Motivating W(H)-Clefts in English and German: A hypothesis-driven parallel corpus study

Volker Gast (Jena), Natalia Levshina (Marburg)*

1 Introduction

This contribution is concerned with the syntactic configuration illustrated in (1), which is often subsumed under the terms "pseudo-Cleft"[1] or "WH-Cleft" in the English literature, in contradistinction to (genuine) Clefts or IT-Clefts[2] as in (2) (note that the term "Cleft sentence" was introduced by Jespersen 1937: 83–89, according to Fischer 2009: 169).

(1) Pseudo-Cleft / WH-Cleft[3]
What I miss in Mr Martens' report are the guidelines and vision[s] of how a new cooperation could be formed. (EPEG-6/Da 1451517)

(2) (IT-)Cleft
Obviously, it is the result that interests us. (EPEG-6/Fr 792462)

*The paper has benefited greatly from comments made by various colleagues, especially the participants of the workshop on Cleft constructions organized by Anna-Maria De Cesare in Basel (June 4, 2012). Moreover, we would like to thank an external reviewer for a very thoughtful review and valuable suggestions for improvements. Any inaccuracies are of course our own responsibility. The paper continues earlier work with Daniel Wiechmann (cf. Gast & Wiechmann 2012), Olga Rudolf and Marie Schneider. We have reused some of the data that we coded jointly with these colleagues at an earlier stage. V. Gast wishes to acknowledge financial support from the German Science Foundation (DFG GA 1288/4-1).

[1] Technical terms such as "Cleft", "Topic", etc. are capitalized.

[2] See for instance Akmajian (1970); Prince (1978); Declerck (1984); Collins (1991); Lambrecht (2001); Collins (2006); Hedberg & Fadden (2007); Dufter (2009); Fischer (2009), among other publications.

[3] The examples used for this study have been taken from the EUROPARL-corpus, cf. Koehn (2005) and Cartoni & Meyer (2012). The label "EPEG-6" stands for "EUROPARL English-German, version 6". The corpus contains the proceedings of the European Parliament. Version 6, which we used for our study, contains the proceedings from April 1996 to December 2010, approx. 55 M words in the English corpus part and approx. 49 M words in the German part. The corpus contains both original language and translations. Until 2003, the texts were translated directly from the source languages into any of the target languages. From 2003 onwards, English has been used as a "pivot language" (Cartoni & Meyer 2012: 3), i.e., all languages were first translated into English and then into the relevant target language. In the indication of the source, the abbreviation after the slash indicates the original language of the example in question, and the following number the line in the corpus.

In German, the term "Sperrsatz" is often used to distinguish examples like (3) from those of the type illustrated in (4), which are commonly subsumed under the term "Spannsatz" (see for instance Altmann 1981, 2009). Alternatively, the terms "W-Cleft" and "ES-Cleft" are used (e.g. by Fischer 2009; Gast & Wiechmann 2012).

(3) Sperrsatz / W-Cleft
Was uns interessiert, ist selbstverständlich das Resultat. (≡ [2])

(4) Spannsatz / ES-Cleft
Es ist zweifellos der Terrorismus, der verhindert, dass die Parteien an einen Verhandlungstisch kommen. (EPEG-6/It 511911)
'It is certainly terrorism that prevents the parties from coming to the negotiating table.'

We use the term "W(H)-Cleft" as a generalization over the corresponding structures of English and German, i.e., English WH-Clefts and German W-Clefts.

In addition to the two types of Cleft sentences introduced above, a third type is standardly distinguished, i.e., reversed pseudo- or W(H)-Clefts (cf. the references in Note 2). Examples of this construction are given in (5) and (6) for English and German, respectively.

(5) Reversed pseudo-Cleft / WH-Cleft
Champaign is what I like. (Lambrecht 2001: 467)

(6) Reversed W-Cleft
Ist es das, was dieser Junge Mann in Indien suchte? (Altmann 2009: 17)
'Is that what this young man was looking for in India?'

We believe that reversed W(H)-Clefts – in English as well as in German – constitute an interesting topic of its own, and that they are not merely a structural variant of uninverted W(H)-Clefts. As has been shown by Hedberg & Fadden (2007), among others, their information structural properties are probably more similar to those of IT-Clefts than to those of uninverted W(H)-Clefts. We will therefore focus on uninverted W(H)-Clefts in this contribution, but a comparison with reversed W(H)-Clefts is certainly an interesting topic for future research.

The present study addresses the following question:

(7) Research question
What determines the distribution of W(H)-Clefts in English and German?

The question in (7) is approached on the basis of data from a parallel corpus, EUROPARL (more precisely, EUROPARL is a translation corpus; cf. Koehn 2005; Cartoni & Meyer 2012; see also Note 3). The design of the study is as follows: We use properties of English WH-Clefts as independent variables that are taken to correlate with one formal property of the corresponding German sentences, captured as the dependent variable of our study, the presence or

absence of a W-Cleft. The properties of the English WH-Clefts constituting the independent variables are intended to reflect motivations for (not) using a W(H)-Cleft.

The rationale underlying this design is the following: We assume that all pairs of sentences from the parallel corpus are (semantically / pragmatically) near equivalent, i.e., they convey basically the same "message", speaking in very general (information-theoretical) terms (cf. Gast forthcoming on the notion of "near equivalence" in contrastive linguistics). This assumption of interlingual near equivalence in a translation corpus is obviously an idealization, but in general, the translations of the EUROPARL-corpus are of a very high quality and certainly come close to that ideal. Given the (*a priori*) assumption of interlingual equivalence in our translation corpus, and given the (theoretically motivated) assumption that the properties of English sentences constituting the independent variables of our study reflect (communicative) motivations for (not) using a W(H)-Cleft, we can expect to find systematic covariation between the independent variables describing the English data on the one hand, and the occurrence or non-occurrence of a W-Cleft in the German data on the other, as the motivations that are reflected in the English data should have observable reflexes in the German data as well.

Given that the messages underlying pairs of sentences are taken to be invariant, the question arises why there should be differences in the distribution of English WH-Clefts and German W-Clefts at all. We believe that the observable differences in the distribution of W(H)-Clefts in English and German result from the fact that English WH-Clefts and German W-Clefts are part of different systems. As a consequence, they compete with different sets of alternative structures. The (non-)availability of alternative structures, in turn, is a crucial determinant of the distribution of W(H)-Clefts and is thus expected to lead to distributional differences between English and German.

What we investigate, thus, is the interplay between motivations for using a W(H)-Cleft and the actual use of W(H)-Clefts in English and German, viewed as the result of a multifactorial decision based on communicative needs and intentions – the invariant or *tertium comparationis* – and available linguistic resources, the most important factor giving rise to distributional variation between English and German.

The relationship between specific motivations underlying the use of a W(H)-Cleft and the type of structure actually found in a given case will be modelled (metaphorically and informally) in terms of a cost-benefit analysis. As will be detailed in Section 4, we assume that W(H)-Clefts provide specific *benefits* which are (partly) independent of their basic information structural function. The benefits of W(H)-Clefts provide motivations for using these structures, in spite of their higher *costs* (in comparison to canonical clauses). Accordingly, a W(H)-Cleft is expected to be used in those cases where it provides a benefit of the type discussed in the present paper which would not be provided by a canonical ("cheaper") structure. Obviously, the various motivating factors are not mutually exclusive but may contribute jointly to the use of a W(H)-Cleft. The more motivating factors there are, the more likely the use of a W(H)-Cleft

will be.

The asymmetrical design of our study – we investigate covariation between English WH-Clefts and their German counterparts in a translation corpus – is due to the fact that W(H)-Clefts are much more widely distributed in English than in German. As Gast & Wiechmann (2012) have shown, the ratio of W(H)-Clefts in these languages is approximately 4:1 in the EUROPARL-corpus. Even though there are cases where we find a W(H)-Cleft in German but not in English – perhaps even systematic and general ones (cf. Section 4.3 for a possible candidate) – it thus makes sense to take the English data as a point of departure. An investigation determining the English structures corresponding to German W-Clefts would probably provide interesting additional information about the distribution of W(H)-Clefts in the languages under comparison, but would not invalidate the findings reported on in the present study.

As the (aymmetrical) design of our empirical study only distinguishes two cases – (i) those where we find a W(H)-Cleft in both languages under comparison, and (ii) those where we only find a WH-Cleft in English – we will be dealing with two major types of benefits as well, i.e., (i) those benefits that are expected to motivate the use of a W(H)-Cleft in English to a (significantly) greater extent than in German, and (ii) those benefits that are not expected to lead to a distributional overrepresentation of W(H)-Clefts in English as compared to German. The first type of benefit will be called "English-specific" and the second type "general", which we use as an abbreviation for "non-English-specific".

Before turning to our quantitative analysis, some remarks on the structure and distribution of W(H)-Clefts are made in Sections 2 and 3. In Section 4, the assumed motivations for the use of W(H)-Clefts in English and German are discussed, and five hypotheses are formulated. These hypotheses are operationalized and tested in Section 5. In Section 6, we identify correlations between independent variables on the basis of a Multiple Correspondence Analysis. As there are clear correlations between two of the variables investigated, we identify three (instead of five) major types of English WH-Clefts, which are associated with specific motivations of use. Only one of these types is also (more or less) widely attested in German, which explains the differential distribution of W(H)-Clefts in English and German. Section 7 contains a summary and the conclusions.

