OVERVIEW ON THE ANNOTATION OF ELLIPSES IN ENGLISH AND GERMAN CORPUS TEXTS (GECCO-CORPUS) WITH MMAX2

This paper gives an overview on how ellipses have been manually annotated in GECCo with the annotation tool MMAX2 (more details on the conceptualisation and annotation of ellipses and related phenomena are set out in the related project working report of the GECCo project (WP2.1/2.4)

Contents

1. Introduction

Halliday/Hasan's (1976) distinction between nominal, verbal and clausal ellipses as omissions that potentially refer to textual antecedent is taken as a starting point for the development of the theoretical framework for this study.

Nominal ellipsis is the omission of the head noun within the nominal group. Therefore, it is the omission of one specific element of the noun phrase. **Verbal ellipsis** is ellipsis within the verbal group. It is optional which element can be left out (modal/auxiliary/operator and/or lexical verb). Verbal ellipsis is often accompanied by the omission of related clause elements such as objects. **Clausal ellipsis** is the broadest subcategory and contains omissions that are not covered under nominal and verbal ellipsis yet. It is defined as the omission of a clause, a part of a clause or an element of a clause (a constituent). It may co-occur with nominal or verbal ellipsis. Halliday/Hasan understand ellipsis primarily as a syntactic phenomenon. Some material is deleted in a sentence and it is licensed by structural identity between the elided element/constituent and its antecedent.

An advantage of utilizing such relatively broad main categories of ellipsis as Halliday/Hasan do is that this provides a suitable framework for a cross-linguistic analysis of ellipses with possible textual antecedents as both English and German have NPs, VPs and clauses where certain elements can be omitted that are deducible from the co-text. The classifications used for the annotation should **not overlap or use gradual categories**. The aim is to place all cases found in the corpus clearly in only one category in order to provide the basis for a meaningful quantitative analysis.

We define ellipsis as a phenomenon where the ellipsis remnant of a syntactic omission is deliberately left grammatically incomplete or deficient to create a sentence fragment or an incomplete phrase. The content of the ellipsis site can be recovered from its **textual antecedent**. What is important for the classification of our ellipsis categories is the underlying sentence / phrase with a complete syntactic / phrasal structure assuming that there is unpronounced material. It has to be determined which cases of ellipsis are used as **text-forming, cohesive devices**, i.e. which cases of nominal, verbal and clausal ellipses endophorically establish textual links.

According to Halliday/Hasan, "cohesion occurs where the interpretation of some element in the discourse is dependent on that of another" (1976: 4). Although Halliday/Hasan state that "[…] cohesive relations are the same whether their elements are within the same sentence or not. […]" (ibid.: 9), all the examples they give are of cohesion **across sentence boundaries** assuming that in those cases the effect is more striking. Halliday/Hasan clearly state that cohesion "is a relation to which the sentence, or any other form of grammatical structure, is […] irrelevant" (1976:9) and that cohesive ties between sentences stand out more clearly because they are the ONLY source of texture whereas within the sentence there are structural relations as well. Therefore the cohesive effect is less pronounced within the sentence (ibid.: 9).

In our analysis of cohesion in GECCo, we specifically look at cases of **ellipses across sentences, but also at ellipses across clauses and across utterance boundaries** although all of those types are sometimes difficult to pin down, particularly in spoken registers. These cases have to be distinguished from ellipsis with **exophoric reference** to the extralinguistic situational context and from **locally bound omissions** referring to antecedents within the same clause or even the same phrase that are mainly the result of intra-clausal grammatical rules. Furthermore, the cases identified as cohesive ellipses have to be distinguished from **other types of fragments and non-clausal units**, other omission phenomena and non-ellipses that might look similar to the categories mentioned above but actually need different

analyses.

The following categories are relevant for the annotation of the corpus with regard to (potentially) cohesive ellipses:

- **nominal ellipsis**
- **verbal ellipsis**
- **clausal ellipsis**
- **mixed cases**

Additionally, it is possible to annotate the following categories that might superficially look very similar to the categories mentioned above. However, there are a various reasons for treating them as specific separate structures.

