SEMANTIC TRANSPARENCY AND ANAPHORIC ISLANDS

MARTIN SCHÄFER

This paper investigates the relationship between semantic transparency of compounds and their status as anaphoric islands. More specifically, I will take a detailed look at the behaviour of German adjective-noun compounds in this respect. The paper argues that semantic transparency plays a crucial role in accessing compound-internal components for anaphoric reference and discusses a number of factors that motivate the actual usage of anaphora.

1 INTRODUCTION

Postal (1969) argued that words, whether monomorphemic or derived, are anaphoric islands.[1] That is, neither internal constituents of morphologically complex words nor entities contained in the meaning of a word can serve as antecedents to a following anaphoric element, or, in Postal's words, allow outbound anaphora.[2] Two pieces of data supporting his claim are reproduced in (1) and (2), his (3) and (53), respectively.

(1) a. Max's parents$_i$ are dead and he deeply misses them$_i$.

     b. *Max is an orphan and he deeply misses them.

(2) a. Harry was looking for a rack for books$_i$ but he only found racks for very small ones$_i$.

     b. *Harry was looking for a bookrack but he only found racks for very small ones.

---

[1] Note that Postal's anaphoric islands are only terminologically similar to Ross's (1967) island phenomena.

[2] Besides outbound anaphora, Postal also discusses inbound anaphora. Harris (2006) provides a comprehensive discussion of inbound anaphora.

In (1a), the pronoun *them* refers anaphorically to the referent of the noun phrase *Max's parents*. In contrast, in (1b) *them* cannot refer to Max's parents, although the meaning of the word *orphan* "involves reference to the parents of an individual." (Postal 1969:206). Example (2) aims to show that even complex words are islands. In the first sentence, the noun phrase *books* can be picked up by the pro-form *one* in the second clause, contrasting with the morpheme *book* contained in the compound *bookrack* in the second sentence, which cannot be picked up by the pro-form *one*.

In contrast to Postal's findings, German adjective-noun-compounds do not act as barriers for anaphoric reference, cf. (3) and (4).

(3)   Ich bin das Grünglas losgeworden, das weiße liegt noch im Auto.

'I am the green-glass got.rid.off, the white lies still in.the car'

i.e. I got rid of the green glass, the white glass is still in the car.

(4)   Ich liebe Großstädte, in kleinen gehe ich ein.

'I love big-towns, in small go I in' [cf. *ein-gehen* 'to perish']

i.e. I love big cities, I cannot exist in small cities.

Standardly, both examples are interpreted as involving anaphoric reference by an empty element to the head of the compound. Thus, we can assume the structure in (5) for (4), where PRO stands for an empty element:

(5)   Ich liebe [Groß[städte] $_i$], in kleinen PRO$_i$ gehe ich ein.

On the other hand, this kind of anaphoric reference does not seem to be available across the board, as e.g. (6) shows, where # marks pragmatic deviance.

(6)   #Mein Vater hat in seinem Garten schon mal einen Grünspecht gesehen, aber noch nie einen schwarzen.

'My father has in his garden already once a green-woodpecker seen but so far never a black.'

Intended: My father once saw a green woodpecker in his garden, but he has never seen a black woodpecker.

In (6), we have the same basic configuration, that is, an A N construction, followed by an A pro-form construction, but the strongly preferred interpretation is with anaphoric reference to the whole compound, not to just its head. Thus, instead of the intended interpretation of the second, elliptic A N construction as *Schwarzspecht* 'black woodpecker', the preferred interpretation is *schwarzer Grünspecht*, 'black green woodpecker'.

These two points, first, that there are counterexamples to anaphoric islandhood, and secondly, that anaphoric reference is sometimes more or less easily available, are not surprising, given that both aspects have been repeatedly pointed out in the literature on English. Why then this paper with its focus on A N compounds in German? Firstly, two recent papers dealing with A N constructions use the status of compounds as anaphoric islands as a test for assigning compound or phrasal status to A N constructions for which no other measures are available, cf. Giegerich (2005) for English and Paul (2005) for Mandarin Chinese A N constructions, making it worthwhile to gather together the counterevidence to the idea of anaphoric islandhood as a syntactic test once again. Secondly, the pattern found in the German data cannot be found in English, allowing us to elaborate on the mechanism behind these types of anaphoric relations.

The paper is organized as follows: section 2 introduces two rival accounts of anaphoric islandhood data and reviews the psycholinguistic evidence discussed in this

context. Section 3 discusses the details of the German data and some findings from an explorative corpus study. Section 4 concludes the paper.

2 ACCOUNTING FOR ANAPHORIC ISLANDS

The pattern in Postal's data has been interpreted in two ways. In the first view, the pattern is taken as evidence for the existence of an anaphoric island and the relationship between the anaphoric element and its antecedent is seen as a syntactic relation. Based on this interpretation, the pattern has also been used as a diagnostic for compoundhood. In the second view, the key to the patterns lies in the pragmatics involved. I will discuss the two accounts in turn, dismissing the former in favour of the latter.

2.1 A syntactic interpretation

A syntactic interpretation contains two core ingredients: (a) a classical anaphora account and (b) a strong version of the lexicalist hypothesis (also referred to as lexical integrity hypothesis). (7) contains a typical statement of the classical anaphora account, a quote from Hankamer and Sag's description of what "by now may be called the 'classical' position" (1976:394).

(7)   Classical anaphora account:

"[A]ll anaphoric processes are transformations that involve deletion (or conversion to a pro-form) of an underlyingly present, fully lexical segment under conditions of identity with an antecedent segment; […]"

Hankamer and Sag (1976:394)

The lexicalist hypothesis can be likewise given in a maximally strong form, cf.(8).

(8) Lexicalist Hypothesis [maximally strong version]

"The syntax neither manipulates nor has access to the internal form of words."