2 Remarks on the structure of W(H)-Clefts

W(H)-Clefts are constituted by a matrix clause headed by a copula in specificational function and a W(H)-clause filling the subject position of that copula (see Declerck 1984; den Dikken 2009; Gast & Wiechmann 2012 for the specificational – as opposed to predicational or equative – interpretation of the copula in such sentences). There is no general consensus concerning the exact function of the W(H)-clause, but mostly it is regarded as a free relative clause (see e.g. Altmann 2009). Alternatively, it could be interpreted as an indirect question (cf. for instance Faraci 1971). We will not take a stance in this matter and simply use

the terms "W(H)-clause" and "Cleft constituent" for the two main constituents of a W(H)-Cleft. This terminology is illustrated in (8), using example (1) above.

(8) [What I miss in Mr Martens' report] are [the guidelines ...]
[W(H)-clause] COP [Cleft constituent]

W(H)-Clefts are restricted in terms of the category of the Cleft constituent. As will be shown in this section, there are only few differences between the structural properties of English and German W(H)-Clefts. English WH-Clefts are found with DPs, sentential Cleft constituents, VPs and *to*-infinitives (cf. [9]–[12]).

(9) DP
What is actually needed however are [$_{DP}$ funds and resources]. (EPEG-6/Sp 425590)

(10) Finite clause (CP)
What this means is [$_{CP}$ that we are in a position to continue the European Union's development in line with what is needed]. (EPEG-6/Ge 281996)

(11) VP
What we have to do is [$_{VP}$ apply the Community acquis and not call other policies into question]. (EPEG-6/Sp 300029)

(12) *To*-infinitive (TP)
What we do not want is [$_{TP}$ to exempt from transparency obligations the small category of highly specialised financial journalists ...] (EPEG-6/Du 334172)

English WH-Clefts are ungrammatical with adverbials and PPs as Cleft constituents, however (cf. [13] and [14]).

(13) Adverb
*When I became a young revolutionary was THEN. (Prince 1978: 885)

(14) PP
*What/where/how many protest is AGAINST PARDONING THESE. (Prince 1978: 885)

German W-Clefts cover basically the same range of possibilities as the English WH-Clefts, as is illustrated in (15)–(18).

(15) DP
Was die Europäische Union jetzt braucht, ist [$_{DP}$ ein schnelleres Wirtschaftswachstum] ... (EPEG-6/Pl 594004)
'This means that faster economic growth [...] is needed in the European Union.'

(16) Finite clause (CP)
Was ich jetzt klar erkenne, ist, [$_{\text{CP}}$ dass es keine wirkliche Chancengleichheit in der landwirtschaftlichen Produktion gibt ...] (EPEG-6/En 595461)
'The one thing I now see clearly is that there are no real level playing fields in agricultural production [...].'

(17) VP
Was wir nicht können, ist, [$_{\text{VP}}$ ein Abkommen schließen, das die WTO dann als nicht kompatibel ablehnt]. (EPEG-6/Po 1543370)
'What we cannot do is come up with an agreement and afterwards WTO tells us it is not compatible.'

(18) *Zu*-infinitive (TP)
Was wir wollen, ist ja, [$_{\text{TP}}$ den Unfallopferschutz zu verbessern]. (EPEG-6/Ge 1731070)
'Surely what we want is to improve the protection of accident victims.'

As in English, adverbs cannot be W-clefted in German:

(19) Adverb
*Wie ich nach Hause kam war schnell.
Lit.: 'How I came home was fast.'

There is one syntactic difference between English and German, however. As is well known, English allows extraction out of PPs, so complements of prepositions can be WH-clefted (cf. [20]).

(20) Extraction out of PP in English
Today, what$_i$ we are talking [$_{\text{PP}}$ about e$_i$], to put it simply, is the ownership of ideas. (EPEG-6/En 634287)

Such extraction is not possible in German. German, in turn, has a lexical option for these cases which English lacks, and which renders W-clefting of PPs possible. The W-forms of so-called "conjunctional adverbs", which consist of the prefix *wo(r)-* and a preposition, stand for a syntactic unit of the type [$_{PP}$ P PRO$_{wh}$], and they can be W-clefted, as is illustrated in (21).

(21) W-form of conjunctional adverb in German
Worüber$_i$ wir heute sprechen [$_{\text{PP}}$ e$_i$], ist – um es einfach auszudrücken – das Eigentum an Ideen. (≡ [20])

In terms of frequency, the difference between English and German that is illustrated in (14), (20) and (21) is not particularly relevant, as the corresponding structures are relatively rare, at least in the corpus used for the present study, the EUROPARL-corpus. We will thus assume that by and large, W(H)-Clefts can be used in the same range of contexts in English and German, as far as their syntactic restrictions are concerned, and that actual distributional differences

are due to differences in the motivating factors, or the interplay between those factors and the "linguistic ecologies" of English and German. A quantitative comparison of the types of Cleft constituents found in the English and German part of the EUROPARL-corpus has been provided by Gast & Wiechmann (2012).

3 The standard motivation of W(H)-Clefts

3.1 On Givenness and Topic-Comment structure

In the comprehensive literature on that construction, W(H)-Clefts are mostly assumed to have an information-structural function, i.e., a function relating to "common ground management" (cf. Krifka 2007 for that term). One of the most influential information structural descriptions of W(H)-Clefts has been provided by Prince (1978), who describes their function as follows:

> A WH-cleft will not occur coherently in a discourse if the material inside the (subject) WH-clause does not represent material which the coöperative speaker can assume to be appropriately in the hearer's consciousness at the time of hearing the utterance.
>
> (Prince 1978: 888)

There is a broad consensus that WH-Clefts are associated with stricter conditions of use than IT-Clefts, as was also claimed by Prince (1978). However, the claim that the W(H)-clause needs to be Given (in a technical sense, hence capitalized) has been qualified and challenged. In their empirical investigation, Hedberg & Fadden (2007) have pointed out that the W(H)-clause of a W(H)-Cleft need not be "referentially Given", in terms of Gundel & Fretheim (2004). What is required is that the it be "relationally Given". In a nutshell, this means that the W(H)-clause is a Topic and the Cleft constituent a Comment (about the Topic). We will adopt this terminology below, even though our analysis differs slightly from the one provided by Hedberg & Fadden (2007).

The analysis of Prince (1978) – in particular, her observation that W(H)-Clefts are subject to more specific conditions of use than IT-Clefts – has been challenged by Declerck (1984), who claims that W(H)-Clefts cover the same range of information structural possibilities as IT-Clefts. He observes that the W(H)-clause of a Cleft sentence need not be Given, as in (22), and that there are cases where W(H)-Clefts can even be used as discourse openers (cf. [23]).

(22) A: Those apples are good, aren't they?
B: So they are! What keeps me from eating all of them is that mother would be furious if I left none for the others. (Declerck 1984: 259)

(23) My dear friends, what we have always wanted to know, but what the government has never wanted to tell us, is what exactly happens at secret conferences like the one you have been reading about in the papers this week. (Declerck 1984: 257)

Similar cases were discussed by Prince (1978) already (e.g. her example 14 on p. 888), and it seems clear that there is accommmodation at work. We will thus assume that the Prince-Hedberg-Fadden analysis is basically correct.

Gast & Wiechmann (2012) have slightly rephrased the analyses of Prince (1978) and Hedberg & Fadden (2007) by using the concept of "Quaestio" from discourse analysis (cf. Klein & von Stutterheim 1987; Gast & van der Auwera 2011; cf. also Büring's 2003 concept of "[current] question under discussion / QUD"). Roughly speaking, the Quaestio-theory assumes that every utterance in a discourse corresponds to some (mostly implicit) question that is under discussion. This question is called the "Quaestio" of the relevant utterance (cf. Klein & von Stutterheim 1987). The answer given to the Quaestio is called the "Responsio" (cf. Gast & van der Auwera 2011). In canonical cases, the Quaestio of an utterance can be read off the information structure of the corresponding sentence by replacing the focus with an appropriate question word. More often than not, however, the Quaestio is not immediately obvious and needs to be inferred from contextual information. The sequence of sentences in (24) (from the EUROPARL-corpus) will help to clarify the concept of "Quaestio".

(24) a. You have presented a sound economic programme, Commissioner Almunia. I am able to assent to everything contained in it. All I should like to say is that I should very much like to see an additional dimension.

b. What I should dearly like to hear come out of the economic guidelines is a constructive message to all the Member States, saying "let us now, together and in each individual country, invest in the Lisbon Objectives over the next three or four years, and let us do so at one and the same time", for doing so simultaneously would be Europe's secret weapon. (EPEG-6/Da 17923)

In (24a), the speaker mentions "an additional dimension" that (s)he would like to see. This raises a question, namely: "What additional dimension would you like to see?" This – here, implicit – question can be assumed to be the Quaestio of the following sentence. The sequence of (24a) and (24b) could be thought of as a dialogue of the following form:

(25) A: I should very much like to see an additional dimension.
B: What additional dimension would you like to see?
A: What I should dearly like to hear come out of the economic guidelines is a constructive message to all the Member States [...]

The question asked by B in (25) is precisely the WH-clause of the Cleft sentence in (24b). Accordingly, the effect of using a W(H)-Cleft in (24b) can be described as follows: The Quaestio of an utterance and the corresponding Responsio are merged into a single sentence as it were. The W(H)-clause of the Cleft sentence corresponds to the Quaestio, the Cleft constituent to the Responsio. This analysis is compatible with the one proposed by Prince (1978), as the Quaestio is obviously under discussion (or "in the hearer's consciousness", as Prince 1978:

888 puts it), and it is similar to the one of Hedberg & Fadden (2007), insofar as it is based on information structural relations rather than properties of constituents or referents (i.e., referential Givenness). Still, the Quaestio-Responsio analysis of Gast & Wiechmann (2012) and the Topic-Comment analysis of Hedberg & Fadden (2007) are not equivalent, as a Quaestio need not be an "address", in terms of Jacobs (2001). For ease of reference, we will nonetheless adopt the (more familiar) terminology of Hedberg & Fadden (2007) and thus phrase our analysis in terms of the dichotomy "Topic vs. Comment".

