Chapter 3

Compositionality in English deverbal compounds: The role of the head

Gianina Iordăchioaia
University of Stuttgart

Lonneke van der Plas
University of Malta

Glorianna Jagfeld
Lancaster University

This paper is concerned with the compositionality of deverbal compounds such as *budget assessment* in English. We present an interdisciplinary study on how the morphosyntactic properties of the deverbal noun head (e.g., *assessment*) can predict the interpretation of the compound, as mediated by the syntactic-semantic relationship between the non-head (e.g., *budget*) and the head. We start with Grimshaw's (1990) observation that deverbal nouns are ambiguous between compositionally interpreted argument structure nominals, which inherit verbal structure and realize arguments (e.g., *the* assessment *of the budget by the government*), and more lexicalized result nominals, which preserve no verbal properties or arguments (e.g., *The* assessment *is on the table.*). Our hypothesis is that deverbal compounds with argument structure nominal heads are fully compositional and, in our system, more easily predictable than those headed by result nominals, since their compositional make-up triggers an (unambiguous) object interpretation of the non-heads. Linguistic evidence gathered from corpora and human annotations, and evaluated with machine learning techniques supports this hypothesis. At the same time, it raises interesting discussion points on how different properties of the head contribute to the interpretation of the deverbal compound.

Gianina Iordăchioaia, Lonneke van der Plas & Glorianna Jagfeld. 2020. Compositionality in English deverbal compounds: The role of the head. In Sabine Schulte im Walde & Eva Smolka (eds.), *The role of constituents in multiword expressions: An interdisciplinary, cross-lingual perspective*, 61–106. Berlin: Language Science Press. DOI:10.5281/zenodo.3598558

1 Introduction

This paper contributes a study on how constituents influence the compositionality of multiword expressions from the perspective of deverbal compounds in English with a focus on the role of their head nouns.

1.1 Deverbal compounds (DCs)

DCs are noun-noun compounds with a deverbal head as illustrated in (1). In contrast to root compounds (RCs) (see 2), whose head nouns are typically simple (non-derived), DCs usually receive an interpretation in which the non-head establishes a syntactic-semantic relationship with the verb from which the deverbal noun is derived (i.e., as a direct object, subject or other argument/adjunct). RCs often receive a fixed interpretation (see 2a) or one depending on the immediate context (see 2b). *Tomato bag* in (2b) may refer to a bag of tomatoes, a bag having the shape or color of a tomato, or any other connection between a bag and tomatoes mentioned in previous context. The same holds for *jelly bottle*.

(1) a. *budget* assess**ment** – to assess (a) *budget*(s) (Object)
b. *police* question**ing** – *police* questions sb. (Subject)
c. *college* educat**ion** – to educate sb. in *college* (Adjunct)

(2) a. train station, bookstore
b. tomato bag, jelly bottle

Nominal DCs may be headed by deverbal nouns built with a variety of suffixes, including those that form participant-denoting nominals, as in (3a) for agents and in (3b) for patients (see Lieber 2016: 73). For reasons that will be given in Section 3.2, we concentrate here on DCs headed by eventive deverbal nominals as in (1), formed by means of the suffixes *-al*, *-ance*, *-(at)ion*, *-ing*, and *-ment*.

(3) a. dog train**er**, flight attend**ant**
b. bank employ**ee**, award nomin**ee**

1.2 Argument structure nominals and result nominals

Grimshaw (1990) points out that the majority of deverbal nouns exhibit an ambiguity between an argument structure nominal (ASN; her *complex event nominal*) reading, which perfectly mirrors the corresponding verb phrase with its argument structure, and a result nominal (RN) reading, which is more lexicalized and

departs from the base verb at various degrees.[1] The crucial difference between the two originates in the availability of verbal event structure, which enforces and constrains argument realization in ASNs (see (6) below), and its absence in RNs. The examples in (4) illustrate the two readings, building on Grimshaw (1990: 49).

(4) a. The **examination**/exam was [on the table/in the bag]. (RN)
b. The **examination**/*exam *of the patients* took a long time. (ASN)
c. * The **examination** *of the patients* was [on the table/in the bag]. (ASN)

In the absence of the object argument *of the patients*, the noun *examination* receives an RN reading, in which, similarly to *exam*, it denotes a concrete entity, which can lie on a table or be in a bag (see 4a). When the argument is realized, the synonymy with *exam* is lost, and the noun behaves like a nominalized verb, expressing an event, which can take a long time (see 4b), but cannot be on a table or in a bag (see 4c). In combination with *exam*, the phrase *of the patients* in (4b) could receive a possessive interpretation, i.e., the exam that belongs to the patients, but not that of an object argument of an examining event, since *exam* lacks such a reading. A similar interpretation would be possible in (4c) with *examination* on its RN reading.[2]

1.3 Compositionality and transparency in deverbal compounds

Compositionality has long been a prominent issue in theoretical linguistics with a first formalization offered in Montague's (1970) *Universal Grammar*. A simple formulation of the principle of compositionality in this tradition is given in (5).

(5) The principle of compositionality (PoC, Partee 1984: 281)
The meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined.

According to the PoC in (5), the interpretation of a complex expression relies on the individual meanings of its parts and their syntactic combination. Leaving technical details aside, an expression like *to kick the bucket* will be interpreted compositionally from the meanings of the verb *to kick* and of the noun phrase *the bucket*, via a verb–direct object syntactic relationship and the corresponding

[1] For the sake of simplicity, we leave aside Grimshaw's third possible reading of deverbal nouns as simple event nominals, since, from the perspective of the properties we consider here, they pattern with RNs and contrast with ASNs in similar ways.
[2] In her examples, Grimshaw strictly uses *of the patients* on its argument interpretation.

semantic relation. On this compositional reading, this expression is semantically transparent both with respect to the meanings of the parts and the syntactic-semantic relationship: the object *the bucket* is semantically interpreted as a patient of the kicking. However, *to kick the bucket* also has the idiomatic reading *to die*, on which neither the meanings of the two parts, nor any syntactic relationship between them can be compositionally retrieved. There is nothing particular about kicking or buckets or the verb–direct object relationship between them to be found in the meaning of *to die*. This reading is non-compositional and opaque.