Anderson (1992:84)[3]

In weaker variants of the lexicalist hypothesis, anaphoric reference to word-internal constituents is allowed, cf. Selkirk's (1982:70) *Word Structure Autonomy Condition*.

With (7) and (8) in place, the difference in the pair of example sentences in (2), repeated here for convenience, is accounted for as follows.

(9) a. Harry was looking for a rack for books$_i$ but he only found racks for very small ones$_i$.

b. * Harry was looking for a bookrack but he only found racks for very small ones.

The underlying form of *small ones* in Sentence (9a) is *small books*, where *books*, stripped of its inflectional ending, is a fully lexical segment that is identical to the preceding segment *book* in the phrase *a rack for books*. The conditions of the classical anaphora account are fulfilled, and the second occurrence of books can be replaced by the pro-form *ones*. The very same conditions seem to be met in (9b): we have the underlying *small books*, with *book* again being identical to *book* in *bookrack*. However, this time, it is blocked by the Lexical Integrity Hypothesis: the syntactic process of anaphora formation is not allowed to look into the compound *bookrack*. Therefore, no identity can be established and no pro-form can be introduced. Following the logic of this explanation, this pattern or rather the unavailibiltiy of this pattern has been used to establish whether A N constructions in languages with few if any morphosyntactical marking of compounds are compounds or not. A simple

---

[3] Note that this version of the Lexicalist Hypothesis is used by Anderson as an illustration of a maximally strong position.

example from English will clarify the logic behind this test. English has no clear morphosyntactic criteria for A N compounds, because the wordforms of the adjective and the noun are unaffected by phrasal or compound status, and secondary criteria like spelling and stress placement are not decisive, that is, a compound does not have to be spelled as one word, nor does it necessarily need to carry main stress on its lefthand constituent.[4]

Thus, an A N construction like *young dogs* could, in principle, be either a phrase or a compound. However, employing lexical integrity and the classical anaphora account, we can settle this issue. First, we need to find a sentence containing a second A N construction with identical head, e.g. *old dogs* in *I like young dogs, but Sue prefers old dogs*. Secondly, we need to check whether it is possible to delete or replace the second head. In this case, it is possible, cf. *John likes young dogs, but I prefer old ones*. Combining the classical anaphor account and lexical integrity, there is only one possible way in which this deletion/substitution can take place, schematically given in (10).

(10) Schema for head noun deletion/substitution

    a. A N . . . A N   [Two A N constructions]

    b. A N_i . . . A N_i [Head-Head identity can be established]

    c. A N_i . . . A $\cancel{N}_i$ or A pro-form$_i$

       [Head of second A N is deleted or replaced by a pro form]

In order to end up with the configuration in (10c), that is, with a deleted or replaced head of the second A N construction, we need to start with two A N constructions in (10a) where the head to be deleted or replaced is still present. Only then can we

---

[4] Note that these criteria usually work the other way around, though: one-word-spellings usually indicate one-word status and stress on the left-most member of an A B construction usually indicates compound status.

identify two identical lexical segments , as required by the classical anaphora account, in this case the two heads, as indicated in (10b) through the use of the identical subscript. At the same time, the identity of the two heads can only be established if the first head can be seen by the synactic component. According to the maximally strong version of the Lexical Integrity hypothesis, this is only possible if the first A N construction is phrasal. Therefore, the availabilty of the A N_i . . .A $\bcancel{N_i}$/A pro-form$_i$ pattern shows that the first A N construction is phrasal.

Recently, this argument has been employed in Giegerich (2005) and Paul (2005) for A N constructions in English and Mandarin Chinese, respectively. Thus, according to the pattern outlined in (10), Giegerich's data in (11), cf. (4) in Giegerich (2005:579f), can be taken to show that *medical appointment* is a phrasal construction, whereas *mental disorder* is a compound.[5]

(11) a.   Do you have a medical appointment or a dental one?

      b.   *Is this a mental disorder or a nervous one?

Paul also uses the anaphoric island diagnostic in her discussion of Mandarin A N compounds, a representative minimal pair is (12), taken from (51a) in Paul (2008) and (19) in Paul (2005).

---

[5] Giegerich (579ff.) interprets this data pattern in a different way, in that he believes that it is actually the underlying second A N construction that is shown to be phrasal or compound-like, that is, for (11a), the availability of *dental one* shows us that *dental appointment* is phrasal. On this view, replacement by *one* itself is enough to show phrasal status, a test criterion also used by Bauer (1998:76ff.) Technically, this is based not on the prohibition of outbound anaphoras into anaphoric islands, but to the prohibition of what Postal calls inbound anaphora, that is, anaphoric elements that occur inside of words. In order to make his point, though, Giegerich would have to show that the antecedents in (11) both allow anaphoric access to their head. That is, if *mental disorder* is a compound and therefore an anaphoric island, its head *disorder* cannot serve as an antecedent in the first place. Note that Levi (1977:332) uses A *one* constructions similar to those used by Giegerich for arguing that complex nominals do in fact allow inbound anaphora.

(12) a. *Wǒ xǐhuān lǜ-chá,   hóng   de   yě   kěyǐ

     I   like   green-tea   red   SUB   also   possible

     i.e.'I like green tea, but black [lit. red] tea is also OK.'

   b. Amei bù   xǐhuān huáng méigui, hóng de   hái   kěyi

      Amei NEG like   yellow rose   red   SUB   still   acceptable

      i.e. 'Amei doesn't like yellow roses, red ones are still OK.'

According to the syntactic reading of the contrast in (12), which is based on the parallel A N constructions *lǜ-chá* 'green tea' and *hóng -chá* 'black [lit. red] tea', the ungrammaticality of (12a) shows that *lǜ-chá* 'green tea' is a compound, whereas the grammaticality of (12b) shows *huáng méigui* 'yellow rose' (paired with *hóng méigui* 'red rose') to be a phrasal A N construction.