3.2 Linear synchronization of Topic and Comment

According to the observations made in Section 3.1, W(H)-Clefts have the effect of rendering the Topic (as well as the Comment) of a sentence as a continuous sequence of words. Consider, once again, (1) for illustration. A corresponding canonical sentence is given in (26). The W(H)-clause / Topic is rendered in italics, the Cleft constituent / Comment is enclosed by brackets ($[\ldots]_C$).

(1) Quaestio: What do you miss in Mr Martens' report?
What *I miss in Mr Martens' report* are [the guidelines and visions $\ldots]_C$

(26) I *miss* [the guidelines and visions $\ldots]_C$ *in Mr Martens' report.*

As (26) illustrates, the W(H)-clause or Topic would be discontinuous in a canonical sentence. In a W(H)-Cleft, it forms an uninterrupted sequence of words, as does the Responsio / Comment. Topic-Comment structure and linear sentence structure are synchronized as it were. This effect of W(H)-Clefting will be called the "linear synchronization of Topic and Comment" (cf. also Gast & Wiechmann 2012), and it is regarded as the "standard motivation" for the use of a W(H)-Cleft.

(27) The standard motivation of W(H)-Clefts
A W(H)-Cleft is used to render the Topic and the Comment (each) as uninterrupted sequences of words (linear synchronization of Topic and Comment).

The linear synchronization of Topic and Comment is called the "standard motivation" of W(H)-Clefts insofar as it can be recovered in most cases, even though many examples obviously imply some degree of accommodation. It is thus (basically) a necessary, though not sufficient, condition for the use of a W(H)-Cleft.[4] In German, the linear synchronization of Topic and Comment can sometimes be achieved without forming a W(H)-Cleft where this would not be possible in English. In fact, the German example corresponding to (1) in the EUROPARL-corpus – given in (28) below – is not a W(H)-Cleft. As in the English examples given above, the W(H)-clause / Topic is rendered in italics.

[4] We use the adverbial hedge "basically" because there might be cases where the primary function of W(H)-Clefts is marginal at best and the use of a W(H)-Cleft is mainly motivated by (originally) secondary benefits like those discussed in Section 4 – perhaps cases of "exaptation", speaking in evolutionary terms; cf. also Gast & Wiechmann (2012: 338).

(28) *Im Bericht Martens vermisse ich* deshalb [die übergeifenden Entwicklungslinien und Visionen für die Gestaltung einer neuen Zusammenarbeit]$_C$.
(EPEG-6 1451517, ≡ [1])

There are thus examples where we find a W(H)-Cleft in English but not in German, as the verb-second order of the latter language allows for a certain freedom of word order already. Still, in many cases a German W-Cleft seems to be motivated by the synchronization of Topic-Comment structure as well. Consider (29):

(29) Quaestio: Womit konnten wir die Sache für uns gewinnen?
'What helped us win (that matter)?'
Womit wir die Sache für uns gewinnen konnten, war
[eine schriftliche Erklärung der Kommission ...]$_C$.
'*What helped us win* was
[the Commission's pledge, in the form of a written statement ...]$_C$'

As is illustrated by (30), a corresponding canonical verb-second sentence would exhibit a discontinuous Topic:

(30) *Wir konnten die Sache* [mit einer schriftlichen Erklärung der Kommission]$_F$ *für uns gewinnen.*
'*We were able to win* [with with a written statement of the commission]$_C$.'

While many examples of W(H)-Clefts can be explained on the basis of the standard motivation as described in (27) above, we will argue that they are often motivated by other, additional factors, including ones that do not relate to information structure in a narrow sense. Such motivating factors are the *benefits* of our study, a concept to be explained in the next section.

4 Motivating W(H)-Clefts: Benefits, costs and obstacles

We assume a very simple model of language production which is obviously inspired by ideas from Optimality Theory (cf. Prince & Smolensky 1993) without making use of the formalisms provided by that framework. Speakers want to express specific meanings or functions, and they have a repertoire of structures which they can use for that purpose. Each structural option comes with specific benefits and with specific costs. The choice of a given construction is thus a matter of finding a balance between costs and benefits. Moreover, there may be obstacles to the use of a specific structure, e.g. syntactic constraints.

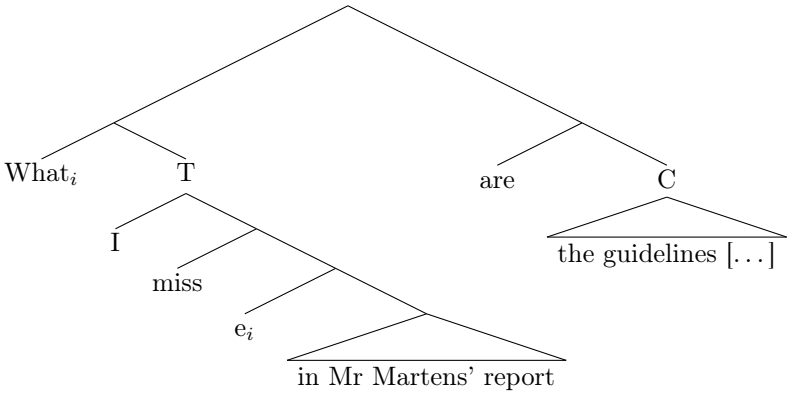
The benefits of using a W(H)-Cleft, as well as one obstacle, are described in this section. We will not provide any quantitative measure for the costs of W(H)-Clefts in comparison to canonical sentences (SVO in English and verb-second in German), but we will assume a general economy constraint on W(H)-Clefts, much in the sense of Optimality Theory. W(H)-Clefts, thus, need to be

motivated in *some* way, otherwise they will be avoided. As mentioned in Section 3.2, the standard condition of W(H)-clefting – the linear synchronization of Topic and Comment – is not regarded as a benefit, but rather as the primary function of that construction, and as a necessary, but not sufficient condition for its use. The actual use of W(H)-Clefts is regarded as a probabilistic choice: The more benefits a W(H)-Cleft provides in a given context, the more likely it will occur.

4.1 Creating an IP boundary for separate nuclear accents: Phonological separation of Topic and Comment

The linear synchronization of Topic and Comment as described in (27) is not the only structural effect of forming a W(H)-Cleft. In addition to forming a continuous chain of words, the Topic and the Comment are contained in different constituents, and are thus also separated at the paradigmatic level. This is illustrated in (31) for example (1) ("T" and "C" stand for "Topic" and "Comment", respectively).

(31)



One of the major benefits of putting the Topic and the Comment in different constituents concerns the syntax-phonology interface, in particular the organization of "tonality" and "tonicity", following Halliday's (1967) terminology (cf. Wells 2006 for a more recent overview of basic notions of intonation). The constituent boundary between the Topic and the Comment allows the speaker to align the two constituents with separate intonation phrases and, accordingly, to have separate nuclear accents in each constituent. This benefit of a W(H)-Cleft will be called the "phonological separation of Topic and Comment". Given that this rather abstract benefit manifests itself in the presence of (at least) two nuclear accents, and given that nulear accents are associated with some type of contrast, the phonological separation of Topic and Comment will be operationalized as the presence of contrast in the Topic (the Comment is assumed to necessarily contain at least one nuclear accent; for a discussion of contrastive Topics in English and German, see Gast 2010).

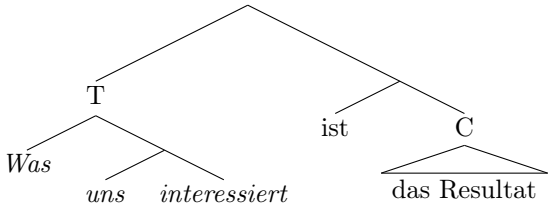
For illustration, consider, once again, (a simplified version of) (3) and its canonical counterpart in (32):

(3) Was uns interessiert, ist das Resultat.
'What interests us is the result.'

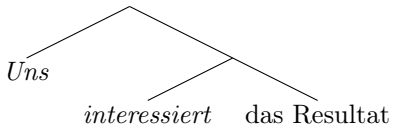
(32) Uns interessiert das Resultat.
Lit.: 'Us interests the result.'

In both (3) and (32) the Topic and the Comment are (each) adjacent; but only in (32) are the two information structural components contained in separate constituents (cf. [33] vs. [34]; the Topics are rendered in italics).

(33)



(34)



The structure in (3) / (33) allows the speaker to align the Topic and the Comment with separate intonation phrases more easily than the one in (32) / (34). As a consequence, the WH-Cleft in (3) / (33) allows for the placement of a nuclear accent in the WH-clause / Topic, and thus for the indication of contrast, more easily than the canonical clause in (32) / (34). The intonation in (35) is therefore more natural than the one in (36) (brackets indicate constituents, "|" an IP-boundary, and small caps the nuclear accent).

(35) [Was uns interesSIERT] | [ist das ResulTAT].

(36) ?Uns [interesSIERT | das ResulTAT]

In English, the constituent boundary between the WH-clause and the Cleft constituent is often used for the placement of adverbs that are little integrated into the sentence structure, e.g. "conjuncts" in terms of Quirk et al. (1985). For instance, the adverb *however* is often inserted into the "gap" between the WH-clause and the Cleft constituent, as in (37).