- **texttype specific ellipses and fragments**
- **sentence splits**
- **short yes / no replies where a particle as conveys affirmation or negation**
- **non-clausal units that are not considered to be the result of syntactic omission**
- **other**

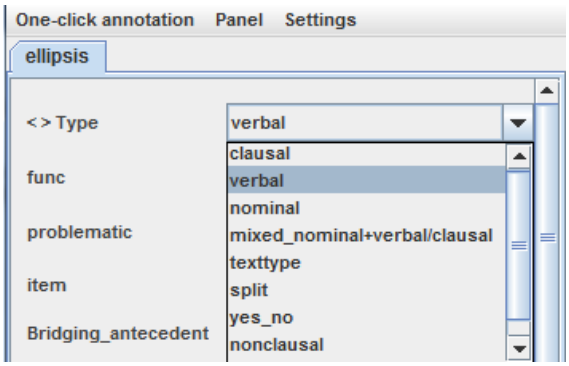

Fig. 1: Annotation of ellipsis types and similar categories in MMAX2

Within these categories, it can be distinguished between:

- **non-cohesive**
- **cohesive**
- **clause-internal**

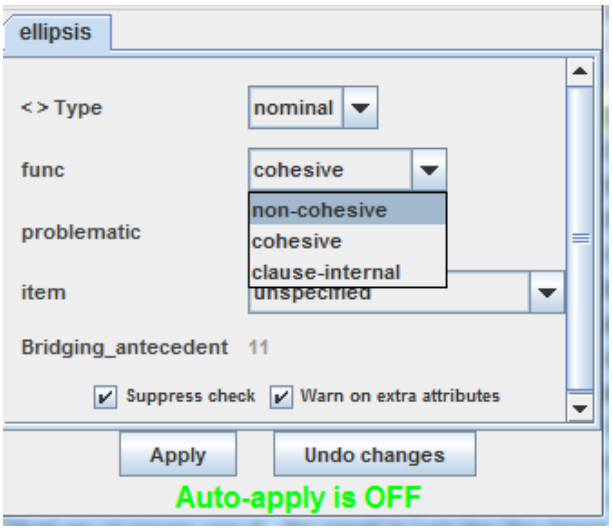

Fig. 2: Annotation of function of ellipsis in MMAX2

A cohesive ellipsis refers endophorically to a textual antecedent. Ideally, the antecedent does not occur in the same clause so that a textual link between different clauses or sentences is created. **Ellipsis remnants (not the omitted elements/the ellipsis sites) and their antecedents** are both annotated in MMAX2 and a **pointer relation** is used to link a remnant to its antecedent. The antecedent is marked under the level "item" as "antecedent" (or „antecedent_ambig" if the exact antecedent cannot be determined due to ambiguity, cf. Fig. 3 and 4).