Some idiomatic expressions, however, may be partially compositional. For instance, in *to spill the beans* 'to divulge a secret', the verb–object relationship is preserved in the idiom meaning and, while the object *beans* is lexico-semantically unrelated to *secret*, the verb *to spill* shares lexical semantic properties with *to divulge* (i.e., 'to let out'), which can be viewed as its figurative meaning. In this expression, the non-head is opaque, the head is partially transparent, and the relationship is compositional and transparent. The head is only partially transparent because it is ambiguous and the meaning *divulge* is not its basic meaning.

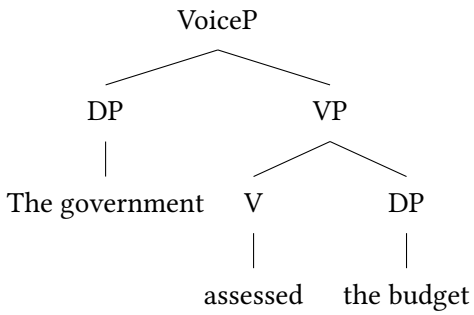
Deverbal compounds offer another pattern of expressions that are not fully compositional – yet, one different from the idioms above. The interpretation of DCs usually relies on a syntactic-semantic relationship between the base verb of their head noun and their non-heads, as shown in (1). Unlike in the corresponding verbal phrases, however, the syntactic relationship is not overt in DCs: e.g., *budget* in the DC *budget assessment* is not marked with accusative case as in the corresponding verb phrase in (1a), and *police* in *police questioning* is not marked by nominative case in (1b).[3] In the absence of overt marking, it is often unclear how to interpret the non-head of a DC, as, for instance, in *police killing*, where *police* could be either the object or the subject of *kill*. The indeterminacy of the syntactic relationship leads to ambiguity, which reduces the transparency of DCs from the perspective of syntactic compositionality, even though the meanings of the parts are transparent (by contrast with *beans* in *to spill the beans* or *kick* and *bucket* in the idiom *to kick the bucket*).

Yet, following the PoC and the compositional make-up of a sentence, if a particular DC is built up compositionally in parallel to the corresponding verbal phrase, then an object interpretation of the non-head is expected. This is the thesis we will follow and support here. But why does an object non-head indicate compositionality and, e.g., a subject does not? The reason follows from simple sentence structure. Transitive verbs form immediate constituents with their direct objects

[3]In a morphologically poor language like English, case marking comes from the syntactic position of the noun phrase, which is also missing in DCs, given their fixed word order.

but not with their subjects, which is why in sentence structure we first form a VP from the verb and its object, and the subject attaches afterwards, usually under a different projection such as VoiceP (or little vP), as in (6) (see Chomsky 1995 and Kratzer 1996 for a discussion on the differences between objects and subjects with respect to the event structure of verbs).

(6)



A DC based on the construction in (6) contains *two* nouns: one is the head derived from the verb and the other is the non-head. The latter can realize only one of the two arguments of the verb. Given the hierarchical structure in (6), this must be the object: see *budget assessment*. Nothing prevents the original subject from being realized as a non-head (e.g., *government assessment*). In that case, however, the DC does not follow the compositional make-up in (6), since the object is missing and the subject cannot form a constituent with the transitive verb alone. Such a DC will be interpreted by means of world knowledge, similarly to RCs as in (2). From this perspective, the subject behaves just like an adjunct/modifier, since it does not play any role in the compositional make-up of the DC.[4]

The importance of compositionality in language use is undebatable: without recursive compositional rules, speakers would not be able to produce and understand infinitely many sentences (Dowty 2007). That compositionality in DCs imposes an object interpretation, as predicted by the structure in (6), is supported by the fact that the default reading of a possibly ambiguous DC like *police killing* is that with an object non-head; the subject reading becomes available if established by a particular context, as, e.g., recent discussion in the U.S. about police killing unarmed civilians. Similarly, out of context, *student evaluation* also receives an object interpretation. The subject reading is brought about by a particular social environment in which people talk about students evaluating their teachers. Moreover, as shown in the linguistic literature (Grimshaw 1990; Borer

[4]Indirect objects are included here as well, since they also attach to the verb after the direct object does: see Larson (1988).

2013; Iordăchioaia et al. 2017), if a DC type is compositionally derived from a VP, it should also be fully productive: that is, any verb–object combination should be able to form a compositional DC, which is confirmed, for instance, by (7b). By contrast, not any subject–verb combination can form a DC: the non-heads in (7c) may at best receive a peculiar object interpretation, but not the subject reading of the corresponding sentence in (7a).

(7) a. A *boy*/*girl* broke the **window**/**pen**.
b. **window** breaking, **pen** breaking
c. * *boy* breaking, **girl* breaking

To summarize, ambiguous DCs as in (8) below are partially opaque, as the relationship between the two nouns is not explicit, and may receive several interpretations. However, if the DC is interpreted compositionally (in parallel to the verbal construction), it will be fully transparent and involve an object reading. The task remains to find independent evidence for the compositionality of a DC. In this respect, we will follow Grimshaw's (1990) distinction from Section 1.2 concerning the head nouns of DCs, as specified below in Section 1.5 and Section 4.1.

(8) a. policy/police/radio announcement (Object/Subject/Adjunct)
b. marketing approval, security assistance (Object/Subject/Adjunct)

1.4 Terminology

Before we introduce our research program, a few terminological clarifications are in order. The term *compositionality* is often used without particular focus on the syntactic-semantic relationship between the parts of the complex expression, an aspect that is of crucial importance in our study. Natural Language Processing literature on (root) noun-noun compounds, for instance, occasionally speaks of *compositionality ratings*, in which annotators evaluate how accessible the lexical meaning of the two nouns is in the overall meaning of the compound (see Section 2.2.2 for details and references). This notion of compositionality is similar to what we call *lexico-semantic transparency* below.