(37) What we are talking about here, however, are medium-sized vessels [...] (EPEG-6/Ge 585863)

As there is no obvious reason why the phonologically motivated use of W(H)-Clefts as described in this section should be more common in English than in German, we will assume that the benefit of phonologically separating the Topic and the Comment is a general one, i.e., a benefit which will not contribute to the use of W(H)-Clefts in English to a greater extent than in German.

4.2 Structural separation of propositional content and of utterance comment

The factor discussed in this section emerged upon inspection of the data used for this study. In a relatively high number of examples, the W(H)-clause does not contribute anything to the propositional content of the sentence, but merely provides a frame of evaluation, insofar as it contains information about the contextual embedding, the modal or epistemic source of the propositional content, etc. For instance, the W(H)-clause often contains a doxastic predicate like *think* (*what I think is ...*), a verb of saying (*what I want to say is ...*), or similar expressions not affecting the truth conditions of a sentence. In (38) and (39), the W(H)-clause links the propositional content conveyed in the Cleft constituent to the preceding utterance, characterizing it as a conclusion or consequence.

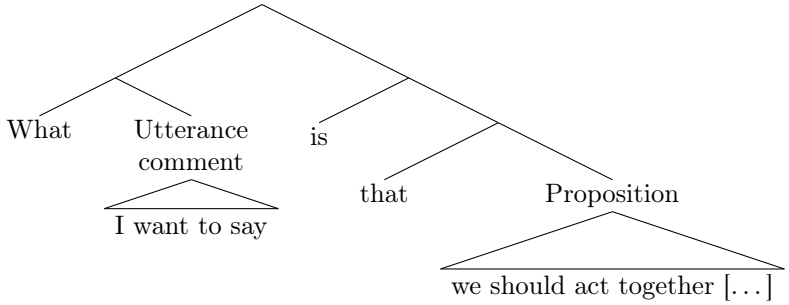
(38) What this means is that we are in a position to continue the European Union's development in line with what is needed. (EPEG-6/Ge 281996)

(39) Was ich sagen will, ist, dass wir zusammenstehen müssen [...] (EPEG-6/Sp 273398)
'What I am saying is that we should act together [...]'

In such cases W(H)-Clefts do not seem to primarily reflect Topic-Comment structure, but the difference between the level of propositional content (in the Cleft constituent) on the one hand, and the frame of evaluation (in the W[H]-clause) on the other. The matrix predication can be regarded as a comment on the proposition expressed in the Cleft constituent – an "utterance comment", as we call this type of predication (cf. also Gast & Wiechmann 2012). Utterance comments are crucially characterized by the fact that they do not contribute anything to the propositional content expressed in the clause (cf. also Section 5.2).

W(H)-Clefts make the distinction between propositional content on the one hand, and utterance comment on the other, transparent, insofar as the two pieces of information are contained in separate constituents. This type of structural separation is illustrated in (40).

(40)



The separation of utterance comment and propositional content seems to motivate W(H)-Clefts to a greater extent in English than in German. In the latter language, we often find embedded verb-second structures instead.[5] The English example (41), for instance, corresponds to an embedded verb-second sentence in the German corpus part (cf. [42]).

(41) What all this means is that we are in fact incapable of doing much at all. (EPEG-6/Pl 637137)

(42) Das heißt, eigentlich können wir nicht viel tun.

The most salient property of embedded verb-second structures in German is that the embedded clause has (representative) illocutionary force. In other words, it is asserted. A comparable stucture is not available in English. While English also allows complement clauses without an introducing complementizer, the relevant clauses are not distinguished structurally from "ordinary" (unasserted) complement clauses. Moreover, such *that*-omission is uncommon in high-register language as represented in the EUROPARL-corpus (cf. Biber et al. 1999: 680-683).

Given that German can use a more economic structure (embedded verb-second clauses) in order to paradigmatically separate the utterance comment and the proposition, while English does not have an obvious structural alternative of this type, it is to be expected that this factor will motivate W(H)-Clefts to a greater extent in English than in German. In other words, the structural separation of propositional content and utterance comment is regarded as an English-specific benefit or motivation.

4.3 Linear synchronization of constituency and operator scope

In both languages under comparison, W(H)-Clefts often lead to a reordering of scope-bearing operators, either relative to each other or relative to their scope

[5] German standardly uses verb-second structure in main clauses and SOV-structure in subordinate clauses. Verb-second structure is also found in embedded clauses, however, as in (42). In this case, a complementizer cannot be used.

domains. Consider the W-Cleft in (43), which is attested in the EUROPARL-corpus, and the corresponding canonical main clause (with extraposition) in (44).

(43) Was nicht passieren darf ist, dass es eine Vorgabe der Standards durch die Kommission durch die Hintertür gibt. (EPEG-6/Ge 358731)
'What must not be allowed to happen is standards being imposed by the Commission through the back door.'

(44) Es darf nicht passieren, dass es eine Vorgabe der Standards durch die Kommission durch die Hintertür gibt.
'It must not be allowed to happen that standards are imposed by the Commission through the back door.'

There are two scope-bearing operators, the deontic possibility (permission) modal *darf* and the negator *not*. The negator takes scope over the modal. While in a W-Cleft, the order of elements reflects the scope-relations directly, in the corresponding main clause the element with narrow scope precedes the one with wide scope. (45) illustrates the (transparent) scope relations in the W-Cleft.

(45) Was nicht passieren darf ist, [dass ...]
¬ [◊ [*p*]]

The effect that can be observed in (43) and (45) will be called "linear synchronization of constituency and operator scope".

The fact that (in German) scope relations are mirrored more directly in W-Clefts than in verb-second clauses is no coincidence but obviously related to the fact that the W-clause is a subordinate clause – with SOV order – and that in German, scope relations are more transparent in subordinate clauses than in main clauses (cf. König & Gast 2012: Ch. 11 for a comparison of English and German sentence structure with reference to scope relations). The effect in examples like (43) is that of a complete "outsourcing" of propositional operators. The W-clause contains no lexical material at all – the verb *passieren* is basically a place holder for the predicate of the Cleft constituent. This separation of propositional operators and lexical material probably makes the relevant sentences more parser-friendly. In the context of political speech – a register which is characterized by a high degree of editorial elaboration and geared towards hearer-friendliness – transparency at the level of sentence operators is a welcome effect. As a matter of fact, the quantitative investigation in Section 5 will show that the linear synchronization of constituency and operator scope is a strong factor motivating W-Clefts in German.

Even though English does not exhibit a difference between main clause order and subordinate clause order, synchronization effects between form and function can also be observed here, though probably to a lesser extent. Sometimes propositional operators take scope over a clausal constituent in the subject position, while following that constituent in terms of linear order. The example in (46) is

a case in point. The scope relations are indicated in (47). In the corresponding canonical clause in (48), the scope-bearing operators occur towards the end of the sentence, even though they take scope over the gerund headed by *being* in subject position.

(46) What must not be allowed to happen is standards being imposed by the Commission through the back door. (EPEG-6/Ge 358730)

(47) What must not be allowed to happen is [standards ...]
$\Box$ [$\neg$ [$\Diamond$ [p]]]

(48) Standards being imposed by the Commission through the back door must not be allowed.

As there is no reason to assume that the reorganization of scope relations in W(H)-Clefts is more important in English than in German, we will assume that it is a general benefit (remember that there is no category of "German-specific" benefit in our investigation, for which the linear synchronization of constituency and operator scope would be an obvious candidate).

4.4 Avoiding left-heavy constituent structure

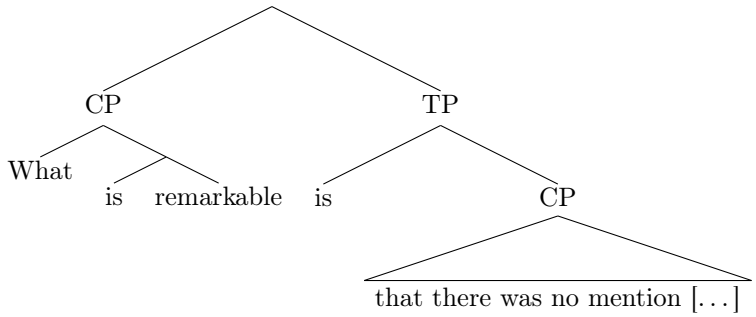
In some cases, W(H)-Clefts seem to be motivated in basically syntactic terms. In particular, they are sometimes used to avoid "left-heavy" constituent structure and have a function similar to that of extraposition. Consider the discourse sequence in (49). There is no obvious information structural motivation for the use of a WH-Cleft in this example, as remarkability is not, in any way, under discussion or topical, there is no contrast involved, and there is no scope-bearing operator, either (some discourse pragmatic motivation can of course always be accommodated).

(49) a. The first version of the report was used as a chance for them to get on their hobbyhorse and oppose nuclear energy.

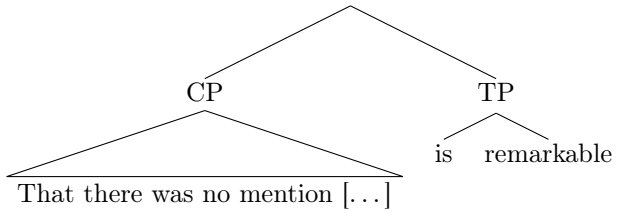
b. What is remarkable is that there was no mention, whatsoever, in the report of the Commission's action programme, which contains many sensible initiatives. (EPEG-6/Fr 119772)

As Hawkins (1994, 2004) has argued, languages tend to exhibit a homogeneous branching direction in their constituent structure, as such structural homogeneity facilitates parsing and, by hypothesis, production (cf. also Gast 2007 for discussion). English is a basically right-branching language and, like most other languages of this type, it avoids sentential arguments in a pre-verbal subject position. (50) provides a simplified tree diagram of the WH-Cleft in (49b). For comparison, the structure with a pre-verbal clausal subject is given in (51).