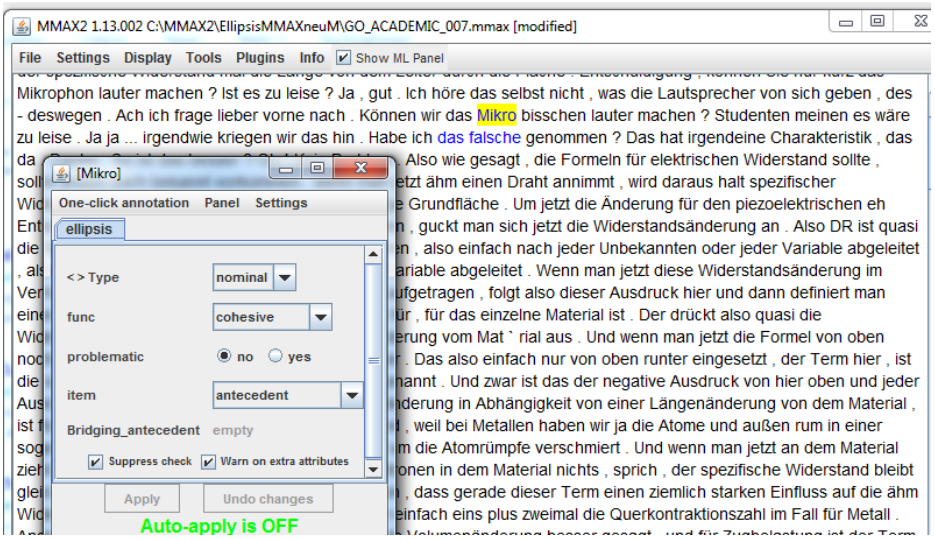

Fig. 3: Annotation of antecedent in MMAX2

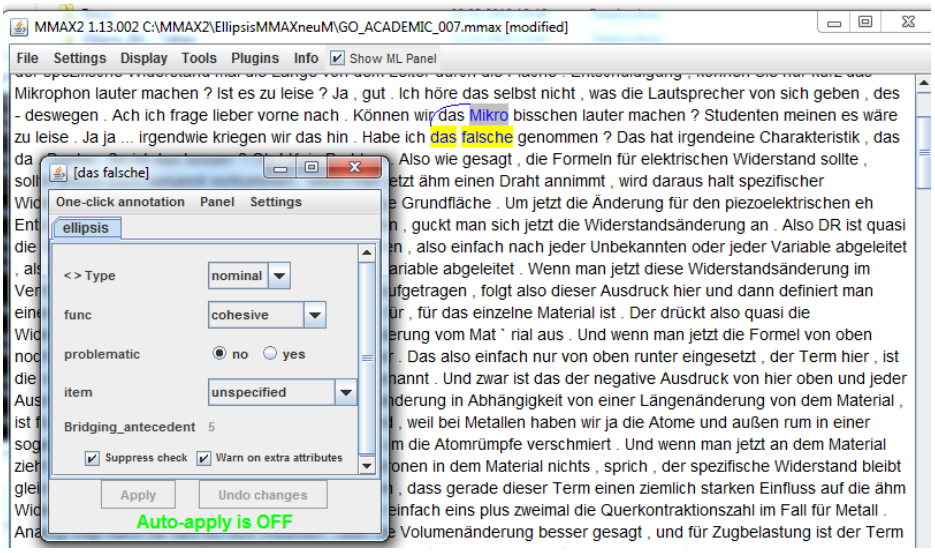

Fig. 4: Link between elliptical phrase and antecedent in MMAX2

For ellipses, the level "item" is left as "unspecified". Problematic cases that require further discussion, for example, cases on the borderline to other cohesive devices or structures where it is debatable whether they should be analysed as ellipses at all as there might be different opinions in the literature can be marked as **problematic** (problematic: ‚yes' / ‚no', cf. Fig. 5).

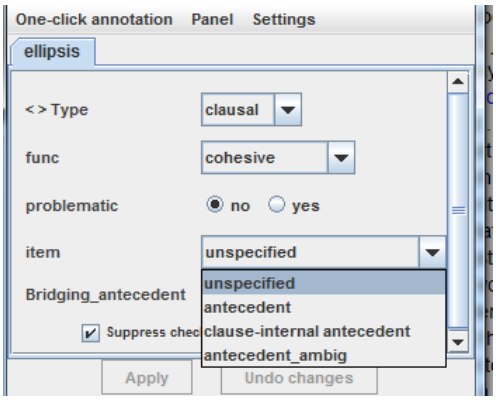

Fig. 5: Distinction between problematic and clear cases and annotation as "unspecified" if the element is an ellipsis and no antecedent

It should be kept in mind that the actual anaphoric element is the ellipsis itself. The ellipsis remnant is connected to the nearest antecedent. Typically there is an anaphoric relation. The annotation does not distinguish explicitly between anaphoric and cataphoric relations as cataphoric ellipses are quite rare in the corpus data or they fall under the category of right node raising. If the element the ellipsis refers to occurs after the ellipsis site in the text this element will also be marked as an "antecedent" (it is actually a "postcedent"). If there are chains of nominal ellipses where several ellipses refer to one common antecedent, these ellipses are all linked with the same antecedents.

An ellipsis referring to an antecedent within the same clause should be marked as clause-internal. Its immediate antecedent is marked as clause-internal as well. If it is possible to argue that the ellipsis refers equally to a cross-clausal antecedent and creates a textual link to a context larger than the immediate clause, it is possible to link such an ellipsis to an additional extra-clausal antecedent (Fig. 6). Cases without any textual antecedents (e.g. exophoric nominal ellipsis) can be marked as 'non-cohesive'.

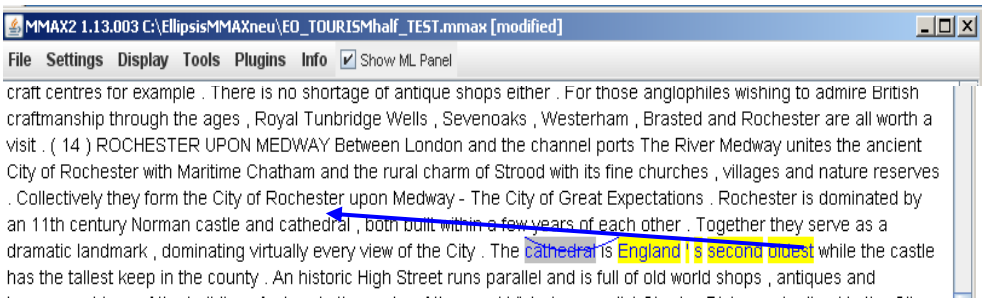

Fig. 6: Nominal ellipsis with anaphoric reference to antecedent occuring both in the same clause and one of the previous sentences (local ambiguities with regard to antecedent are possible, e.g. ellipsis could theoretically also refer to 'landmark' in the previous sentence)