The problem with the argument is that there are counterexamples for almost every prediction of the syntactic account, which are briefly summarized below, where we will start with some general remarks, and then focus on counterevidence involving compounds in English.

In general, the counterevidence shows two important things. Firstly, it shows that the syntactic explanation as stated by Postal is insufficient. Secondly, it shows that any explanation for the data must be able to explain different degrees of acceptability. This cline in acceptability was observed in one of the first reactions to Postal's original paper, a paper by Lakoff and Ross (1972) giving the data and judgements reproduced in (13), their (2b) and (3a-b), where *one* in (13a) and *it* in (13b) and (13c) are intended to refer to the guitar.

(13) a. *A guitarist bought one yesterday

   b. ?*The guitarist thought that it was a beautiful instrument.

c. ?John became a guitarist because he thought that it was a beautiful instrument.

Clearly, this cline cannot be explained by Postal's original proposal, which makes a categoric difference between islands and non-islands. Other authors offering counterexamples to Postal's strong claim include Tic Douloureux (1971), Corum (1973), Browne (1974) and Watt (1975), whose main claims and accounts are discussed in Ward et al. (1991), as well as Levi (1977). A representative set of counterexamples involving English compounds is presented below, first with anaphoric references to the first element of the compound, cf. (14), secondly with anaphoric reference to the second part of the compound, cf. (15).

(14) a. Although casual cocaine$_i$ use is down, the number of people using it$_i$ routinely has increased.

b. Patty is a definite Kal Kan$_i$ cat. Every day she waits for it$_i$.

c. I was an IRS$_i$-agent for about 24 years. ..I stopped working for them$_i$.

The examples in (14) are from the appendix to Ward et al. (1991). *Cocaine use* in (14-a) is a synthetic compound which should disallow anaphoric reference to its two constituents. However, *it* refers back to *cocaine* and not to *cocaine use*. Similarly, *it* in (14-b) refers back to the denotation of *Kal Kan*, that is, to a specific brand of catfood, where *Kal Kan* is embedded in a standard endocentric compound, and finally, in (14-c), *it* back to the IRS.[6]

The data discussed by Ward et al. (1991) is restricted to anaphoric reference to the first element of the compound, whereas Levi (1977) presents data where we see

---

[6] Note that it does not seem to be accidental that in two of the three examples in (14) the prononimal element refers back to the referent of a proper noun. Ten Hacken (1994:76) points out that this occurs in two thirds of the cases involving pronominal reference to non-heads in the corpus investigated by Ward et al.

that reference to the second element is also possible, cf. (15), her (17b), (18b), and (19a).

(15) a. State taxes$_i$ were higher than municipal ones$_i$.

b. Steam irons$_i$ need more maintenance than those$_i$ that iron dry.

c. Student power$_i$ is insignificant compared to that$_i$ of the Dean.

In (15-a), *ones* refers back to the denotation of *taxes* and not to that of *state taxes*, *those* in (15-b) refers to that of *irons* and not to the denotation of *steam irons*, and finally, *that* in (15-c) refers back to the denotation of *power* and not to the denotation of *student power*.

The German data presented in section 1 strengthens Levi's observation insofar as the first A N constructions are clear compounds according to morphosyntactic criteria, whereas for English similar clear-cut criteria are not available. In addition, the difference between the data in (3)-(4) and (6) shows also a cline in acceptability.

Thus, while the patterns observed by Paul and Giegerich still need an explanation, it seems clear that anaphoric islandhood of the antecedents is not the correct one. The main alternative to the original proposal is the pragmatic account outlined below.

2.2 A pragmatic interpretation

Ward et al. (1991) offer a pragmatic account for the anaphoric island data, summed up in the following quote: ". . . the degree to which outbound anaphora is felicitous is determined by the relative accessibility of the discourse entities evoked by word-internal lexical elements, and not by any principle of syntax or morphology." (1991:449). The question that emerges for this account is then, naturally, what determines the relative accessibility of the discourse entities involved. Ward et al.

(1991) distinguish two sets of factors: (a) morphosyntacic/semantic factors and (b) pragmatic factors, that is, contrastiveness and topicality. They discuss semantic transparency as the key morphosyntactic and semantic factor. Semantic transparency of a compound is understood by Ward et al. to involve accessing the meanings of both parts of the compound. Thus they claim that interpreting *cocaine use* and *Kal Kan cat* in the examples (14-a) and (14-b) above both require the hearer to access the meanings of the two constituents of the compounds. Due to this access to the meaning of the individual constituent, the corresponding discourse entities are evoked and, in turn, available for anaphoric reference. In contrast, institutionalized, idiosyncratic compounds can acquire opaque meanings, that is, meanings that cannot be "straightforwardly" (1991:454) interpreted on the basis of the meanings of the constituents of the compound. Once a compound has acquired an idiosyncratic meaning, "a hearer may access the meaning of the compound directly, i.e. without morphologically decomposing it." (1991:454). In consequence, the potential discourse entities are not evoked. This explains the pragmatic deviance of (16), their (23a).

(16)   Fritz is a [cow]$_i$boy.

       # He says they$_i$ can be difficult to look after.

While this explanation itself rests on a categorical contrast (either the meanings of the individual constituents are accessed or not), Ward et al. are careful to point out that opacity is a gradient phenomenon: "…, the distinction between transparent words and opaque or institutionalized words is gradient rather than categorical. We would therefore expect word-internal morphemes to evoke discourse entities with a greater or lesser degree of accessibility depending, inter alia, upon the relative transparency of the containing word." (1991:455). However, a relative measure of semantic transparency is not introduced by the authors.

Interestingly, Ward et al. argue that "While semantically transparent compounds do allow felicitous outbound anaphora, it is also true that anaphora involving antecedents within compounds is, other things being equal, more difficult to construe than anaphora involving non-word-internal antecedents." (1991:455). To account for this, they speculate that it may have to do with the fact that in modifier-head constructions, which the compounds in their data essentially are, the modifiers are more backgrounded than the heads, a claim they support by referring to psycholinguistic evidence from McKoon et al. (1990).[7] Note that this explanation only targets anaphoric reference to the non-heads of compounds, whereas the German data introduced in section 1 (examples (3), (4) and (6)) in all cases involves anaphoric reference to the head of the compound.