A notion of *compositionality* that is closer to ours appears in some Distributional Semantics (DS) approaches, which, in view of the PoC in (5), seek to identify linguistically-informed composite functions to combine the individual parts of complex expressions (Marelli & Baroni 2015; Baroni & Zamparelli 2010). Like

us, these authors take a closer look at the relationship between the parts; however, their focus is more on the technical implementation (i.e., the DS correspondent of function application from theoretical linguistics) rather than on the linguistically relevant constraints that are at play. Although we share the interest in the relationship between the parts with this literature, we are not concerned with the technical details of the function, but with how this relationship interacts with other morphosyntactic properties of the head, as explained in Section 1.5.

We use the terminology as follows: *compositional* refers to DCs that encode the structure in (6). Some may call this "syntactic compositionality". The term *transparent* is broader and allows two specifications. First, *lexico-semantically transparent* characterizes compounds whose parts are semantically fully recoverable from the compound meaning. These include all DCs as in (1) and (8), as well as some RCs like those in (2).[5] Second, what we would call *compositionally transparent* applies to DCs that, besides being lexico-semantically transparent, also follow the structure in (6). These correspond to our *compositional DCs*, since all the DCs we consider here are lexico-semantically transparent.

1.5 Our contribution

We start with the assumption that an important source of ambiguity in DCs such as in (8) is the ambiguity of their deverbal head nouns as in (4) and the correlated ambiguous relationship that they establish with the non-heads. The non-head is entirely transparent in DCs: its lexical semantics is present in the DC meaning, and, as an argument or adjunct, it brings no syntactic constraints to influence its syntactic-semantic relationship with the head noun. By contrast, the head noun is more complex. Its lexical semantics is also visible in the DC; yet, following Grimshaw's distinction in (4), its ambiguity between ASN and RN readings has a great impact on its syntactic-semantic relationship with the non-head. As perfect transpositions of verb phrases, ASNs follow the compositional structure in (6) and require objects to be realized first. RNs maintain only remote lexical connections to the verb base and do not inherit their compositional structure. Thus, RNs impose no syntactic requirements on the non-heads and are compatible with any syntactic-semantic relationship allowed by their lexical semantics.

Following this reasoning, our hypothesis is that DCs with ASN heads will obey the constituent structure in (6) and realize only objects as non-heads. These DCs will be both compositional and lexico-semantically transparent. DCs whose

[5]Other RCs like *hogwash* are substantially less transparent: see the previous literature in Section 2.2.

heads are RNs do not respect this structural condition and allow any interpretation that a context or world knowledge provide – whether related to the base verb or not (cf. *police building* 'building that hosts the police department'). In this respect, DCs headed by RNs are semantically similar to RCs; their deverbal morphology is irrelevant for their interpretation, since they are lexicalized. Such DCs are lexico-semantically transparent, but they are not fully compositional.

To test this hypothesis, we use a series of morphosyntactic properties that Grimshaw argued to be ASN-specific (see Section 4.1) and check their presence in the behavior of DC heads, on the basis of evidence from a large corpus of naturally occurring text. Since it is not a given fact that the ASN-features defined by Grimshaw can be reliably informed by corpora, we also gathered human judgments on ASN-hood – namely, we asked annotators to indicate to what extent the deverbal head refers to a process (or verbal event). By asking annotators directly for their judgments, we try to get an estimate for the latent variable that underlies the ASN properties defined by Grimshaw. We use these different types of data as features in a logistic regression classifier, by which we aim to predict the syntactic-semantic relation between the head and the non-head. These results are compared with the manually annotated interpretation of DCs.

Given our hypothesis and methodology, we expect that the ASN-features extracted from the corpus, as well as that based on human judgments, will point to an object interpretation of the DC (as predicted by 6) and will have high predictive power in determining whether the DC's non-head is an object or not. A high predictive power of the features will additionally show us that compositionality is an important aspect in the disambiguation of DCs.

First of all, our results indicate that all the ASN-hood features have predictive power above the chance level when tested individually and together. The most stable individual features point to an object interpretation, as expected under our hypothesis. Second, the ablation experiments show that many features overlap in the identification of ASN-hood, inviting to theoretical reflection on the individual contribution of these features. Third, the best feature is the manual annotation of ASN-hood, which confirms the importance of this property for interpreting DCs; it also indicates that either the morphosyntactic features are comparatively weaker or our corpus did not offer enough material for better results. Fourth, some weaker features raise stimulating questions especially relevant for linguistic investigation.

Our study investigates transparency strictly from the perspective of the compositional structure in (6). The degree of (lexico-semantic) transparency of DCs that do not receive such a verb-related compositional interpretation (i.e., those headed by RNs) goes beyond the scope of our present study and must be left for a future endeavor. As mentioned above, the role of world knowledge and context

is essential for such DCs. Therefore, such an investigation would need to employ a different methodology, more similar to that pursued in several computational studies as presented in Section 2.2.2. We also do not aim to measure speaker intuitions about the transparency degrees of DCs (as done in some of these computational approaches), although it would be interesting to compare such ratings with our relation-based annotations in the future. Our present study conceptually differs from these computational approaches, as it addresses the transparency of DCs from a structural perspective. We use insights from theoretical linguistics on the morphosyntactic properties of the deverbal noun heads of DCs and general principles of syntax-semantics mapping, and test these theoretical hypotheses with corpus-based and computational methods.

We start with an overview of relevant previous studies from theoretical linguistics (TL) and natural language processing (NLP) in Section 2. Sections 3 and 4 describe our data collection and methodology; Section 5 presents our experiments, followed by a discussion in Section 6. We draw our conclusions in Section 7.

2 Previous literature

In Section 2.1 we introduce the main theoretical concepts that have guided our investigation and briefly refer to previous analyses of DCs relevant to our assumptions. Section 2.2 presents the NLP literature on deverbal and root noun-noun compounds and the extent to which these studies can be compared with ours.

2.1 Theoretical approaches to DCs

Deverbal compounds have been at the forefront of theoretical linguistics since the early days of generative grammar. Especially beginning with the 1970s, after Chomsky's (1970) *Remarks on nominalization*, the theme of the theoretical debate has been whether word formation is part of the syntax or the lexicon. Syntactic approaches have argued that DCs behave systematically enough to be accounted for by syntactic rules (Roeper & Siegel 1978; Ackema & Neeleman 2004); lexicalist approaches have pointed out peculiar properties of DCs, which would require their analysis as part of the lexicon (Selkirk 1982; Lieber 2004).