(50)



(51)



In German, W-Clefts are also sometimes used to restructure left-heavy sentences, i.e., sentences with a sentential argument in initial position. An example of this type is given in (52), the corresponding verb-second structure is shown in (53).

(52) Was dem Kautabak einen Vorteil verleiht, ist, daß der Nikotinabhängige ihn konsumieren kann, ohne seine Umgebung in Mitleidenschaft zu ziehen. (EPEG-6/Fr 58757)
'What gives snuff an advantage is the fact that nicotine addicts can use it without affecting the environment.'

(53) Daß der Nikotinabhängige den Kautabak konsumieren kann, ohne seine Umgebung in Mitleidenschaft zu ziehen, verleiht ihm einen Vorteil.

It should be noted that German, unlike English, has basic OV-order and is thus not *prima facie* right-branching. However, clausal constituents in the forefield (the position preceding the finite verb in verb-second languages) are also avoided, just like pre-verbal clausal subjects in English. Still, structural benefits like those illustrated in (49) and (50) are expected to have a weaker effect in German than in English. The reason is that in many cases, German does not need a W-Cleft to avoid left-heavy constituent structure as in (53). For instance, the German sentence corresponding to (the English example) (49) – here given in (54) – is not a W-Cleft. It exhibits canonical verb-second structure, with the adjective (in its predicative superlative form) in the forefield and the clausal subject in the middle field.[6] German can thus avoid a left-heavy constituent structure without resorting to the structural choice of a W-Cleft.

[6]The middle field is the field between the finite verb (in second position) and the non-finite verb (if any) at the end of the sentence.

(54) Am erstaunlichsten ist aber vielleicht, wie wenig Aufmerksamkeit dieses Thema bei den Eigentümern der Bank [...] findet [...].
Lit.: 'Most remarkable of all, however, is perhaps how little all this is discussed by the Bank's owners [...].'

We can thus classify the (structural) motivation of avoiding left-heavy constituent structure as English-specific, as it is expected to have a stronger positive distributional effect on the occurrence of W(H)-Clefts in English than in German.

4.5 An obstacle to the use of W(H)-Clefts: *Horror aequi*

So far, we have only considered the benefits associated with the use of W(H)-Clefts. In some cases, factors preventing the use of a W(H)-Cleft have also been mentioned, e.g. the fact that German can often achieve specific effects (like the avoidance of sentence-initial clausal constituents) without the use of a W(H)-Cleft. In this section, we will consider an additional factor that leads to asymmetries in the distribution of W(H)-Clefts in English and German, i.e., the tendency to avoid the immediate sequence of identical words (*horror aequi*) in sentences of the latter language.

The phenomenon of *horror aequi* often seems to prevent the use of Clefts in cases of nominal predication. If the W-clause contains a copula, the resulting sentence contains a sequence of two copulas (because of the verb-final order in subordinate clauses), i.e., a sequence of the type [... *ist, ist, dass* ...]. This is, obviously, often avoided for stylistic and perhaps also linguistic reasons. One avoidance strategy found in the German corpus part is the division of the sentence into two parts separated by a colon, as in (56), which corresponds to (the English example) (55). (57) provides the German sentence realized as a W-Cleft, with the mentioned *horror aequi*-problem.

(55) What is even more ridiculous is that surpluses are produced even under normal conditions. (EPEG-6/Ge 1737040)

(56) Noch viel unsinniger: Schon unter normalen Bedingungen werden Überschüsse produziert.

(57) Was noch viel unsinniger ist, ist, dass schon unter normalen Bedingungen Überschüsse produziert werden.

As there are no comparable effects observable in the English data, the *horror aequi* obstacle is expected to be a factor that decreases the chances of finding a W-Cleft in German as a counterpart of an English WH-Cleft.

4.6 The hypotheses tested in this study

Let us take stock of the observations made so far and formulate some hypotheses for our empirical study. We have proposed four motivations for the use of W(H)-Clefts. Moreover, we have identified one obstacle which prevents the use of a

W-Cleft in German, i.e., the adjacency of two copulas in German W-Clefts with nominal predication in the matrix clause. The benefits have been categorized into two classes, English-specific ones and general ones. English-specific benefits are assumed to motivate W(H)-Clefts to a greater extent in English than in German, while general benefits are not expected to have any such "pro-English" effect. The two classes of benefits and the obstacle are summarized in Table 1.

Table 1: Benefits of and obstacles to the use of W(H)-Clefts

English-specific benefits
1. Structural separation of propositional content and utterance comment (Sect. 4.2)
2. Avoidance of left-heavy constituent structure (Sect. 4.4)
General benefits
3. Phonological separation of Topic and Comment (creation of an IP-boundary, Sect. 4.1)
4. Linear synchronization of constituency and operator scope (Sect. 4.3)
Obstacle (German)
5. Immediate sequence of two copulas (*horror aequi*, Sect. 4.5)

The distribution of W-Clefts in the German data is expected to be a function of the interplay between the factors summarized in Table 1 (plus some other factors which have not been taken into account). Overall, the ratio of German W-Clefts to English WH-Clefts has been shown to be approximately 1:4 in the EUROPARL-corpus (cf. Gast & Wiechmann 2012). Given that in the present (asymmetrical) study we have disregarded those cases where we find a W-Cleft in German while not finding a WH-Cleft in English, the ratio is even lower in our data. Of the 722 cases of WH-Clefts used for this study, 116 correspond to German W-Clefts (cf. Table 2). The ratio of German W-Clefts to English WH-Clefts is thus approximately 1:5.

Table 2: W-Clefts and alternative structures in the German data

	W-Cleft		ratio of W-Clefts
	TRUE	FALSE	to alternative structures
German correspondences of English WH-Clefts	116	606	1 : 5.2 (= 116/606)

The proportion of German W-Clefts in the dataset is approximarely 0.16 (=116/722). This figure is assumed to be the result of two antagonistic forces. The English-specific benefits as well as the *horror aequi*-obstacle pull the figure towards a proportion even lower than 0.16. By contrast, the general benefits – while obviously not leading to a totally even distribution of W(H)-Clefts in English and German – are expected to have a positive impact on the occurrence

of W-Clefts in the German data, in comparison to the overall proportion of W-Clefts.

On the basis of the observations made so far, we can formulate the hypotheses in (58). Each hypothesis implies a claim about the frequency of W-Clefts in the German data, under specific conditions observed in the English data. The terms "over"- and "underrepresentation" do not refer to the German as opposed to English data, but to the distribution of W-Clefts within the German data. The descriptive labels of the hypotheses are introduced for future reference.

(58) **Hypotheses**

a. **The phonological separation hypothesis**
If an English WH-Cleft is motivated by the placement of a nuclear accent in the WH-clause, which implies the presence of an IP boundary between the Topic and the Comment, German W-Clefts will be overrepresented in the corresponding sentences, relative to the distribution of W-clefts in the entire German data set.

b. **The utterance comment hypothesis**
If an English WH-Cleft is motivated by the structural separation of propositional content and utterance comment, German W-Clefts will be underrepresented in the corresponding sentences, relative to the distribution of W-Clefts in the entire German data set.

c. **The scope hypothesis**
If an English WH-Cleft is motivated by the synchronization of constituency and operator scope, German W-Clefts will be overrepresented in the corresponding sentences, relative to the distribution of W-Clefts in the entire German data set.

d. **The structural hypothesis**
If an English WH-Cleft is motivated by the avoidance of left-heavy constituent structure, German W-Clefts will be underrepresented in the corresponding sentences, relative to the distribution of W-Clefts in the entire German data set.

e. **The *horror aequi* hypothesis**
If the WH-clause of an English WH-Cleft contains a nominal predication, German W-Clefts will be underrepresented in the corresponding sentences, relative to the distribution of W-Clefts in the entire German data set.

We will now proceed to the operationalizations and testing of the hypotheses made in (58).

5 Testing the hypotheses

In order to test our hypotheses, we extracted 800 examples of English WH-Clefts from the EUROPARL-corpus (cf. Gast & Wiechmann 2012 for some details of

the process of data extraction). During the process of coding, several instances of presumable WH-Clefts were identified as not actually representing a structure of this type. Most importantly, the original data set contained a number of examples which were instances of nominal predication with a free relative clause in subject position. Consider (59) and (60).

(59) What we need is a hero.

(60) What you just said is a lie.

On the face of it, the two examples are structurally entirely parallel. There is an important difference in the interpretation of the copula, however. In (59), the copula is interpreted as "specificational" (cf. Declerck 1984; den Dikken 2009; Gast & Wiechmann 2012), i.e., it specifies a variable in an open proposition ("We need x, x = a hero"). By contrast, (60) attributes a property to an abstract referent described by the free relative clause *what you just said* (i.e., "the utterance that you just made"). While being more or less identical structurally, the two types of predication are quite different in terms of their semantics. Most importantly, predicational uses of the copula imply the attribution of a property to a referent, and are thus *about* this referent, whereas specificational uses provide information about a predicate or open proposition.

In order to differentiate the two uses of a copula, we need a test. While a copula is systematically ambiguous between a predicational and a specificational reading, other, near equivalent predicates taking nominal complements do not display this ambiguity. We have used the predicate *consider* in order to determine whether a given instance of a copula is specificational or predicational and, hence, whether or not the corresponding sentence qualifies as a W(H)-Cleft. A copula is regarded as predicational iff "A is B" can be paraphrased as "I consider A B". This paraphrase is possible for (60), but not for (59):

(61) ?I consider what we need a hero. ($\neq$ [59])

(62) I consider what you just said a lie. ($\sim$ [60])

On the basis of this test, we excluded some data points, thus reducing the data set to 722 occurrences of WH-Clefts in English.