As for pragmatic factors, Ward et al. (1991:456) discuss contrast and topicality. Discourse entities seem to be more accessible if in salient opposition to some other discourse entity (see also Watt, 1975): "topical discourse entities evoked by word-internal elements facilitate outbound anaphora …" (Ward et al. 1991:456). The role of topicality in the facilitation of outbound anaphora is also discussed by Bosch (1983:238, note 116), who notes that this factor might be responsible for the difference in acceptability in the following example sentences from the original Postal paper:[8]

(17) a. Prejudice against Jews makes rich ones donate money.

b. *Anti-semitism makes rich ones donate money.

(18) a. Max hunts for wild animals but Pete only kills domesticated ones.

b. *Max is a wild-animal hunter but Pete only kills domesticated ones.

---

[7] This manuscript is not available to me.
[8] Cf. for (17) Postal's (59a-b), for (18) his (54a-b). (19) corresponds to (27a-b) in the appendix to Postal's paper.

(19) a. Harry solicits for prostitutes and Pete arrests them.

b. *Harry is a pimp and Pete arrests them.

All three b-sentences contain different examples for anaphoric islands, e.g. a neo-classical derivation in (17b), an [[AB]C] compound in (18b), and a monomorphemic word in (19b), yet for all three 'islands', the same, simple explanation can account for the missing availability of anaphoric reference: "The (a) sentences can be about Jews, wild animals, or prostitutes respectively, and accordingly anaphoric relations linking up to the corresponding referents are possible. In the (b) sentences there is no possibility for corresponding aboutness relations and hence no anaphora either" (Bosch 1983:238). In the German data, topicality and contrastiveness are inherent features of the pattern under investigation.

2.3 Psycholinguistic evidence

The pragmatic account is partially supported by findings from psycholinguists. On the one hand, Ward et al. (1991) report the results of a number of psycholinguistic experiments[9] that show the importance of the two factors of transparency and topicality. On the other hand, their claims can be fruitfully linked to other psycholinguistic research on compounds.

Ward et al. (1991) report experiments where the accessibility of discourse units serving as antecedents for anaphora was manipulated by (a) varying morphosyntactic structure (the antecedent either occurred in a nominal compound or in a verb phrase) and (b) varying topicality and contrast. The two most important results of the experiments are (a) when the antecedent is topical, there is no significant

---

[9] In reporting these experiments, Ward et al. always refer to the aforementioned manuscript McKoon et al. 1990. McKoon et al. (1993) seems to be the publication that this manuscript resulted in, as it contains discussion of all the experiments mentioned below.

difference in reading times regardless of the antecedent occurring in a compound or not and (b) when the antecedent is non-topical, the reading times for the compound internal versions and the VP version differ significantly. In McKoon et al. (1993), these findings are accounted for by assuming that we use a discourse model, i.e. a continually updated representation of information built up in comprehension of texts. The model "is made up of the entities evoked by linguistic and contextual information, the relations among the entities, and their accessibilities relative to potential referential cues." (McKoon et al. (1993:72)). Importantly, they propose that a number of factors, such as topicality, morphosyntactic but also extralinguistic context, influence the accessibility of a given discourse entity, and that the referential cue, i.e. the anaphoric element itself, also determines relative accessibility. Therefore, what might be, at a specific position in the text, accessible through use of a personal pronoun might not be accessible by other pro-forms in the same position.[10]

Another aspect of compounds, namely their internal semantics, has also been subject of psycholinguistic investigations. Zwitserlood's (1994) study on Dutch compounds addresses exactly this point. Zwitserlood classifies the compounds used in her study into three different groups: (a) fully transparent compounds, (b) truly opaque compounds, and (c) partially opaque compounds. In the case of fully transparent compounds, the meaning is synchronically related to the meaning of the individual constituents, as in e.g. *milkman*. The truly opaque compounds, in contrast, bear no semantic relation with any of the constituents. Zwitserlood uses the English

---

[10] A small example from native speaker judgements on English A N compounds illustrates this sensitivity to the anaphoric element itself: While the native speakers I asked all judged (i) as out, most thought (ii) was OK.

(i)   Peter uses a blackboard, but I prefer the white ones.
(ii) Blackboards require more care than those with a white surface.

However, the target of the anaphoric references is in both cases identical.

blackguard 'one of the idle criminal class' to illustrate this; a Dutch example she uses is *klokhuis*, lit. 'clock house', but meaning 'core of an apple'. Finally, partially opaque compounds are linked to the original meaning of one of the constituents, e.g. *jailbird* refers to a person that is often in jail. The most important finding of Zwitserlood's experiments is that all compounds, including the truly opaque compounds, are represented as morphologically complex at some level. Transparent compounds and partially opaque compounds[11] facilitate semantic access to the meanings of their constituents. This is interpreted by Zwitserlood as evidence that transparent and partially opaque compounds are linked (a) to their own semantic representation and (b) to the semantic representation of their constituent words. Truly opaque compounds, in contrast, behave semantically like monomorphemic words in that they are only linked to their own semantic representation. Libben et al. (2003), in a study on English compounds, also found that all compounds show morphological constituency. In addition, they found that the transparency of the head played a decisive role for the time it takes to make a lexical decision, i.e., the time it takes to determine whether or not a string of letters is a word of one's language.