The syntax vs. lexicon debate is relevant for our study in so far as recognizing a syntactic component in DCs leads to their compositional analysis, while specifying lexical rules for them suggests that they are like RCs and lack a systematic morphosyntax that preserves phrase-like compositionality. Meanwhile, both theoretical trends have argued for both kinds of analysis of DCs, and we will abstract away from the type of framework to focus on the properties of DCs.

Noteworthily, in theoretical studies the problem of compositionality in DCs is not addressed with respect to the contribution of the two individual nouns as done in recent NLP studies (see Section 2.2). If available at all, implications on compositionality come indirectly from the claims on the make-up of DCs and the structural relationship between their parts as in (6) (see Section 2.1.2).

2.1.1 Morphosyntactic properties of ASNs

In support of the contrast illustrated in (4), Grimshaw (1990) argues that deverbal nouns in their ASN reading exhibit a special morphosyntactic behavior, which is not shared by RNs. Table 1 is a summary of the main contrastive properties of ASNs (vs. RNs) from Grimshaw (1990) that are relevant for our study, adapted from Alexiadou & Grimshaw (2008: 3). These properties are positively specified for ASNs only, since RNs behave like non-derived lexical nouns and do not present any such particularities. The reasoning is that ASNs have verbal properties (i.e., event structure as in 6), which will impose restrictions on their nominal behavior (e.g., must appear in the singular) or make them compatible with verb-specific modifiers (e.g., aspectual adverbials).

Table 1: Morphosyntactic properties of ASNs vs. RNs

	Morphosyntactic property	ASN	RN
i.	Obligatory object arguments realized as *of*-phrases	Yes	No
ii.	Agent-oriented modifiers (*deliberate*, *intentional*, *careful*)	Yes	No
iii.	*By*-phrases are (subject) arguments	Yes	No
iv.	Aspectual *in*/*for-X-time* adverbials	Yes	No
v.	*Frequent*, *constant* appear with singular	Yes	No
vi.	Must appear in the singular	Yes	No

The realization of object arguments is a necessary and sufficient condition for ASNs. It indicates the presence of verbal event structure, which associates with the other ASN-properties. However, the morphosyntactic means to introduce an object argument in nominals is an *of*-phrase, which may also express possession. Given this ambiguity, using an *of*-phrase in combination with other ASN-properties is more reliable. For instance, in (4b), the predicate *took a long time* requires an event as a subject, which shows that *the examination of the patients* is an ASN, while *the exam of the patients* is not. As mentioned above, in the latter case *of the patients* expresses a possessor of the entity *exam*.

Agent-oriented adjectives like *deliberate, intentional, careful* are also taken by Grimshaw (1990: 51–52) to depict ASNs. Like *of*-phrases, possessive marking is ambiguous between expressing subject arguments, as in (9b), and possessive modifiers, as in (9c). Agentive modifiers, however, require verbal event structure with a subject (agent) argument, which cannot be available in the absence of the object argument in (9a) and (9c) (cf. the hierarchy in 6). The contrast between (9a) and (9b) shows that the possessive *the instructor's* cannot introduce the subject argument, if the object argument is not realized.

(9) a. * The instructor's *intentional/deliberate* examination took a long time.
b. The instructor's *intentional/deliberate* examination *of the papers* took a long time. (ASN)
c. the instructor's (**intentional/*deliberate*) book

In ASNs, *by*-phrases have a function similar to that of the possessive in (9b): they introduce the subject argument. Yet, like the possessive and *of*-phrases, *by*-phrases may also introduce modifiers. In (10a), the *by*-phrase acts as a modifier of the lexical noun *book*, which has no event structure. In (10b), however, it introduces the subject argument of an ASN, the same way the possessive does in (9b). (10c) is ungrammatical, because the agent-oriented modifiers *intentional/deliberate* require a subject argument, which the *by*-phrase cannot introduce in the absence of event structure and the object: (10c) parallels (9a).

(10) a. a book *by Chomsky*
b. The intentional/deliberate examination *of the papers by the instructor* took a long time. (ASN)
c. * The intentional/deliberate examination *by the instructor* took a long time.

Given the verbal event structure and the correlated aspectual properties of ASNs, they are expected to allow aspectual adverbials and to obey the aspectual restrictions of their base verbs. In (11a), the telic verb *destroy* allows *in-* but not *for-*adverbials. The correlated ASN in (11b) exhibits the same constraint. By contrast, simple nouns that lexically denote events such as *trip, process* are incompatible with such modifiers in (11c), although they occupy time, as shown by (11d). The latter pattern with RNs (Grimshaw 1990: 58–59).

(11) a. The bombing destroyed the city *in/*for only 2 days*.
b. The total destruction of the city *in/*for only 2 days* appalled everyone.

c. * The process/John's trip *in/for 5 hours*
d. The process/John's trip *took 5 hours*.

Finally, Grimshaw argues that, due to their verbal structure, ASNs, in general, disallow plural marking, and when plural is available it indicates an RN reading. This is illustrated in (12) from Grimshaw (1990: 54). Related to this and the aspectual contrast in (11), Grimshaw notes that aspectual modifiers like *constant, frequent* will combine with a singular ASN, but with a plural RN. These modifiers require habitual/iterative aspect, which is made available by the event structure of ASNs, but not by the lexicalized RNs. The latter need the plural to contribute the iterative meaning: see (13a)/(13b–c).

(12) a. The *assignments* were long. (RN)
b. * The *assignments of the problems* took a long time. (ASN)
c. The *assignment of that problem* always causes problems. (ASN)

(13) a. * The *constant* assignment is to be avoided. (RN)
b. The *constant* assignment of unsolvable problems is to be avoided. (ASN)
c. The *constant* assignments were avoided by students. (RN)

In (9) to (13), the contrasts between ASNs and RNs are clear. Yet, depending on the lexical semantics of the individual nouns, the application of these tests may exhibit quite a bit of variation, which led many to challenge Grimshaw's generalizations. For instance, Alexiadou et al. (2010) show that in some languages, ASNs may pluralize provided particular aspectual properties, while Grimm & McNally (2013) and Lieber (2016) challenge some of Grimshaw's claims with counterexamples attested in corpora. However, a general tendency of ASNs to exhibit the properties in Table 1 cannot be denied. At least so far, no corpus study has offered a quantitative analysis to prove that these properties are irrelevant for ASNs. From this perspective, our study can also be viewed as testing the relevance of these properties on the basis of deverbal compounds, which, according to Grimshaw, are headed by ASNs (see Section 2.1.2).