The data was coded for the variables described in this section. They are the independent variables of the study. Moreover, we coded the data for the type of construction found in the corresponding German sentence, the dependent variable. For the purposes of the present study, this variable was binary, i.e., we determined whether or not there was a W-Cleft in the relevant German sentence.

We will now discuss each hypothesis separately, considering the operationalizations used in each case as well as the results obtained.

5.1 The phonological separation hypothesis

The phonological separation hypothesis says that W(H)-Clefts are formed in order to create an IP-boundary between the Topic and the Comment, to a more

or less comparable extent in English and German. Unfortunately, EUROPARL is a written corpus, so we cannot use any phonological evidence to test whether W(H)-Clefts are in fact associated with the formation of separate intonation phrases.

What we have done, instead, is determine whether or not there is any type of contrast in the W(H)-clause which is recoverable from the immediate discourse environment. Such contrast is expected to be reflected phonologically in a contrastive focus accent. A contrastive accent, in turn, by definition requires the formation of a separate intonation phrase. Our operationalization of the phonological separation hypothesis is described in (63).

(63) The phonological separation hypothesis (operationalization)
W-Clefts are expected to be overrepresented in the German data (relative to the entire German data set) if in the English data there is an explicit contrast between an element of the WH-clause and some other element in the discourse environment.

The notion of "contrast" was interpreted narrowly. A data point was coded as "TRUE" iff an overt constituent was found in the immediate discourse environment which contrasted with the Cleft constituent of the WH-Cleft. In typical cases, there is a binary contrast, e.g. between positive and negative polarity. In (64a), there is an explicit contrast between *what* [*the report*] *says and does not say*. The WH-clause of (64b) picks up the positive option (*what the report does say*). In this case, the contrast is also overtly marked through *do*-support, which is specialized for the function of *verum focus*.

(64) a. Therefore, what remains of concern to us is not in the main the contents or analysis within the report but what it says and does not say about the performance of the structural funds programmes themselves.

b. What the report does say is that there are significant and damaging delays in payments to localities and regions [...] (EPEG-6/En 13261979

The results of our quantitative study are shown in Table 3. The frequencies that are expected on the assumption that the variable *contrast* does not have any effect are given for comparison, with the figures rounded to whole numbers.

Table 3: The influence of the independent variable *contrast* on the occurrence of W-Clefts in German ($p = 0.0004$, $OR = 2.18$, $\phi = 0.14$)

	W-Cleft in the German data			
	observed frequencies		expected frequencies	
contrast	TRUE	FALSE	TRUE	FALSE
TRUE	48	148	23	165
FALSE	68	458	85	441

Fisher's exact test delivers a p-value < 0.001, i.e., the impact of the variable *contrast in the* WH-*phrase* on the distribution of W-Clefts is very highly significant. The odds ratio is higher than 2, which means that, in our data set, the presence of contrast in the WH-phrase increases the chances of finding a W-Cleft in the corresponding German sentence by a factor of more than 2. Given that the odds ratio is not sensitive to absolute numbers and, hence, not a very good indicator of effect size, we also used the phi-coefficient (ϕ), which is calculated as follows:

(65) The phi-coefficient ϕ

For a contingency table of the form

	¬A	A
¬B	a	b
B	c	d

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

In the case of 2 by 2 tables, ϕ is equal to the absolute value of another popular measure, Cramer's V. Like Pearson's correlation coefficient r, ϕ ranges from –1 (complete negative association) to +1 (complete positive association). 0 indicates no association. The following relation holds between ϕ and χ^2:

(66) $\phi^2 = \chi^2/\mathrm{n}$

For the benefit considered in this section, the presence of contrast in the WH-clause and the associated presence of an IP-boundary, ϕ has a value of 0.14. As all other ϕ-values determined for our data, this is considered a weak association (values lower than .25 are considered weak). Given that we are dealing with five variables each of which has a relatively low number of TRUE cases, this is not surprising. Determining ϕ-values for our data will allow us to rank the variables in terms of importance, however (cf. Section 5.6).

5.2 The utterance comment hypothesis

According to the utterance comment hypothesis, W(H)-Clefts are sometimes used to structurally separate the propositional content from a frame of evaluation, more commonly in English than in German. This hypothesis can be operationalized as in (67).

(67) The utterance comment hypothesis
W-Clefts are underrepresented in the German data (relative to the entire German data set) if

a. the WH-clause of an English WH-Cleft contains a propositional complement taking predicate,

b. the Cleft constituent contains a clause complementing the matrix predicate, and

c. the embedded clause is truth-conditionally equivalent to the entire sentence.

For illustration, consider (68).

(68) What I mean is that we are not starting from square one and, as a producer of several tonnes of beef a year, I know what I am talking about. (EPEG-6/Ge 37948)

The truth conditions of the embedded clause and of the entire sentence are the same, as is illustrated by the equivalence of (68) and (69):

(69) We are not starting from square one and, as a producer of several tonnes of beef a year, I know what I am talking about.

Our results are displayed in Table 4. As the p-value indicates, the impact of the independent variable under analysis is highly significant, and the hypothesis can be assumed to have been confirmed. The odds ratio is 0.3. In other words, if an English WH-Cleft is used to separate an utterance comment from the propositional content, the probability of finding a W-Cleft in the corresponding sentence of our German data decreases by a factor of more than three. The phi-coefficient is –0.1.

Table 4: The influence of the independent variable *utterance comment* on the occurrence of W-Clefts in German ($p = 0.0065$, OR = 0.3, $\phi = -0.1$)

	W-Cleft in the German data			
	observed frequencies		expected frequencies	
utterance comment	TRUE	FALSE	TRUE	FALSE
TRUE	5	79	12	71
FALSE	111	527	103	535

5.3 The scope hypothesis

The scope hypothesis has so far been phrased generally in terms of scope-bearing operators. More specifically, we have tested whether the co-occurrence of a modal and a negator in the W(H)-clause of a Cleft sentence has any effect on the distribution of W-Clefts in the German data. The scope hypothesis is thus operationalized as follows:

(70) The scope hypothesis
W-Clefts are overrepresented in the German data (relative to the entire German data set) if the WH-clause of an English WH-Cleft contains both a modal and a negator.

The number of WH-Clefts with a modal *and* a negator in the WH-clause was obviously rather small. Still, Fisher's exact test shows that this variable has a very highly significant (positive) effect on the occurrence of W-Clefts in the German data ($p < 0.001$). The odds ratio is 10.4, and the phi-coefficient is 0.206. The data are shown in Table 5.

Table 5: The impact of the independent variable *modal and negator in* WH-*clause* on the occurrence of W-Clefts in German ($p < 0.001$, OR = 10.4, $\phi = 0.206$)

	W-Cleft in the German data			
modal and negator	observed frequencies		expected frequencies	
in WH-clause	TRUE	FALSE	TRUE	FALSE
TRUE	11	6	3	14
FALSE	105	600	113	592

5.4 The structural hypothesis

Left-heavy structures of the type discussed in Section 4.4 arise when a clause functions as the subject of a sentence. According to our "structural hypothesis", W(H)-Clefts are used to avoid such left-heavy structures. The benefit of avoiding left-heavy constituent structure can thus simply be operationalized in terms of the category and function of the Cleft constituent, and the structural hypothesis can be rephrased as follows:

(71) Structural hypothesis
W-Clefts are underrepresented in the German data (relative to the entire German data set) if the Cleft constituent is a clause in subject function.

The results of our study are shown in Table 6.

Table 6: The impact of the independent variable *clausal subject* on the occurrence of W-Clefts in German ($p < 0.001$, OR = 0.18, $\phi = -0.16$)

	W-Cleft in the German data			
	observed frequencies		expected frequencies	
clausal subject	TRUE	FALSE	TRUE	FALSE
TRUE	5	124	17	107
FALSE	111	482	96	499

Fisher's exact test delivers a p-value lower than 0.001, so we can consider our hypothesis confirmed. The odds ratio is approximately 0.2. In other words, if the Cleft constituent of an English WH-Cleft is a clause in subject function, the probability of finding a W-Cleft in the corresponding German sentences is five times lower than otherwise. The phi-coefficient is –0.16.

5.5 The *horror aequi*-hypothesis

The obstacle of having two identical words following each other emerges when the W(H)-clause contains a nominal predicate headed by a copula. We can thus rephrase the corresponding hypothesis as in (72).

(72) The *horror aequi* hypothesis

W-Clefts are underrepresented in the German data (relative to the entire German data set) if the WH-clause contains a nominal predicate headed by a copula.

The notion of "nominal predicate" is interpreted relatively broadly, as it covers passive predications of the type *what is needed / required*, in addition to more prototypical cases like *what is necessary*.

The results of our study are shown in Table 7. The variable *governing nominal predicate* is the only variable that we have investigated which delivers a p-value higher than 0.01, but rather minimally so (p = 0.013). The result is still significant and we consider our hypothesis confirmed. The odds ratio is 0.5, which means that the presence of a nominal predicate in the WH-clause of an English Cleft sentence decreases the chances of finding a W-Cleft in the German data by a factor of two. The phi-coefficient is –0.093.

Table 7: The impact of the independent variable *governing nominal predicate* on the occurrence of W-Clefts in German (p = 0.013, OR = 0.51, ϕ = –0.093)

	W-Cleft in the German data			
governing	observed frequencies		expected frequencies	
nominal predicate	TRUE	FALSE	TRUE	FALSE
TRUE	18	160	22	149
FALSE	98	446	87	457

5.6 Summary

The results obtained in our study and reported on in the previous sections are summarized in Table 8. The hypotheses are ordered by the absolute (rounded) phi-values of the relevant distributions.