3 THE GERMAN DATA

Ward et al.'s (1991) data contains a large number of compounds, but anaphoric reference is never made to the right-most constituent. Instead, the anaphoric reference is made to the first constituent, as in the examples in (14). In addition, in Ward et al.'s examples, the anaphoric element is always (a) a pro-form on the surface and (b) a form serving as noun phrase (or prepositional phrase) by itself. As a consequence, the anaphoric elements are mostly personal pronouns or *there*, and only in a few cases is

---

[11] In Zwitserlood's study, only partially opaque comounds with a transparent first constituent and an opaque second constituent were used.

one used as an anaphoric pro-form. The anaphoric element thus typically refers to a concrete discourse entity introduced by the antecedent, whereas the German pattern as introduced in examples (3) and (4) always involves the deletion of the head noun. The anaphoric element is a pro-form that is not visible on the surface. This pro-form does not serve as a noun phrase, but just replaces the head noun. In addition, the whole test pattern already rests on contrastive topicality, which allows focusing on other factors that lead to anaphoric reference in these examples.

Below, I will first introduce a standard classification of German A N compounds and discuss their behaviour with regard to anaphoric reference. In a second step, I will discuss the results of an explorative corpus search for the pattern under investigation.

3.1 Subclasses of German A N compounds and anaphoric islandhood

We get a first feel for the interrelationship between semantic transparency of German A N compounds and the availability of our deletion pattern if we go through examples of the standard subclasses of A N compounds. An endocentric compound is understood here as allowing the paraphrase '$[AB]_N$ is a B_N', that is, the compound is a hyponym of its head. The classes described here are based on the classification schemes for A N compounds in Fahim (1977), who employs simple paraphrase tests for the distinction between all in all five classes. Three of the five classes are subcategories of endocentric compounds, here referred to as endocentric class A, endocentric class B, and endocentric class C. The other two classes cover the subcategories of exocentric compounds, here referred to as exocentric class A and exocentric class B.

Endocentric class A compounds are compounds that correspond in meaning to the respective phrasal versions, cf. (20).

(20) Endocentric class A

$[AN]_N = [AN]_{NP}$

Rotwein = roter Wein 'red wine'

This class, which includes the compound pair *Grünglas* 'green glass' and *Weißglas* 'white glass' from example (3), allows for anaphoric reference, cf. (21).

(21) Ich hab keinen Rotwein gekriegt, es gab nur noch weißen.

'I have no red.wine received, there was only still white.'

i.e. I didn't get any red wine, they only had white wine left.

Other examples for compounds of this class are *Grüntee* 'green tea' and *Schwarztee* 'black tea'.

The second subgroup of endocentric compounds, endocentric class B, still allows phrasal paraphrases, but the meaning of compound and corresponding phrase is not completely equivalent, cf. (22).

(22) Endocentric class B

$[AB]_N \approx [AB]_{NP}$

Großstadt ≈ große Stadt 'big city'

The meaning of the phrase *große Stadt* 'big city' is less specific than the meaning of the compound and can be used to refer to any large-sized city. In contrast, the compound *Großstadt* has two different idiosyncratic meanings. On the one hand, it is the technical term in Germany for any city with at least 100,000 inhabitants. On the other hand, it can be used for cities in the sense of 'metropolis', that is, a large, bustling city.

Despite the specialized meaning in this class, the corresponding patterns are acceptable, cf. (4), repeated here as (23), where *Großstadt* contrasts with *Kleinstadt*, which exhibits the same two kinds of specialized meaning, i.e. it either refers to a town with between 5,000 and 20,000 inhabitants or to a town with a provincial feel to it.[12]

(23) Ich liebe Großstädte, in kleinen gehe ich ein.

'I love big-towns, in small go I in' [cf. *ein-gehen* 'to perish']

i.e. I love big cities, I cannot exist in small cities.

The possibility of pragmatically unmarked anaphoric reference might be connected to the fact that for endocentric A as well as for endocentric B compounds the contribution of the adjective fully overlaps with its predicative usage. For class A compounds, although *Rotwein* is a specific kind of wine, it is also wine that is red, a feature perhaps even more obvious for the three common colour adjective + *glas* compounds, *Weißglas* 'white glass', *Grünglas* 'green glass' and *Braunglas* 'brown glass'. These three compounds refer to the three kinds of colored glass bottles relevant for the recycling schemes in Germany, and one encounters them when standing in front of the correspondingly labelled containers in order to get rid of one's old bottles, where one typically checks the colour of one's bottles before disposing of them in the corresponding containers. For endocentric class B compounds, the pattern still holds, i.e. a *Kleinstadt* is small, although the specific meanings mentioned above are lost.

Endocentric class C compounds, finally, are those that are so opaque that the phrasal version does not correspond to the compound any more, cf. (24).

---

[12] Not surprisingly, there is also a technical term for the towns between 20,000 and 100,000 inhabitants: *Mittelstadt* 'middle city'. However, this term is not used in everyday language.

(24) Endocentric class C

$[AB]_N \neq [AB]_{NP}$

Grünspecht ≠ grüner Specht 'Green woodpecker'

That is, the *green woodpecker* is a member of the woodpecker family *Picidae*. Green woodpeckers all have green upperparts, paler yellowish underparts, and a red crown, but the properties expressed by the adjective do not hold for the whole entity referred to by the head noun and the corresponding predications are false (i.e. a green woodpecker is NOT green). In contrast, the intuitive interpretation of the phrasal *grüner Specht* is intersective, its referent is a woodpecker, and it is green. Anaphoric reference to the head is pragmatically heavily marked, cf. (6), repeated here as (25).

(25)   #Mein Vater hat in seinem Garten schon mal einen Grünspecht gesehen, aber noch nie einen schwarzen.

My father has in his garden already once a green-woodpecker seen, but so_far never a black.

Intended: 'My father once saw a green woodpecker in his garden, but he has never seen a black woodpecker.'

On its preferred reading, (25) is interpreted with anaphoric reference to the whole compound, that is, to a black green woodpecker.

For the two groups of exocentric compounds, Fahim (1977) distinguishes between those where the first part is true of the referent of the whole compound and those where this is not the case. Thus, exocentric class A conforms to the pattern in (26).