2.1.2 Deverbal compounds between ASNs and RNs: Grimshaw (1990)

Let us now consider DCs from the perspective of the documented ASN vs. RN contrast. We focus on Grimshaw's analysis of DCs and on Borer (2013), the latter of which reviews Grimshaw's arguments to support an opposite position.

In her study of nominalization, Grimshaw (1990) argues that the heads of DCs (i.e., her *synthetic compounds*) are ASNs. Her reasoning relies on the observation that DCs obey argument structure constraints in the realization of their non-heads. In her model of argument realization, she proposes the hierarchy of argument roles in (14), such that the lower arguments (from right to left) must be realized syntactically before the higher ones. This means that the theme, i.e., the syntactic direct object, must be realized before the goal (indirect object) and the agent (subject). This thematic hierarchy reminds us of the constituent structure of verb phrases in (6).

(14) Agent (subject) > Goal (indirect object) > Theme (direct object)

Grimshaw argues that DCs obey the hierarchy in (14), since they disallow non-heads that realize other arguments than the theme (object). (15) repeats two of her examples. Her explanation is that, when occurring in DCs, deverbal nouns such as *giving* and *reading* are disambiguated to an ASN interpretation.

(15) a. They give **gifts** to *children*.
DC: **gift**-giving to children vs. **child*-giving of gifts
b. *Students* read **books**.
DC: **book**-reading by students vs. **student*-reading of books

In contrast to suffix-based deverbal nouns as in (15), she considers zero-derived nouns like *a sting* and *a bite* to always be RNs. She shows that the compounds these may head need not obey the hierarchy in (14) and allow agent non-heads. The grammatical compounds in (16) are RCs for Grimshaw.

(16) **bee** sting (vs. *bee-stinging), **dog** bite (vs. *dog-biting)

2.1.3 Deverbal compounds between ASNs and RNs: Borer (2013)

In spite of her extensive study on ASNs, Grimshaw does not go to great lengths to compare DCs with ASNs in terms of morphosyntactic properties such as those in Table 1. Di Sciullo (1992) investigates some of these tests in further support of the similarity between DC heads and ASNs. However, two decades later, Borer (2013) challenges Grimshaw's analysis of DCs by using some of these morphosyntactic tests. She argues that the behavior of DCs essentially differs from that of ASNs, and proposes that all DCs are headed by RNs.

We retain three of Borer's arguments. First, she argues that, unlike ASNs, DCs disallow aspectual *in*/*for*-adverbials and, second, that they also disallow argumental *by*-phrases. This contrast is illustrated in (17) (cf. 11 and 10). In Borer's

system, the unavailability of aspectual modifiers indicates that event structure (with arguments) is entirely missing from DCs, so they cannot involve ASNs. Her conclusion is that DCs are headed by RNs and behave just like RCs.

(17) a. the demolition of the house **by the army** *in 2 hours* (ASN)
b. the stabbing of the emperor **by Brutus** *for 10 minutes* (ASN)
c. the house demolition (***by the army**) (**in 2 hours*) (DC)
d. the emperor stabbing (***by Brutus**) (**for 10 minutes*) (DC)

Third, Borer claims that the object reading of non-heads in DCs is just as available as a subject reading, depending on context. As evidence, she quotes DCs as in (18), parallel to those in (1b), whose non-heads may correspond to subjects.

(18) teacher recommendation, court investigation, government decision

Some criticism and re-interpretation of Borer's facts is found in Iordăchioaia et al. (2017) and Iordăchioaia (to appear). We briefly note here that aspectual adverbials are barely ever attested in corpora even with ASNs (Lieber 2016: 39–42), so an extensive empirical study is necessary to determine how much DCs differ from ASNs in this respect. Furthermore, *by*-phrases are broadly attested with DCs in corpora, as Grimshaw's (15b) also predicts, but they usually involve bare plurals and not definite noun phrases or proper names as in Borer's (17c–d). Given that DCs are often generic, this restriction is natural.

Having summarized these two theoretical approaches to DCs, we may add that we do not aim to argue for one or the other. Instead, we use morphosyntactic properties whose pertinence for ASN-hood is accepted by both to guide us in evaluating the impact of the head noun on the interpretation of the DC. Our hypothesis that a high level of ASN-hood in DC heads correlates with an object reading of the non-heads, however, follows Grimshaw's intuition that "true" DCs involve ASN heads and are fully compositional. By contrast, Borer's claim is that DCs are always ambiguous like RCs and never as compositional as ASNs. Given that our results support the correlation between ASN-properties and an object reading in DCs, they also bring some evidence against Borer's analysis.

2.2 Computational approaches to compounds

Compounds have been the focus of quite a number of papers in the field of computational linguistics (CL) and NLP. In view of the topic of this paper there are two strands of research that are most relevant. The first focuses on determining the relation between the two components of a compound, the head and the non-head.

For our study this work is relevant to the extent that it discusses compounds whose head is a deverbal noun. The second strand of research is concerned with modeling the lexico-semantic transparency of noun-noun compounds. We will start by discussing the former and finish with an overview of the work that predicts the degree of transparency in compounds.

2.2.1 Predicting the interpretation of deverbal compounds

The goal of computational work on deverbal compounds (referred to as nominalizations) has been to predict the relation between the non-head and the deverbal head. The relation inventory has varied from two classes, OBJ and SUBJ, in Lapata (2002), to three classes, OBJ, SUBJ and prepositional complement in Nicholson & Baldwin (2006), and to 13 classes – OBJ, SUBJ and further specifications of the prepositional complement in Grover et al. (2005).