2. The different types of ellipsis

2.1 Nominal ellipses

Nominal ellipsis is the omission of a head noun in a noun phrase (NPs) and its purpose is to avoid explicitly mentioning or repeating a noun. Often nominal ellipses involve deletions of nouns after, numerals, quantifiers or adjectives. Compared with other ellipsis types, nominal ellipses typically occur in various registers of spoken language as well as in different text written text types such as narrative, technical or business writing.

Examples of nominal ellipses after numerals in English:

- *This copy is defective but the other two [copies] are fine.*
- *We have two keys but we need three [keys].*

Examples of nominal ellipses after numerals in German:

- *Im Erbfall hätte unsere Tochter viel mehr bekommen als seine drei [Töchter].*
- *Vor den Spielen 2008 lagen zwischen EM und Olympia viereinhalb Monate, diesmal sind es nur zwei [Monate].*

Das Hotel hat zwar nur drei Sterne, hätte aber vier [Sterne] verdient.

Examples of nominal ellipses after adjectives in English:

- *Which hat will you wear? This is the best [hat].*
- *Although Helen is the oldest girl in the class, Julie is the tallest [girl]*

In German, the ellipsis remnant has to show strong morphological agreement in order to license the elided noun. In English, nominal ellipses after adjectives are less frequent than in German and mainly restricted to ellipses after some frequently used modifiers (e.g. adjectives describing size, age, material or colour, especially if they express contrast to other modified NPs in the immediate context) or after comparative and superlative adjectives as those forms have a richer morphology. To avoid explicitly mentioning or repeating a noun, English can use the nominal substitute "one" (which in a few cases is optional).

Examples of nominal ellipses after adjectives in German:

- Müllers haben zwei Töchter. Die jüngere [ ] geht noch in den Kindergarten. *- Die größere Katze schlief in dem braunen Körbchen. Die kleinere [ ] schlief in dem schwarzen [ ].*

Nominal ellipses can be introduced by possessive -s or genitive marker:

- It did come across as if she thought her opinion was worth more than John's [ ]. *- Everyone's favourites (well surely someone's [ ]) are the cupcakes she made.* *- Ich mag Antonias Auto, aber Martins [ ] ist besser.*

English has the possibility to omit nouns after classifier nouns. These nouns behave like adjectives and refer to materials / substances when there is no specific adjective such as *wooden / golden / linen / brazen* (archaic). German has fewer zero-derivational relationships and the distinction between noun/adjective is clearer.

Example of nominal ellipses after classifier nouns in English:

- I prefer cotton shirts to nylon [ ].

In both languages nouns can be omitted after quantifiers and indefinite pronouns.

Example of nominal ellipses after quantifier / indefinite pronouns in English:

- While Kim had lots of books, Pat had very few [ ].

Examples of nominal ellipses after quantifier / indefinite pronouns in German:

- Kim hatte viele Bücher, aber Pat hatte nur sehr wenige [ ].

Quantifiers and indefinite pronouns referring (exophorically) to people in general and not explicitly to nouns mentioned in the immediate co-text will not be treated as cohesive ellipses (e.g. *'many / some / all / viele / manche / alle / jeder / einer'*).

As mentioned above, it can be distinguished between "non-cohesive", "clause-internal" and "cohesive". A nominal ellipsis referring to a textual antecedent from a different clause is marked as "cohesive". If the antecedent of a nominal ellipsis is in the same clause (which is quite often the case), the effect of textual cohesion is less significant, and those cases are marked as "clause-internal" so that it is possible to treat them separately. The ellipsis can sometimes occur in a phrase after a linking verb, e.g. a predicate noun phrase.

Clause-internal nominal ellipsis:

- Our tax rates are comparable to Germany's [ ].

Non-cohesive nominal ellipses might refer exophorically to extra-linguistic elements that can be recovered from to the situational context. These ellipses do not have textual antecedents

and are not used endophorically.