(26) Exocentric class A

$[AB]_N(x) \rightarrow [A]_N(x)$

Ein Dummkopf ist dumm. 'A stupidhead is stupid'

As in (25), anaphoric reference into exocentric A class compounds is pragmatically marked, intuitively bordering on the ungrammatical, cf. (27), based on the pair *Dummkopf* 'stupid head'– *Schlaukopf* 'smart head'.[13]

(27) #In dieser Klasse gibt es reichlich Dummköpfe, aber zum Glück auch schlaue.

In this class has it abundantly stupidheads, but luckily also smart.

Intended: 'This class has lots of boneheads, but, luckily, also some smart students.'

Exocentric class B compounds, finally, are those compounds where neither the first part nor the second part is a predicate of the referent of the whole, cf. (28).

(28) Exocentric class B

$[AB]_N(x) \text{ NOT} \rightarrow [A]_N(x)$

Ein Rotkehlchen ist nicht rot. 'A robin [red.throat] is not red.'

Thus, the referent of the noun phrase *ein Rotkehlchen* 'a robin' in (28) is a bird with a red throat. It is not a red entity nor is it a throat. It is thus a prototypical instance of a possessive compound: the head, *Kehlchen* 'throat', is metonymically reinterpreted and stands for the whole bird, and the property expressed by the adjective is a property of the entity normally referred to by the reinterpreted head. As was the case for exocentric class A compounds, anaphoric reference is heavily marked, cf. (29),

---

[13] One additional problem with this pair is that *Schlaukopf* 'smart.head' is often interpreted with the somewhat negative tinge also present in e.g. English *smart alec*.

with the pair *Braunkehlchen* [lit. brown throat] 'whinchat'– *Rotkehlchen* [lit. red throat] 'robin' .

(29) #Guck mal, ein Braunkehlchen und ein rotes!

Look, a brown.throat and a red

Intended: 'Look, a whinchat and a robin!'

The paraphrase tests used to establish the five different classes are one way to operationalize the notion of semantic transparency. On this view, endocentric class A compounds are the most semantically transparent subclass of A N compounds, while exocentric class B compounds represent the least transparent subclass. The intuitive judgements of the possibility of anaphoric reference into these different classes of compounds support the idea that semantic transparency plays a key role. In the next section, we attempt to give empirical backing to these intuitions by looking at corpus data.

3.2 The Data in the corpus

The data so far was based solely on intuitive judgements. In order to give some empirical support to these intuitive judgements, I report here the result of an exploratory corpus study. I used the COSMAS II$_{web}$ corpus tool[14] developed and maintained by the Institut für Deutsche Sprache (Mannheim), accessing all written corpora of the Deutsche Referenz Korpus (DeReKo)[15] available through that system.

---

[14] COSMAS II (*Corpus Search, Management and Analysis System*), http://www.ids-mannheim.de/cosmas2/, © 1991-2010 Institut für Deutsche Sprache, Mannheim.

[15]Das   Deutsche   Referenzkorpus   DeReKo,   http://www.ids-mannheim.de/kl/projekte/korpora/,
am Institut für Deutsche Sprache, Mannheim.

The main idea behind this exploratory study is quite simple: if there is anything to the intuitions reported in the previous sections, then we should find a reflection in the corpus. First of all, we should find the pattern of interest as such, i.e. occurrences of an A N compound with a following inflected adjective and a deleted nominal head. Secondly, there should be a higher relative number of tokens exhibiting this pattern for the endocentric class A compounds, and a lower relative number for those compound classes were the pattern was judged to be pragmatically deviant. Trying to find this kind of data in the corpus proved to be difficult. First, the corpus is not tagged for the internal structure of compounds, so that no search for A N compounds as such is possible. Secondly, the second element of the pattern, the occurrence of an inflected adjective followed by a zero element provides no help, since, unsurprisingly, the zero element is not encoded, and inflected adjectives occur in large numbers. The only reasonable strategy would be to search standard lists of A N compounds for same-headed pairs with contrasting first adjectival element, and then using these pairs as a basis to search for the patterns individually. While this involves a rather large scale study, I present below the results of an exploratory corpus search for all the patterns discussed in the examples so far. The search pattern used was always similar, I illustrate it here for the pair *Großstadt/Kleinstadt* 'big city/small city'. In order to capture both possible linearizations, two searches were carried out, one with the search pattern "Großst?dt+ /+s0 kleine*" and one with the search pattern "Kleinst?dt+ /+s0 große*".[16]

---

[16] The search syntax and the special characters in these two search patterns have the following functions: /+s0 requires the two other constraints to be matched in the same sentence, the question mark stands for any single character, allowing simultaneous search for the singular wordforms and the plural forms, which require an umlaut. The + at the end of the compounds stands for zero or one more character, capturing thus all inflectional endings except the dative plural, the star stands for any number of further characters.

The results can easily be summarized: there were no hits for the pattern involving any of *Weißglas/Grünglas* 'white glass/green glass', *Grüntee/Schwarztee* 'green tea/black tea', *Schwarzspecht/Grünspecht* 'black woodpecker/green woodpecker', *Dummkopf/Schlaukopf* 'stupidhead/smarthead' and *Rotkehlchen/Braunkehlchen* 'robin/whinchat' as the first element. For the four remaining compounds, I obtained the results presented in table 1.

Pattern	*Number of Occurrences*
Rotwein(e) ...weiße(n/r) [Wein]	25
Weißwein(e) ...rote(n/r) [Wein]	55
Großstadt/städte ... kleine(n/re) [Stadt/Städte]	1
Kleinstadt/städte ...große(n)/größere [Stadt/Städte]	2

Table 1: Absolute occurrences of the anaphoric reference search pattern

These results can partly be explained by the absolute frequencies of the corresponding compounds in the corpus (2.3 billion words), cf. table 2.