These works have mostly focused on encyclopedic, usage-based features such as the syntactic relations attested between the base verb of the head noun and the non-head in large corpora. The underlying assumption is that the frequency distribution of syntactic relations between a given noun and a verb, for example, between *taxi* and *drive*, is a good estimate for the distribution of the underlying relation between *taxi* and *driver*. Additional pragmatic knowledge is obtained from the direct context of the compound. In selecting these pragmatic features, these works are in line with lexicalist theoretical approaches that list several covert semantic relations typically available in compounds (cf. most notably, Levi 1978; see Fokkens 2007, for a critical overview). In addition to these pragmatic features, some straightforward morphological features are selected, such as the suffix of non-heads ending in *-ee* and *-er* (Lapata 2002).

Our study differs from these works in several ways. First, our aim is not to reach state-of-the-art performance in prediction, but to test linguistic hypotheses by measuring the predictive power of the various features discussed in theoretical linguistics, which are also indicative of the compositionality of the compound.

Second, and related to the previous point, our features are all head-specific. This is because, following Grimshaw's theory, the behavior of the derived nominal heads (as ASNs or RNs) should mirror the structural correlation between DCs and the compositional structure of the original verb. The presence (or absence) of such a correlation is expected to have a great impact on the relation between the head and the non-head. In order to measure the individual impact of these theoretically-defined features, we do not rely on pragmatic features that involve both the head and the non-head as in the studies above.

Lastly, because our goal is to uncover in how far the behavior of the derived nominals (as ASNs or RNs) can predict the relation between head and non-head, we carefully selected equal numbers of DCs with the suffixes *-al*, *-ance*, *-ing*, *-ion*, and *-ment*. These suffixes are all ambiguous in their formation of ASNs and RNs, so we eliminate any bias for particular readings (cf. *-ee* and *-er*, Section 3.2).

2.2.2 Predicting the degree of transparency in noun-noun compounds

For the transparency of compounds two types of CL work are relevant, which focus on different tasks, but share the same assumptions. One type aims to predict the meaning of compounds based on composite functions between the vector-based representations of their parts, e.g., Ó Séaghdha (2008) and Mitchell & Lapata (2010). These works compare different types of mathematical functions for the combination of the vectors for heads and non-heads to best represent the meaning of compounds. In the same spirit, but closer to our interest in the syntactic-semantic relationship between the parts, Marelli & Baroni (2015) and Baroni & Zamparelli (2010) investigate linguistically-informed composite functions.

The other line of work aims to predict the degree of lexico-semantic transparency (i.e., what they call "compositionality"; cf. Section 1.3) of compounds. For this, they compare the vector-based representations of the parts and composite functions to the vector-based representations of the compound as a whole, e.g., Schulte im Walde, Hätty & Bott (2016); Reddy et al. (2011).

This second line of work also draws upon psycholinguistic insights, such as Libben et al. (1997; 2003), which groups noun-noun compounds into four different categories, depending on the transparency of the head and the non-head. The four classes are: TT for compounds with both a transparent head and non-head, OO for compounds with opaque heads and non-heads, and OT and TO for compounds whose parts differ along the dimension of transparency. They found that both semantically opaque and semantically transparent compounds show morphological constituency. However, they found the semantic transparency of the head to play a significant role. This confirms previous results from the psycholinguistic literature (Zwitserlood 1994).

In this literature, several datasets have been created, which collect human ratings on the degrees of lexico-semantic transparency of compounds with respect to their constituents: e.g., in English (Reddy et al. 2011; Juhasz et al. 2015) and in German (Schulte im Walde, Hätty, Bott & Khvtisavrishvili 2016). Schulte im Walde, Hätty, Bott & Khvtisavrishvili (2016) have enriched the semantic transparency ratings with several empirical features related to the constituents of

the compound in order to measure the influence of these features on the transparency of the compound. These features include:

- Corpus frequencies of the compounds and their parts;
- Productivity of the parts, as in the number of compound types the part (head/non-head) appears in;
- Number of senses for the parts as retrieved from GermaNet (Hamp & Feldweg 1997; Henrich & Hinrichs 2010) for the German dataset and WordNet (Fellbaum 1998) for the English dataset.

Schulte im Walde, Hätty & Bott (2016) use vector space models to model the meaning of the compounds and their parts. Subsequently, they model the transparency of the compound by measuring the distance between the composite vector of its parts and the vector for the actual compound. The assumption behind this work is that the vectors of transparent compounds should be closer to the composite function of the vectors of their parts than the vectors of opaque compounds.

The main question Schulte im Walde, Hätty & Bott (2016) try to answer is whether the above-mentioned properties (frequency of the compound and its parts, productivity, and ambiguity of its parts) play a major role in the quality of the predictions. They found that for the head all properties had a significant effect on the predictions, whereas for the modifier the effect was not consistent. This converges with our results in predicting the compositionality of DCs from the properties of the head.

Furthermore, they attribute the influence of these features to the underlying ambiguity that they seem to be correlated with: e.g., frequent heads that are highly productive are often highly ambiguous. We note, however, that these studies are not concerned with DCs, as ours is, but especially with what we call RCs, some of which are lexico-semantically less transparent than our DCs (cf. *hogwash*).

3 Methodology

In this section we present the corpus and the tools for automatic pre-processing, the procedure in the DC extraction, as well as the annotation and post-processing of our collection of DCs.

3.1 Corpus and tools

For the selection of DCs and to gather corpus statistics on them, we exploited the Annotated Gigaword corpus (Napoles et al. 2012), one of the largest general-domain English corpora, which contains several layers of linguistic annotation. This corpus encompasses ten million documents from seven news sources and more than four billion words. We made use of the following available automatic preprocessing steps and annotations, which we accessed via the Java API provided along with the corpus: sentence segmentation (Gillick 2009), tokenization, lemmatization and POS tags (Stanford's CoreNLP toolkit[6]), and constituency parses (Huang et al. 2010) converted to syntactic dependency trees with Stanford's CoreNLP toolkit. The POS tags adhere to the Penn Treebank tagset (Santorini 1990); the dependency relations follow the Stanford typed dependencies (de Marneffe & Manning 2008). As news outlets often repeat news items in subsequent news streams, the corpus contains a considerable amount of duplication. To improve the reliability of our corpus counts, we removed exact duplicate sentences within each of the 1010 corpus files, reducing the corpus size by 16%.