Example of exophoric ellipsis:

- (Woman pointing to a shop window:) *Soll ich das rote [Kleid] oder ein blaues [Kleid] nehmen?*

2.2 Verbal ellipses

Verbal ellipses are omissions of operator (i.e. auxiliary or modal verb) or lexical verbs in VPs (in contrast to nominal ellipsis where only one specific element can be left out). In general, English has more possibilities than German to leave out VP-elements due to the different and often longer structure of VPs in English. The omission of VP-elements often also involves the omission of other constituents in English and German, e.g. the subject in operator ellipsis (which is also called 'ellipsis from the left' in English) or the object in some cases of lexical verb ellipsis (also called 'ellipsis from the right' in English).

Examples of operator ellipsis:

English:
What have you been doing? – [ ] Swimming.
German:
Hat er geweint? – Nein, [ ]gelacht.

After modal verbs and after the auxiliaries *'be/sein', 'have/haben', 'werden/will',* lexical verbs can be left out in both languages, often including the omission of additional VP complements (e.g. object, adverbial phrase). In the annotation, structures are annotated as verbal ellipsis if

a) a verb is missing in the VP
b) if the whole VP has been omitted (e.g. if it consisted only of a lexical verb)
c) the VP or a part of the VP + complement has been omitted
d) if the subject + a part of the VP have been omitted.
If at least the whole VP and the subject are left out, it can be annotated as clausal ellipsis.

Omission of lexical verb after auxiliary and modal verbs (lexical verb ellipsis):

English:
- *Have you eaten? - Yes, I have [ ].*
- *Are you listening? – Yes, I am [ ].*

German:
- *Hast du gegessen? Ja, [ ] habe ich.* (here: subject-auxiliary inversion in German as in many other cases of lexical verb ellipsis, instead of a verb 'dies/das' might be inserted)
- *Er kann Ihnen nicht sagen, woher er dies weiß. Er könnte es vielleicht [ ], aber er möchte es nicht [ ].*

Remnants of verbal ellipses that refer endophorically to textual antecedents in texts of the GECCo corpus are relevant for the annotation with MMAX2. They are linked to their antecedent in the annotation:

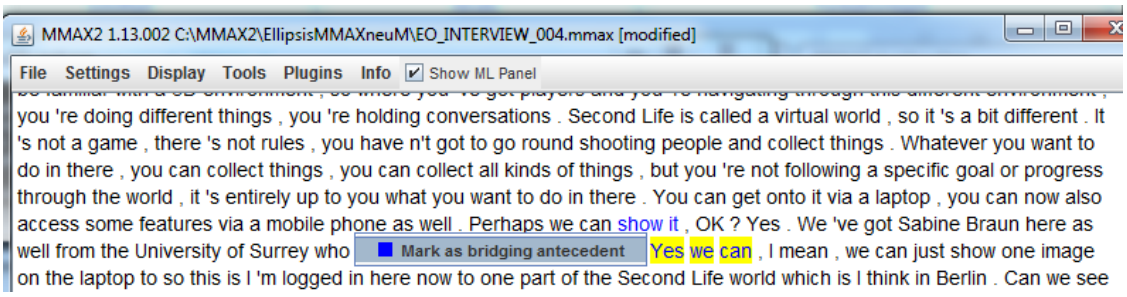

Fig. 7: Verbal ellipsis and its antecedent in MMAX2

Typically, verbal ellipses in the corpus refer anaphorically to an antecedent **from a different clause**. These cases are marked as "cohesive". Clause-internal endophoric verbal ellipses would be very rare in contrast to clause-internal endophoric nominal ellipses. There are some cases of tautosentential[1] cataphoric verbal ellipsis, particularly in German, but they would fall under right node raising.

Examples of right node raising ("Linkstilgung") that might alternatively be analysed as cataphoric lexical verb ellipsis:

- *Jim can but Jerry cannot make the meeting.*
- *Er wollte und sie konnte nichts dagegen tun.*

2.3 Clausal ellipses

The line drawn between verbal and clausal ellipsis by Halliday/Hasan is not very sharp. As the classifications used for our annotation should not overlap, and the aim is to place all cases found in the GECCo corpus clearly in only one category, clausal ellipsis is defined here as omissions that have not been covered under nominal and verbal ellipsis yet.

Certain sentence-internal omissions such as **subject ellipsis in coordination** are not marked in the annotation as there is no specific focus on these constructions in our analysis, but they occur in highs number in coordinated sentences. If the subject and a verb are left out in gapping, this type of gapping falls under clausal ellipsis.