Rotwein 'red wine'	12,465	Weißwein 'white wine'	5,209
Weißglas 'white glass'	86	Grünglas 'green glass'	65
Grüntee 'green tea'	336	Schwarztee 'black tea'	328
Großstadt 'big city'	17,638	Kleinstadt 'small city'	12,276
Schwarzspecht 'black woodpecker'	304	Grünspecht 'green woodpecker'	319
Dummkopf 'stupidhead'	948	Schlaukopf 'smarthead'	76
Rotkehlchen 'robin'	873	Braunkehlchen 'whinchat'	281

Table 2: Absolute frequencies of compounds in the corpus

The two members of the pair *Weißglas/Grünglas* occur 86 and 65 times, compared to *Rotwein/Weißwein* with 12,465 and 5,209, respectively. In fact, the absolute frequency for the compounds where the pattern does not occur are all relatively low,

except for the pair *Dummkopf/Schlaukopf*, which shows a high asymmetry with 948 to 76 occurrences. The absolute frequency of the *Großstadt/Kleinstadt* 'big city/small city' pair, 17,638 and 12,276, respectively, is the highest of the sample and clearly higher than that of white/red wine.

Taking a closer look at the data, a few other aspects are noteworthy. First, the two instances of the "Kleinstadt/städte ...große(n)/größere" pattern both involve the comparative form of the second adjective, cf. the example below.

(30) Erst mit der Erschließung des Umlandes durch Eisen- und Straßenbahn konnten Kleinstädte zu größeren wachsen, so auch Salzburg. N97/AUG.33561

'Only with the development of.the surroundings through trains and trams could small.cities to bigger grow, as also Salzburg.'

i.e. Not until the urban hinterland had been developed through trains and trams could small cities grow into bigger ones, as did Salzburg.

Here, the corresponding compound could not have been used (the phrase could, though). Similarly, the single example for the "Großstadt/städte ... kleine(n/re)" pattern contains multiple modification, again not realizable through a compound, cf. (31).

(31) Nein, nicht in einer Großstadt, in einer kleinen, überschaubaren, ohne Hektik und Trubel. O95/JUL.69073

'No, not in a big city, in a small, manageable [one], without hustle and bustle. '

Thus, although in all three cases the compounds in question are topical and the anaphoric usage is contrastive, that is, the general conditions for anaphoric reference are fulfilled, we have, in addition, clear grammatical constraints that make the usage of the corresponding compound impossible. This differs radically from the picture

that presents itself when looking through the sentences involving *wine* as the head of a compound. In all but one instance, the compound could be used instead of the adjective + pro-form. A typical example is given in (32).

(32) Die viele Sonne ist gut für den Rotwein, schlecht für den Weißen.

     NON07/JUN.16950

     'The large amount of sunshine is good for the red wine, bad for the white'

And while the choice of the construction with a deleted head in these sentences can thus be seen as an instance of linguistic economy, it is the only available encoding in the sentences involving *Stadt* 'city' as the head of a compound, where the usage of the corresponding phrasal variant would lose the specific meaning characteristics of the compound. If it is true that we can thus distinguish between two different motivations for the usage of the pro-form anaphora, on the one hand grammatical necessity in the case of the *city* examples, on the other hand simple linguistic economy in the case of the *wine* examples, the difference between the sentences with compounds headed by *Wein* 'wine' and compounds headed by *Tee* 'tea' comes as a surprise. Looking at the number of absolute occurrences given above, one is led to believe that it is simply an effect of the low overall number of the two types of *tea* compounds. However, a closer look at the data reveals a fundamental difference between the two cases. *Tea* compounds co-occur frequently with full *tea* phrases, as in (33).

(33) Wie Schwarztee ist Grüner Tee koffeinhaltig, ... HMP08/NOV.01318

     'Like black tea is green tea caffeine-containing, ...',

     i.e. Like black tea, green tea contains caffeine ...

Out of the 17 instances that fit the general search pattern used for the detection of anaphoric reference, 16 instances show this pattern. In contrast, this pattern never

occurs for the *wine* examples, where we have 181 instances matching the search pattern, of which 79 show the anaphoric reference phenomena. Why might this be? The key seems to lie in the relation between the phrasal and the compound versions. As was mentioned in section 3.2, the defining characteristic of endocentric class A compounds is the full equivalence of the phrasal and the compound version. This is true for all four compounds. However, the absolute number of occurrences of the compound variants and the phrasal variants differ asymmetrically, cf. table 3.

Rotwein	12,465	roter Wein	668
Weißwein	5,209	weißer Wein	106
Grüntee	336	grüner Tee	1,082
Schwarztee	328	schwarzer Tee	549

Table 3: Asymmetry in the absolute frequency of phrasal and compound A N

constructions

Thus, *Rotwein* is 19 times more frequent than *roter Wein* and *Weißwein* is 49 times more frequent than its phrasal counterpart *weißer Wein*, contrasting with the relation between *grüner Tee* and *Grüntee*, where the former is 3.2 times more frequent than the latter, and between *schwarzer Tee* and *Schwarztee*, where the former is 1.7 times as frequent as the latter.

The picture that emerges is that the relative collocational strength of the phrasal *tea* constructions is far greater than that of the *wine* constructions. In other words, for the A N constructions with *tea*, the phrasal variant is more entrenched than the compound variant, whereas for the *wine* A N constructions, it is the other way around.[17] The effect of this difference in collocational strength is that once we decide

---

[17] For the notion of entrenchment, cf. Langacker (2000:3).

to use the phrasal A N construction to refer to black or green tea, we are forced to add the head, even if this head is easily anaphorically recoverable, cf. (33).