As can be seen from Table 8, the presence of a modal and negation in the WH-clause (as an operationalization of the scope hypothesis) is the variable showing the strongest correlation with the presence or absence of a W-Cleft in German, followed by the variables *clausal subject* and *contrast in* W*(*H*)-clause*. *Governing metalinguistic predicate* and *nominal predicate within the* W*(*H*)-clause* are the variables showing the weakest correlations with the dependent variable, although the association is still statistically significant.

The results displayed in Table 8 obviously have to be taken with caution. It is well known that the individual effect of some independent variable can be distorted by correlations with other variables. We have therefore carried out a multifactorial analysis that is intended to show which variables are intercorrelated, and which variables form their own dimensions of variation.

Table 8: Effect sizes and p-values for all independent variables, ordered by importance (the absolute value of ϕ)

Hypothesis	variable	ϕ	OR	p-value (Fisher)
The scope hypothesis	MOD+NEG in WH-clause	0.21	10.4	< 0.001
The structural hypothesis	clausal subject	-0.16	0.18	< 0.001
The phonological separation hypothesis	contrast	0.14	2.18	< 0.001
The utterance comment hypothesis	gov. metaling. pred.	-0.1	0.3	0.0065
The *horror aequi* hypothesis	nom. pred. in WH-clause	-0.09	0.51	0.013

6 Determining correlations between variables: Multiple Correspondence Analysis

The method that we have used to determine correlations between variables is Multiple Correspondence Analysis (MCA) with supplementary points. MCA is a technique suitable for the visualization and exploration of multidimensional contingency tables. This is exactly what we need if we cross-tabulate all our variables, and not only pairs of variables, as in the previous section. Like Multidimensional Scaling and Principle Component Analysis, MCA represents multidimensional data in models with a lower number of dimensions, most commonly two or three. However, unlike its relatives, MCA was developed specifically for categorical (non-numeric) variables with such values as TRUE or FALSE, heads or tails, male or female, smoker or non-smoker, etc. In our data, all variables are categorical and binary (TRUE or FALSE). The distances on a correspondence analysis plot represent associations between variables, formalized as chi-squared distances. The dependent variable, i.e., the presense or absense of a W-Cleft in the German data, will be treated as a supplementary variable. In the Correspodence Analysis jargon, a supplementary status means that the variable will not infuence the orientation of the dimensions of variation. In this way we will preserve the special status of the dependent variable and separate the properties of the English sentences from those of the German ones, and we can map the German structures to the English ones. For more details and another application of this method in a contrastive study, see Levshina et al. (2013).

The resulting map is shown in Figure 1. The two-dimensional solution accounts for 73.6% of variation in the data. This means that 26.4% of information is not captured by the solution. For Multiple Correspondence Analysis, this can be considered an acceptable result. The features that are near the centre of the plot are those with the highest frequencies in the data. Those are the default values, such as a lack of both negation and modality (*scopeNo*), the absense of

an utterance comment (*meta_commNo*), and the lack of contrast in the WH-clause (*contrastNo*). The majority of observations (60%) contain those default values.

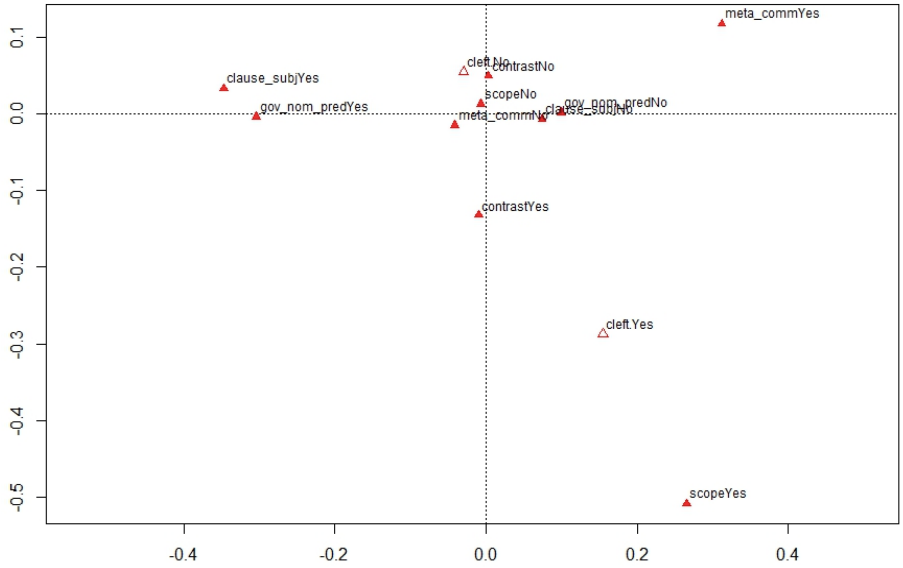


Figure 1: A Multiple Correspondence Analysis of the variables investigated in the present study

Most variation (60.8%) is captured by the horizontal dimension, and only 12.8% by the vertical dimension. This suggests that intercorrelated variables are mostly spread along the horizontal dimension. Especially strongly associated are the variables *clausal subject* and *governing nominal predicate*: *clause_subjYes* and *gov_nom_predYes* are located very closely to each other in the left part of the plot, whereas their negative counterparts are found in the right part. A glance at our data reveals that the two values indeed often go hand in hand, as clausal subjects are often headed by a copula: Of the 155 cases with a clausal subject, 92 have a nominal predicate in the WH-clause (59%). An example of this type is given in (55):

(55) [What is even more ridiculous$_{\text{NOM_PRED}}$] is
[that surpluses are produced even under normal conditions]$_{\text{SUBJ}}$.

The presence of utterance comments (*meta_commYes*), which is represented in the top right corner of the map, is not strongly associated with any other features, and it co-occurs extremely rarely with clausal subjects (*clause_subjYes*) and with nominal predicates in the WH-clause (*gov_nom_predYes*), as can be seen from their opposite locations on the map. In a sense, *meta_commYes* constitutes its own functional dimension.

As for the vertical axis of the map, it is formed by the presence of modality and negation (*scopeYes*), and to some extent by the presence of contrast (*contrastYes*). These features do not co-occur frequently with any other variables and therefore constitute the third major type of context. The relatively close association of these variables is not surprising, since negation, as the reader may recall from Section 5.1, can be used to create a contrast. A relevant example with this combination (modal / negation in W[H]-clause and contrast) is given in (73b). The preceding context is given in (73a).

(73) a. I can also assure the honourable Member that there is never any problem with purely local initiatives, since they can never be regarded as distorting trans-frontier competition. There is no reason why they should not go ahead.

b. What we cannot allow are measures specifically designed to distort competition with neighbouring countries in particular, or measures which are claimed to be aid to employment, but which are really also designed to give firms an advantage in a given sector. (EPEG-6/Du 1306258)

The contrast is between the *admissibility* of local iniatives and the *inadmissibility* of trans-frontier competition, or, more specifically, between positive and negative polarity (*What we can*NOT *allow ...*). Note that there is, of course, also a contrast in the Cleft constituent, but we have disregarded such contrastive constituents in the Cleft constituent, as they do not (necessarily) introduce an IP-boundary.

To summarize our results, it seems that we can distinguish three major types of English WH-Clefts, each of them associated with a specific benefit:

(74) a. **Structural W(H)-Clefts**
W(H)-Clefts with clausal subjects
Benefit: avoidance of left-heavy constituent structure

b. **Metalinguistic W(H)-Clefts**
W(H)-Clefts with an utterance comment in the matrix clause
Benefit: structural separation of propositional content and utterance comment

c. **Contrastive W(H)-Clefts**
W(H)-Clefts with contrast in the WH-phrase
Benefit: phonological separation of Topic and Comment

W(H)-Clefts of type (74a) often have a nominal predicate in the W(H)-clause, those of type (74c) (relatively) often contain more than one scope-bearing operator.

Let us now examine the mapping of the dependent variable, i.e., the presence or absence of a W-Cleft in the German data, which is plotted onto the space constituted by the English Clefts in Figure 1. The positions of the two levels – *cleftYes* and *cleftNo*, see the empty triangles – is determined by the

association with the independent variables. Obviously, German W-Clefts (*cleftYes*) are associated predominantly with the third functional dimension of the corresponding English construction, represented by the features *scopeYes* and *contrastYes*. This is not surprising if we consider that both scope interactions in the WH-clause and contrast are factors that have been subsumed under the general benefits, i.e., benefits that are expected to be reflected in English and German more or less alike.

With respect to the three types of W(H)-Clefts distinguished in (74), we can thus say that only one of them – type (74c) – is prominent in German, whereas types (74a) and (74b) are characteristic of English but not of German, and are thus mainly responsible for the differential distribution of W(H)-Clefts in English and German.

The linguistic interpretation of Figure 1 is shown in Figure 2. Two of the three types of W(H)-Clefts emerging from this map are labelled "English-specific", one of them "general", in analogy to the attributes used for benefits. Remember that "general" means "non-English-specific", i.e., in the present context, widely attested in both English and German.