Clausal ellipses that are used endophorically as cohesive devices are typically found in dialogic interaction, e.g. fragment answers in question-answer-pairs, reduced recapitulatory echo questions, echo exclamations, corrections, confirmations and other types of adjacency ellipses in utterance pairs. They generally work very similar in English in German. Often only one single constituent is left as an ellipsis remnant, but German ellipsis remnants usually carry more morphological information tying the constituents that are left more explicitly to the syntax of the previous sentence.

[1] belonging to the same sentence

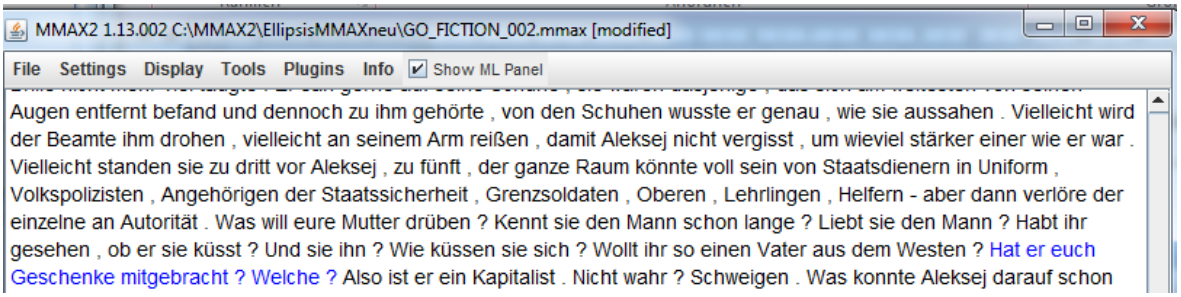

Fig. 8: Clausal ellipsis and its antecedent in MMAX2

Examples of clausal ellipses in adjacency pairs:

question-answer-pairs (wh-questions, alternative questions and yes-no questions):
- *What does it look like? – A tape recorder. / Wie sieht es aus? – Wie ein Tonbandgerät.*
- *In how many are you failing in? – Four. / Und in wie vielen fällst du durch? – In vier.*

Halliday/Hasan do not particularly mention the possibility to omit predicative expressions. Such cases can be treated as a specific subtype of clausal ellipsis. In a way, their structure resembles that of lexical verb ellipsis. However, not a verb, but an adjective or another predicative expression is omitted.

The most frequent type of non-clausal units and sentence fragments are those that do not particularly refer to a specific textual antecedent (but for example can be understood using situational, text type or world knowledge). Quite often they cannot be used as a linking element in a text at all and they are not always the result of an omission. They might be annotated as clausal ellipsis + "incohesive", but many of these will fall under specific categories discussed in the next chapters. Incomplete sentences, headlines, slogans, for example, that do not have to be completed by adding something that was previously mentioned in the same text will not be annotated as ellipses but as separate categories.