4 CONCLUSION AND OPEN ENDS

The aim of this paper was to discuss the relationship between semantic transparency and anaphoric islands. In particular, it discussed the two competing analyses of the phenomena, reproduced the most compelling evidence in favour of the pragmatic account, and introduced German data through which a deeper understanding of the processes behind anaphora into compounds can be gained. Thus, already for English we have data that shows that the first, modifying part, of compounds as well as the second part, the head, can serve as antecedents for anaphora. This can easily be explained by the pragmatic account, whose insistence on the importance of semantic transparency and contrastive topicality is backed up by findings from psycholinguistics. The German data shows that while standard classifications of A N compounds follow the general tendencies predicted by the pragmatic account, we can identify other relevant factors by looking at corpus data. Thus, at least for compounds that are fully transparent, i.e. of the Endocentric A class, the entrenchment of the phrasal variant relative to the compound variant plays a decisive role in opting for a construction involving anaphoric reference. And for compounds that are not fully transparent, starting with Endocentric class B, we apparently need morphosyntactic constraints to force us into the usage of an anaphoric construction.

However, since we did not find any data in the corpus involving the other compound classes, there is still much work left to do. According to the pragmatic account, anaphoric reference should be possible in all cases, something which we do not find reflected in the data. This might ultimately be due to the relatively low

frequency of the base constructions, and maybe only psycholinguistic tests will allow further insights into the detailed workings behind anaphoric reference into A N compounds.

Anderson, Stephen R. (1992). *A-Morphous Morphology.* Cambridge: Cambridge University Press.

Bosch, Peter (1983). Agreement and Anaphora - A Study of the Roles of Pronouns in Discourse and Syntax. (Cognitive Science Series) London & New York: Academic Press.

Browne, Wayles (1974). On the Topology of Anaphoric Peninsulas. *Linguistic Inquiry*, 5(4), 619-620.

Corum, Claudia (1973) Anaphoric peninsulas. *Chicago Linguistic Society* 9, 89-97

Fahim Elsayed, Mohammed Salah (1977) Untersuchungen zum Modell substantivischer Komposita mit einem Primäradjektiv als erster unmittelbarer Konstituente. Dissertation, Leipzig.

Giegerich, Heinz J. (2005). Associative adjectives and the lexicon-syntax interface. Journal of Linguistics 41, 571-591

ten Hacken, Pius (1994), *Defining Morphology: A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection*, Hildesheim: Olms.

Hankamer, Jorge and Ivan Sag (1976). Deep and surface anaphora. *Linguistic Inquiry* 7, 391–428.

Harris, Alice C. (2006) *Revisiting Anaphoric Islands* Language - Volume 82, 1, 114-130

Langacker, Ronald W. (2000). A dynamic Usage-Based Model. In Barlow, Michael and Suzanne Kemmer, (eds.) Usage based models of language. Stanford: CSLI

Lakoff, George, and John Ross. 1972. A note on anaphoric islands and causatives. *Linguistic Inquiry* 3.121-25.

Levi, Judith N. (1977). The constituent structure of complex nominals, or, That's funny, you don't look like a noun! In Woodford A. Veach, Samuel E. Fox, and Shulamith Philosoph (eds.) Papers from the thirteenth regional meetin: Chicago Linguistic Society. Chicago: Chicago Linguistic Society. 325-338.

Libben, Gary and Martha Gibson, Yeo Bom Yoon, and Dominiek Sandra (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, 84, 50-64.

McKoon, Fail, and Gregory Ward, Roger Ratcliff, and Richard Sproat (1990). Morphosyntactic and pragmatic manipulations of salience in the tinterpretation of anaphora. Evanston, IL, and Murray Hill, NJ: Northwestern University and AT&T Bell Laboratories, MS.

McKoon, G., Ward, G., Ratcliff, R., & Sproat, R. (1993). Morphosyntactic and pragmatic factors affecting the accessibility of discourse entities. *Journal of Memory and Language, 32*, 56-75.

Paul, Waltraud (2005). Adjectival modification in Mandarin Chinese and related issues. *Linguistics* 43(4): 757–793.

Paul, Waltraud (2008). Adjectives in Mandarin Chinese: The rehabilitation of a much ostracized category. Manuscript.

Pesetzky, David (1979). Russian morphology and lexical theory. Cambridge, MA: MIT, MS. „http://web.mit.edu/linguistics/people/faculty/pesetsky/russmorph.pdf"

Postal, Paul M. (1969). Anaphoric islands. In Binnick et. al. (eds.), *Papers from the 5th regional meeting of the Chicago Linguistic Society*, Chicago: Chicago Linguistic Society, University of Chicago. 209–39.

Ross, John R. (1967). Constraints on variables in syntax. Doctoral dissertation, Massachusetts Institute of Technology.

Schäfer, M. (2009). A N constructions in Mandarin and the compound versus phrase debate. *Word Structure* 2 (2), 272-293.

Selkirk, Elisabeth O. (1982). *The syntax of words*. Cambridge: MIT Press.

Tic Douloureux, P.R. N. (1971). A note on one's privates. In Arnold Zwicky et al. (eds.) *Studies out in left field: Defamatory essays presented to James D. McCawley on the occasion of his 33rd or 34th birthday.* 45-52. Edmonton and Champaign: Linguistic Research, Inc. 45-52

Ward, Gregory, Sproat, Richard & McKoon, Gail 1991. A pragmatic analysis of so-called anaphoric islands. *Language* 67(3): 439–474.

Watt, W. (1975). The indiscreteness with which impenetrable are penetrated. *Lingua* 37, 95-128.

Zwitserlood, P. (1994). The role of semantic transparency in the processing and representation of Dutch compounds. Language and Cognitive Processes, 81, 358-367.

Corpora and tools

Das Deutsche Referenzkorpus DeReKo, http://www.ids-mannheim.de/kl/projekte/korpora/, am **Institut für Deutsche Sprache, Mannheim**

COSMAS II (*Corpus Search, Management and Analysis System*), http://www.ids-mannheim.de/cosmas2/, © 1991-2010 Institut für Deutsche Sprache, Mannheim