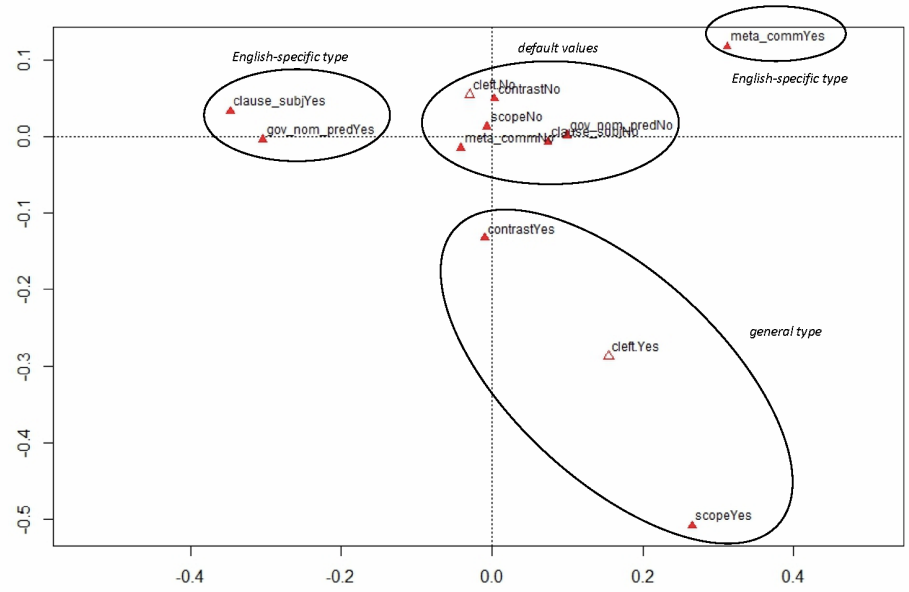


Figure 2: Major types of W(H)-clefts as (sets of) variables

7 Summary and conclusions

In this contribution we have argued that the distribution of W(H)-Clefts is not just a function of information structural factors, in particular Topic-Comment

structure, but also of other effects which are (partly) independent of information structure. We have assumed that the linear synchronization of Topic and Comment represents a necessary condition for the use of a W(H)-Cleft – allowing for some degree of accommodation – but that the probability of finding a W(H)-Cleft in a given data set is also influenced by other factors. Factors favouring a W(H)-Cleft have been called "benefits", and four benefits of using a W(H)-Cleft have been identified: (i) the phonological separation of Topic and Comment, (ii) the structural separation of propositional content and utterance comment, (iii) the linear synchronization of scope relations and constituency, and (iv) the avoidance of left-heavy constituent structure. In addition, the distribution of W(H)-Clefts has been regarded as being determined by "obstacles", and one particular obstacle, the immediate co-occurrence of two copulas in German, has been taken into account. These hypotheses have been tested quantitatively, and it has been shown that all of the factors have a significant effect in our data set, but that two of the variables are closely interrelated. We have therefore distinguished three major types of English WH-Clefts, only one of which is often rendered as a W-Cleft in German.

On the basis of these observations, we can make the following generalizations concerning the differential distribution of W(H)-Clefts in English and German. In addition to the five contingent (i.e., non-necessary) benefits discussed in Section 4, this summary makes reference to the standard condition for W(H)-Clefts as well (cf. Section 3):

(75) Generalizations concerning the differential distribution of W(H)-Clefts in English and in German

a. Both English WH-Clefts and German W-Clefts are standardly motivated by the linear synchronization of Topic and Comment (or Quaestio and Responsio).

b. Both English WH-Clefts and German W-Clefts are often motivated by the phonological separation of Topic and Comment (or Quaestio and Responsio), which is required if the W(H)-clause contains a contrastive constituent. Such occurrences of W(H)-Clefts (relatively) often contain more than one scope-bearing operator.

c. English WH-Clefts are often motivated by the avoidance of left-heavy constituent structure emerging from a clausal subject in preverbal position. In German, such structurally motivated W-Clefts seem to be rare, as left-heavy constituent structure can also be avoided in canonical (verb-second) clauses. Moreover, the relative rarity of structural W-Clefts in German is probably partly due to the immediate sequence of two copulas which results when the W-clause is headed by a copula (*horror aequi*).

d. English WH-Clefts are often motivated by the structural separation of propositional content and utterance comment. Such instances of W-Clefts seem to be rarer in German, where embedded verb-second clauses are often used instead.

Obviously, the generalizations made in (75) do not describe the (differential) distribution of W(H)-Clefts in English and German exhaustively, and it should have become obvious that there is still a lot of work to do if we want to properly understand the (manifold) reasons why speakers use WH-Clefts in English and W-Clefts in German. Still, we hope to have made some progress towards a better understanding of that matter.

References

Akmajian, Adrian. 1970. On deriving cleft sentences from pseudo cleft sentences. *Linguistic Inquiry* 1. 149–168.

Altmann, Hans. 1981. *Formen der Herausstellung im Deutschen: Rechtsversetzung, Linksversetzung, freies Thema und verwandte Konstruktionen*. Number 106 in Linguistische Arbeiten. Tübingen: Niemeyer.

Altmann, Hans. 2009. Cleft- und Pseudocleft-Sätze (Spalt- und Sperrsätze) im Deutschen. In Brdar-Szabó, Rita, Elisabeth Knipf-Komlósi & Attila Péteri (eds.), *An der Grenze zwischen Grammatik und Pragmatik*, 13–34. Frankfurt: Peter Lang.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Büring, Daniel. 2003. On D-Trees, Beans, and B-Accents. *Linguistics and Philosophy* 26. 511–545.

Cartoni, Bruno & Thomas Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*. Istanbul.

Collins, Peter. 1991. *Cleft and Pseudo-Cleft Constructions in English*. London & New York: Routledge.

Collins, Peter. 2006. *It*-clefts and *wh*-clefts: Prosody and pragmatics. *Journal of Pragmatics* 38. 1706–1720.

Declerck, Renaat. 1984. The pragmatics of *it*-clefts and *wh*-clefts. *Lingua* 64. 251–289.

den Dikken, Marcel. 2009. Predication and specification in the syntax of cleft-sentences. Ms., CUNY Graduate Center, 24 pp.

Dufter, Andreas. 2009. Clefting and discourse organization: Comparing Germanic and Romance. In Andreas Dufter & Daniel Jacob (eds.), *Focus and Background in Romance Languages*, 83–121. Amsterdam & Philadelphia: John Benjamins.

Faraci, Robert. 1971. The deep question of pseudo-clefts. *English Linguistics* 6. 48–85.

Fischer, Klaus. 2009. Cleft sentences: Form, function and translation. *Journal of Germanic Linguistics* 21. 167–191.

Gast, Volker. 2007. From phylogenetic diversity to structural homogeneity – on right-branching constituent order in Mesoamerica. *SKY Journal of Linguistics* 20. 171–202.

Gast, Volker. 2010. Contrastive topics and distributed foci as instances of subinformativity: A comparison of English and German. In Carsten Breul & Edward Göbbel (eds.), *Comparative and Contrastive Studies of Information Structure*, 15–50. Amsterdam & Philadelphia: John Benjamins.

Gast, Volker. forthcoming. Contrastive linguistics: Theories and methods. In Bernd Kortmann & Johannes Kabatek (eds.), *Dictionaries of Linguistics and Communication Science: Linguistic Theory and Methodology*. Berlin & New York: de Gruyter Mouton.

Gast, Volker & Johan van der Auwera. 2011. Scalar additive operators in the languages of Europe. *Language* 87. 2–54.

Gast, Volker & Daniel Wiechmann. 2012. W(h)-clefts im Deutschen und Englischen. Eine quantitative Untersuchung auf Grundlage des Europarl-Korpus. In Lutz Gunkel & Gisela Zifonun (eds.), *Jahrbuch des IDS 2011*, 333–362. Berlin & New York: de Gruyter Mouton.

Gundel, Jeanette K. & Thorsten Fretheim. 2004. Topic and focus. In Lawrence Horn & Gregory Ward (eds.), *The Handbook of Pragmatics*, 175–196. London: Blackwell.

Halliday, Michael A.K. 1967. *Intonation and Grammar in British English*. The Hague: Mouton.

Hawkins, John. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, John. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hedberg, Nancy & Lorna Fadden. 2007. The information structure of *it*-clefts, *wh*-clefts and reverse *wh*-clefts in English. In Nancy Hedberg & Ron Zacharski (eds.), *The Grammar-Pragmatics Interface. Essays in Honor of Jeanette K. Gundel*, volume 155 of *Pragmatics & Beyond*, 19–48. Amsterdam & Philadelphia: John Benjamins.

Jacobs, Joachim. 2001. The dimensions of topic–comment. *Linguistics* 39. 641–681.

Jespersen, Otto. 1937. *Analytic Syntax*. London: Unwill.

Klein, Wolfgang & Christiane von Stutterheim. 1987. Textstruktur und referentielle Bewegung. *Linguistische Berichte* 109. 67–92.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. Phuket. MT Summit X.

König, Ekkehard & Volker Gast. 2012. *Understanding English-German Contrasts*. Berlin: Erich Schmidt-Verlag, 3rd edition.

Krifka, Manfred. 2007. Basic notions of information structure. In Manfred Krifka & Caroline Fery (eds.), *Interdisciplinary Studies of Information Structure 6*, 13–56. Potsdam: Universitätsverlag.

Lambrecht, Knud. 2001. A framework for the analysis of cleft constructions. *Linguistics* 39. 463–516.

Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. 2013. Mapping constructional spaces: A contrastive analysis of English and Dutch analytic causatives. *Linguistics* 51. 825–854.

Prince, Alan & Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University Cognitive Science.

Prince, Ellen. 1978. A comparison of wh-clefts and it-clefts in discourse. *Language* 54. 883–906.

Quirk, Randolph, Sydney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *The Comprehensive Grammar of the English Language*. London: Longman.

Wells, John C. 2006. *English Intonation: An Introduction*. Cambridge: Cambridge University Press.