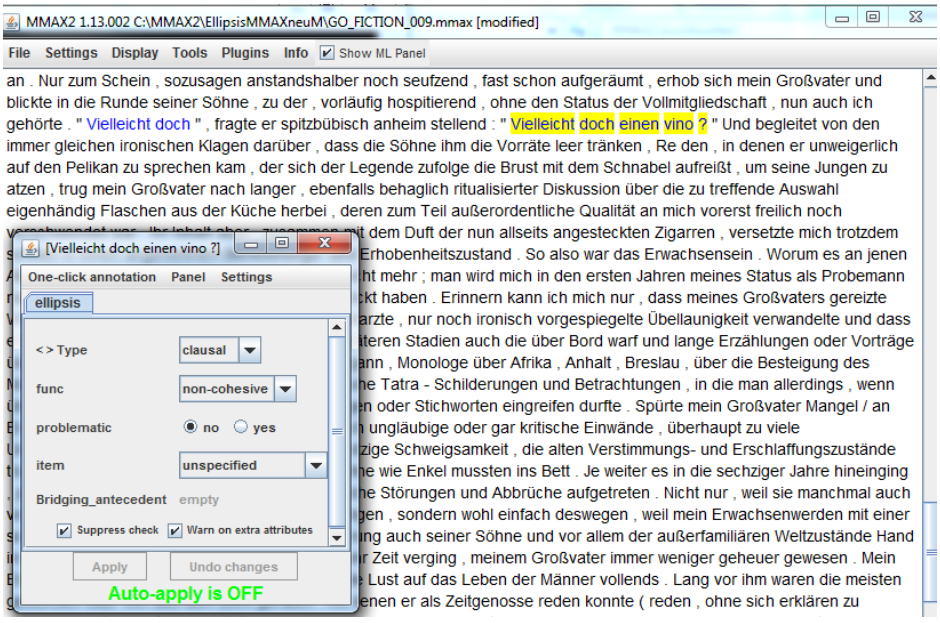

Fig. 9: Clausal (situative) ellipsis without antecedent

2.4 Mixed categories

If nominal ellipsis co-occurs with verbal or clausal ellipsis, it can be annotated as MIXED. It would also have been possible to annotate them twice (a) as nominal ellipsis, b) as verbal/clausal ellipsis). However, this might create the impression that a text has more non-clausal units than it actually has. Cases of potential overlap only between verbal and clausal ellipsis (where VPs and additional whole constituents are left out) are either annotated as verbal or clausal ellipsis as has been explained above.

- *Four Oysters followed them and yet another four [ ].* (Halliday/Hasan mention this as an example of nominal ellipsis although other constituents have been left out as well (VP + object)
- *How many slices do you want? - Two.* (nominal+clausal)
- *The first postulate denotes the goal, the second [ ] the road that leads to this goal.* (nominal+verbal)

3. Other categories similar to ellipses

3.1 Overview

Apart from ellipsis, there are several other reduction strategies and possibilities of non-standard syntax in English and German. One type of incomplete structures is aposiopesis, i.e. if a sentence is deliberately broken off (e.g.: *I don't always use incomplete sentences. But when I do...* / *Wer andern eine Grube gräbt…*). An anapodoton is a stand-alone subordinate clause. An anacoluthon ('Satzbruch' or 'syntactic blend') is an abrupt change in the syntax of a sentence (mainly in spoken language). If one verb refers to different NPs at the same time, this might co-occur with it syllepsis or zeugma (e.g.: *I am leaving for greener pastures and ten days. / Er schlug die Scheibe und den Weg nach Hause ein.*) etc.

In addition to the ellipsis categories explained in the previous chapters, the following categories have been included into the debate about ellipsis and the annotation process. They might superficially look very similar to the categories mentioned above and have sometimes been called 'ellipsis' in the literature. However, there are a various reasons for treating them as specific separate structures, e.g. if they are non-clausal units that are different from ellipses indicating omitted textual material.

- **text type specific ellipses and fragments**
- **sentence splits**
- **short yes / no replies where a particle conveys affirmation or negation**
- **non-clausal units that are not considered to be the result of syntactic omission**
- **other**

The results of the annotation of those additional categories can be used, for instance, to make comparisons between incomplete structures (elliptisch/auslassend/lückenhaft) and other cases of non-standard syntax or other means of *'Sprachökonomie'* (komprimierend/kompakt/kondensierend/verdichtend). The overall syntactic or fragmentary

nature of texts and registers and the tendency towards syntactic standardisation might be other factors that could be assessed. We will not be concerned with purely semantic implications that are not reflected in incomplete syntax.

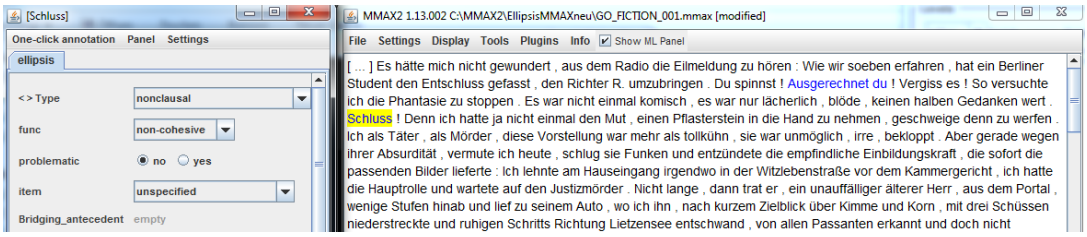

Fig. 10: Non-clausal unit (sometimes also called 'ellipsis', but different from ellipses that indicates omitted textual material)

3.2 Text type specific ellipses and fragments

Text type specific ellipses and fragments are a certain type of non-clausal units related to a certain text type or the beginning / end of a document. Headlines and captions under a figure, image or table fall under this category as well as bulleted items or numbered lists, fragments in recipes, instructions or manuals *("Mit Eigelb bestreichen" / "Beiliegenden Antrag sorgfältig durchlesen")* greetings and other formulaic non-clausal units that clearly belong to a certain text type or mark the beginning or end of a text such as a political speech or a letter *(Dear shareholders, Sehr geehrte Damen und Herren).*

Subject pronouns and/or a form of *'do'* or *'be'* in English and topics in general in German can be dropped in certain written text types (diary, text message, email, fictional dialogues, in German: jokes) and in other forms of informal communication. They might also be analysed as situational clausal ellipsis if no textual antecedent is necessary for the interpretation of the sentence.