3.2 Extraction of deverbal compounds

We created a balanced collection of DCs, which we extracted from the Gigaword corpus. We first gathered 25 nouns (over three frequency bands: high, medium, low) for each of the suffixes *-al*, *-ance*, *-ion*, *-ing*, and *-ment*. The highest frequency band ranges from 4.5 to 3.5 on the Zipf-scale (van Heuven et al. 2014), the medium frequency band ranges from 3 to 2.5, and the lowest one from 2 to 1.5. The suffixes may form both ASNs and RNs according to Grimshaw (1990).

We did not consider zero-derived nouns like *attack*, *abuse*, *bite*, because Grimshaw considers them RNs (see 16). We also excluded deverbal nouns based on the suffixes *-er* and *-ee*, as they denote event participants corresponding to the subject and the object of the base verb, respectively, implicitly blocking this interpretation on the non-head (cf. $police_{subj}$ *train**ee*** – dog_{obj} *train**er***). In our attempt to capture the closeness of DCs to ASNs (and the base verbs), we considered only the suffixes that build eventive nominals, which could realize both a subject and an object argument. DCs headed by *-ee* and *-er* nouns would have been biased for one or the other. However, our selection of suffixes represents the large majority of deverbal nouns. They make up 69.4% of the total number of deverbal nouns in the NOMLEX database (Macleod et al. 1998), which consists of 1025 lexicalized deverbal nouns.

[6]http://nlp.stanford.edu/software/corenlp.shtml

The nouns were selected such that their base verbs present transitive uses, making both subjects and objects available.[7] For illustration, Table 2 offers samples of deverbal nouns per each frequency range and suffix. For each such selected noun we then extracted the 25 most frequent compounds that they appeared as heads of, where available. A few deverbal nouns (in particular those with suffixes *-al* and *-ance*) were less productive in compounds and appeared with fewer than 25 different non-heads. Given these gaps and after removing a few repetitions due to capitalization, we obtained a collection of 3111 DCs.

Table 2: Samples of extracted deverbal nouns

Frequency	*-al*	*-ance*	*-(at)ion*	*-ing*	*-ment*
High	approval withdrawal rental	performance assistance surveillance	protection reduction consumption	building training trafficking	development movement punishment
Medium	renewal survival upheaval	assurance dominance tolerance	supervision cultivation instruction	killing counseling teaching	deployment placement adjustment
Low	retrieval disapproval dispersal	defiance endurance ignorance	demolition expulsion deportation	weighting chasing mongering	reinforcement empowerment abandonment

3.3 Annotation and post-processing of DCs

3.3.1 Interpretation of (non-heads in) DCs

All DCs were annotated by three trained American English speakers, who had a university level background in linguistics. They had to label the DCs as OBJ(ect), SUBJ(ect), OTHER, or ERROR, depending on the syntactic relationship that they considered the DC to establish between the base verb of the head noun and the non-head. For instance, DCs such as in (1) would be labeled as OBJ (1a), SUBJ (1b), and OTHER (1c). OTHER was an umbrella label for prepositional objects (e.g., *adoption counseling* 'somebody counsels somebody *on adoption*'), various adjuncts (e.g., *ultrasound examination* 'to examine somebody *with an ultrasound*', *sea burial* 'to bury somebody *by the sea*', *surprise arrival* 'somebody/something arrived *by surprise*'). ERROR was intended to identify errors of the POS tagger (e.g., *face abandonment* originates in 'they face$_V$ abandonment'), but was also employed by the

[7] *Arrive* is the only intransitive unaccusative verb that realizes the object/internal argument as a subject.

annotators when they considered the DC uninterpretable or ungrammatical. We allowed the annotators to use multiple labels and to indicate ambiguity (using "–") and the preferred order of the readings (using ">").

We used the original annotations to create a final list of compounds with the labels that all three annotators agreed on. For ambiguously labeled DCs we selected the one reading available for all three. If they all indicated the same ambiguity for a DC, we labeled the DC as ambiguous. The labels we used for the final dataset are OBJ, SUBJ, OTHER, DIS(agreement between annotators), AMBIG(uous), and ERROR. In spite of Borer's (2013) claims, we found only two cases of ambiguity which all three annotators agreed on – namely, *police killing* and *doctor referral*, which were both labeled SUBJ–OBJ. In the end we identified 772 DIS, 1377 OBJ, 404 OTHER, 286 SUBJ, and 270 ERROR cases of DCs. After removing the disagreements, the two ambiguous DCs and the errors, we obtained 2067 DCs. We based our study on the agreed-upon relations only. We note, however, that the simple inter-annotator agreement (IAA) among the three annotators, excluding the errors, was 72.8%. In a previous study with only two annotators (Iordăchioaia et al. 2016), the IAA was 81.5%.

We kept two versions of the data: one in which the classes OTHER and SUBJ are separate and one in which we conflated them to NOBJ (non-object). Given the purpose of this paper, i.e., verifying to what extent the OBJ reading of a DC correlates with particular morphosyntactic properties of the head noun, we focus here on the binary classification. The resulting data set is skewed with OBJ prevailing: 1377 OBJ and 690 NOBJ.

3.3.2 Process vs. result readings in DCs

An additional annotation task concerned feature "7. *process-vs-result*" from Table 3 in Section 4.1. This feature was designed to capture the three annotators' judgments with respect to how close the interpretation of the DC comes to the ASN and the verbal expression of a process/event in which the non-head is realized as SUBJ, OBJ, or OTHER. They had to rate DCs from 5 (very prominent process) to 1 (no process = result) (see Grimshaw 1990).

We first explained the difference between an ASN and an RN to them as follows: "*The teacher's* ***assignment*** *of tasks* expresses a process in which the teacher assigns tasks. However, in *this long* ***assignment*** *took several hours to complete*, the noun *assignment* is interpreted as a result of the process of assigning something – namely, the task itself." We then instructed the annotators to check this contrast in DCs like *task assignment* and *Math assignment* and rate the ones that relate to the process as closer to 5 and those that relate to the result as closer to 1. Another example was *apartment building*, which should be rated as closer to 5,

if they interpret it as 'to build apartments', and closer to 1, if they interpret it as 'a building with apartments'. We fully encouraged the annotators to employ the scores 4, 3, 2 for unclear cases.

