lead to ambiguity due to the morphological characteristics of the language. Verbal ellipsis often co-occurs with proper names or personal pronouns and therefore also depends on the text topic and the level of interaction in discourse. Verbal ellipsis also typically occurs in text types with otherwise rich verb phrase structures to avoid verbal repetition; some subtypes require hypotactic or parallel structures, often involving contrasts between two or more members of a semantic category (e.g. *The parents ate cake, and the children* [ ] *cookies*). Due to the lack of exact correspondence between the English and German verbal system, there are more differences between English and German verbal ellipsis than with regard to nominal ellipsis.

Filler words, redundancies, anacolutha and less clear sentence boundaries in spoken language make queries for spoken registers more difficult than for written registers. Often larger parts of texts have to be taken into account to determine whether a certain structure is an ellipsis, an anacoluthon or a sentence break, regional variation or simply an error where people forgot to complete a sentence. The special syntax of spoken language and differences between English and German morphology (high frequency of zero derivation/word-class ambiguities in English, declension of adjective/ pronouns as ellipsis remnants in German) have to be considered in queries as well as the particularities of ellipses as syntactically incomplete or – without an appropriate context – even deficient structures. Tagging errors resulting from these untypical syntactic patterns are another aspect that has to be taken into account when formulating CQP corpus queries.

Table 2 shows the normalized frequencies of ellipsis subtypes per 100.000 words found via cqp queries and/or manually in 4 registers of GECCo.

Cohesive ellipsis in total seems to be more frequent in spoken than in written language. However, nominal ellipsis, which is the most frequent type among the categories mentioned in Table 2 in all registers (surprisingly more frequently in English), is found very often in fictional written texts (due to the high frequency of adjectives and similarity to spoken language, although admittedly fiction is a rather heterogeneous register). Ellipsis of head nouns is often used as a stylistic device in written registers to avoid redundancy.

Table 2. Cohesive ellipsis in 4 registers of GECCo (GO = German Originals, EO = English Originals; spoken registers: Interview + Academic lectures; written registers: Fiction/novels + Tourism leaflets

	Nominal ellipsis	**verbal**	**clausal**	**∑**
GO Interview	62.2	9.7	42.2	114.1
EO Interview	129.3	58.0	42.2	229.5
GO Academic	124.4	9.8	43.9	178.1
EO Academic	131.0	29.7	12.4	173.1
GO Fiction	114.2	38.1	51.7	204.0
EO Fiction	154.1	37.8	27.0	218.9
GO Tourism	24.6	13.7	16.4	54.7

EO Tourism	52.9	5.6	0	58.5

So far, clausal ellipsis was not found very frequently via CQP queries, and manual search also showed that it is not a typical phenomenon in our corpus. That is probably due to the fact that clausal ellipsis is typical for dialogues and spontaneous spoken language in both English and German. Our register of fiction has long narrative passages and few dialogues and the spoken register of academic lectures are rather monologic and quite formal whereas the register of interview has rather long passages before speaker turns take place and the texts had probably been prepared carefully in advance. A difference between English and German corpus registers is that English sometimes uses longer questions (including more hedging instruments and modal verbs as politeness strategies) and often has longer (echo) answers than German. In our spoken data, sometimes quite the opposite of ellipsis was observed: in structures where ellipsis would have been a possible strategy to avoid repetition, it was not used.

4. Conclusion and outlook

While it is generally assumed that ellipsis (e.g. exophoric/ situational) and fragments are a typical feature of spoken language, it is questionable whether the same is true for (all types of) cohesive ellipsis. Looking through the spoken registers of academic discourse and interviews in GECCo, it becomes intuitively obvious that ellipsis as a cohesive device is less frequent than expected in our spoken data. This infrequency can be explained by the fact that both academic lectures and prepared interviews texts are more or less similar to written language and that not all marginal and less typical subcategories of ellipsis were included in the analysis.

Generally speaking, English and German differ with respect to the types and properties of elliptical sentences that they allow. Possibly German allows more «deep anaphora» (antecedents are recovered from conceptual, rather than syntactic representations) vs. more «surface anaphora» in English (structural parallelism constraint, ellipsis and antecedent required to share the same syntactic structure). After analysing the phenomenon of ellipsis cross-linguistically in original texts, it might be interesting to look at what happens in translation, whether there is an effect of shining through[11]. It might be possible that English translations, for instance, include a higher frequency of nominal ellipsis after adjectives where we would normally expect *one*, e.g. The grey fox is not as flamboyant as the red [ ] or higher frequencies of *one* as a substitute where it is not necessary (e.g. after *next*, *second*, *another*, *which*). Therefore it might be possible to show that English is influenced by German through translations.

[11] *Teich E.* Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts. Berlin, 2003.