3.3 Sentence splits

Some cases are on the borderline between sentence split (e.g. by creative use of punctuation marks as in advertisements or fictional texts) and cohesive ellipsis as in most cases theoretically more syntactic material might be added – also in way that takes up the structure of the previous environment. However, often there is no actual omission, but a specification is added that may be integrated into the syntax of the previous sentence. The punctuation mark signals a speaking pause (very frequent in political speeches in GECCo), a pause for effect, emphasis, or reflection. German, in particular, uses many of these constructions.

- *You can even stream your songs to the dorm room next door. Or down the hall.*
- *Many of the people involved would gladly have kept individual aspects of it. Lech Walesa for example, the hero of system change in Poland.*

If a constituent or part of a clause at the end of a sentence is separated from the rest of the sentence by full stop, dash or colon, this also falls under sentence split (however, these are mainly appositions after dash or colon):

- *They installed one of the most sophisticated conflict-management systems in the world: collective bargaining.*

- *Über zehn Jahre lang haben die Exporte den Amerikanern hochwertige Waren zu niedrigeren Preisen verschafft - besonders für einkommensschwache Familien ein Segen.*

A constituent or a part of a clause may also be split at the beginning of a structure. It is separated from the following sentence to create a pause (usually by inserting a colon or a dash).

- *Auch hier zeigt sich: Der Osten ist auf dem richtigen Weg.*
- *Und: Das schlägt durch beim Export.*

The annotation of spoken registers is more difficult with regard to this category. Turn-taking is sometimes not clear from MMAX2-file. Transcribers had made the decisions on where to put sentence boundaries (already an interpretation).

3.4 Short yes / no replies

Short answers with only *'yes'* or *'no'*, their German equivalent, and similar constructions (*'OK' / 'Yes and no.' / 'Indeed' / 'Sowohl als auch.' / 'Genau.'*) are 'Satzäquivalente' (also called 'Wörter mit Satzcharakter' (Duden 1052), sentence substitutes / pro-sentences or sentence words (Helbig/Buscha 1994; Schachter 1985: 32) that have sentential character. Their morphological-syntactic classification as 'particle' or 'adverb' is unclear (Bussmann 1996), as is their connection to ellipsis.

3.5 Non-clausal units

Non-clausal units include exclamation (usually one word or a noun phrase e.g. *Great! / The Beatles! Charming people!* or a subclause: *That it should have come to this!* or a conditional fragment: *If it isn't my old friend!*), congratulations, vocatives, greetings, excuses *(*etc. *Herzlichen Glückwunsch! / Sir! / Bye! / Sorry!),* fragments to express thanks, stand-alone discourse markers, formulaic non-sentences, sayings and slogans (e.g. *The sooner, the better. More haste, less speed.).* They are annotated as non-clausal units (provided they did not fall under the category of text type specific units already). As the spoken corpus texts consist of a huge number of non-clausal units, only a selection of sample cases has been annotated in those registers.

3.6 Other

Cases of aposiopesis, anapodoton and anacoluthon and omissions that do not fall under any of the categories described above can be annotated as "other". In general, there are only very few examples from the corpus that would fall under this category. It might include non-rule based, spontaneous omission in spoken language, for instance but ungrammatical ellipses of prepositions or articles:

- *What else have we used it [ ]?*
- *[ ] Student hat den Unfall überlebt.*
- *[ ] Logik ist klar?*
- *um [ ] Bus zu nehmen*

Question tags as a specific case have been annotated under the category "other". Question tags (e.g. *isn't it? / haven't you?* etc.) might theoretically be seen as verbal or clausal ellipses,

depending on whether a part of the VP or other constituents are missing. Sometimes expressions such as *no?/correct?/right?* were used in the corpus texts as well. German has some invariant all-purpose tags such as *nicht wahr? / oder? / ne?*... Often a discourse particle is used where English might have used a question tag.

Fragments in joint utterance construction can also be put under "other" if a syntactic unit has been completed jointly by two or more participants in interaction. (e.g. *A: Heute ist der? B: Erste. / A: All we had was water. B: Bottled water.*)