During this task, the annotators had access to their previous SUBJ/OBJ/OTHER annotation labels for each DC and could compare different DCs headed by the same head noun. In terms of the variation of ratings between DCs headed by the same noun, one annotator in particular assigned pretty similar scores, although the contrast was clear. This annotator also showed a tendency towards the extremes: either 5 or 1. In general, the task was perceived as difficult, especially by this annotator. We multiplied the scores from 5 to 1 by 20 to use them as percentages. For each DC we calculated the average between the three annotations obtaining values between 20 and 100.

4 Feature selection

4.1 Theoretical considerations

To collect information on the properties of the head nouns in DCs, we defined a total of nine features, given in Table 3.

The first seven features are inspired by Grimshaw (1990), although only the first four directly correspond to the properties in Section 2.1.1. Two adjustments led us to four features instead of the six properties in Table 1: first, *in*/*for*-adverbials were discarded, because we found close to no relevant data; second, we counted agent-oriented and aspectual adjectives together, as they were also very few.[8] In line with our hypothesis, we expect all these seven features to have predictive power and to point to an OBJ interpretation of the DCs.

Feature *of_outside_DC* encodes the first property in Table 1. Here we counted the percentage of occurrences of a (singular) head noun in which it also realizes an *of*-phrase. For feature *by_outside_DC* (i.e., the third property in Table 1), we collected the frequency of a *by*-phrase with a head noun. Feature *sum_adjectives* collects all the (singular form) occurrences of the head nouns in a modifier relation with agent-oriented or aspectual adjectives (cf. second and fifth property in Table 1).[9] Feature *sg_outside_DC* measures the percentage of singular occurrences of the head noun out of its total occurrences in the corpus (cf. last property in Table 1).

[8] We initially collected data on *in*- and *for*-adverbials, but only a few nouns had such occurrences. At closer inspection even these examples turned out not to illustrate *in*- and *for*-phrases that modify the telic/atelic aspect of the head noun, as Grimshaw and Borer used them. Instead, they mostly functioned as temporal modifiers, and we therefore discarded this feature.

[9] Note that, given Grimshaw's assumption that ASNs do not appear in the plural, we counted all of these occurrences in the singular form of the head noun.

Table 3: Indicative features for head nouns

Feature label	Description and illustration
1. of_*outside_DC* (Grimshaw 1990)	Percentage of the head's occurrences as singular outside compounds which realize a syntactic relation with an *of*-phrase. E.g., *assignment **of** problems*
2. by_*outside_DC* (Grimshaw 1990)	Percentage of the head's occurrences in the singular outside compounds which realize a syntactic relation with a *by*-phrase. E.g., *assignment (of problems) **by** teachers*
3. *sum_adjectives* (Grimshaw 1990)	Percentage of the head's occurrences in a modifier relation with one of the adjectives *frequent*, *constant*, *intentional*, *deliberate*, or *careful*.
4. *sg_outside_DC* (Grimshaw 1990)	Percentage of the head's occurrences as singular outside compounds.
5. by_*inside_DC* (≈ 2. by_*outside_DC*)	Percentage of the head's occurrences as singular inside compounds which realize a syntactic relation with a *by*-phrase. E.g., *task assignment **by** teachers*
6. *sg_inside_DC* (≈ 4. *sg_outside_DC*)	Percentage of the head's occurrences as singular inside compounds.
7. *process-vs-result* (≈ ASN vs. RN)	Native speaker annotation of each DC as a process (*car driving*) or result (*apartment building*) on a scale from 5 to 1.
8. *suffix* NEW	Suffix of the head noun: *-al* (rent**al**), *-ance* (insur**ance**), *-ing* (kill**ing**), *-ion* (destruct**ion**), *-ment* (treat**ment**)
9. *head_in_DC* NEW	Percentage of the head's occurrences within a compound out of its total occurrences in the corpus.

Grimshaw's properties in Table 1 characterize deverbal nouns as ASNs when they appear on their own, i.e., *outside* compounds. This is why features 1. to 4. are labeled correspondingly. Yet, if DCs are supposed to resemble ASNs, we considered that their head nouns should preserve these properties also within DCs, i.e., when the head noun is *inside* a DC.[10] For this reason, we also introduced the features *sg_inside_DC* and *by_inside_DC*. The former measures the percentage of singular DCs out of their total occurrences, and the latter the percentage of DCs that realize a *by*-phrase. We did not test *of*-phrases inside DCs, since DCs usually realize the object as a non-head (see our annotation results in Section 3.3.1) and collecting such occurrences would have mostly delivered noise. The adjectives modifying DCs were also left out, because their number was close to inexistent.

There are two caveats to these features inspired by Grimshaw (1990). First, as we noted in Section 2.1.1, the individual ASN-properties are not fully reliable in determining ASN-hood: e.g., there is ambiguity in argument marking (i.e., *of*- and *by*-phrases), and deverbal nouns are easily coerced between the readings. For this reason, Grimshaw used several such properties together in her examples. However, we extracted these data from corpora, and most of the attestations were too few to allow any combined patterns beyond the one we ensured – that of a singular form of the head noun in each of the other properties. Second, and related to this, basing our study on a corpus comes with the risk that, no matter how large the corpus, it may not present enough relevant data. It was for these two reasons that we considered adding three more head-related features to our study. We first gathered native-speaker intuitions about the ASN vs. RN status of the head nouns in DCs (see feature *process-vs-result*) and supplemented Grimshaw's tests with information about the suffix and the frequency of the head noun within compounds (features *suffix* and *head_in_DC*).

We designed feature *process-vs-result* (P-R) in order to grasp Grimshaw's intuition about the contrast between ASNs and RNs by means of introspection. The process vs. result interpretation is the fundamental difference between ASNs and RNs in Grimshaw's understanding. It can be seen as the latent variable that her morphosyntactic properties are intended to identify: ASNs express processes or events like the corresponding verbs, while RNs depart from this meaning and express results. Following this annotation (see Section 3.3.2), we gathered information on how salient the verbal process is in the meaning of a DC and, indirectly, how accessible the compositional structure of the base VP is within the DC.[11]

[10] Di Sciullo (1992) and Borer (2013) apply the same reasoning.

[11] The way we gathered estimates for our P-R feature comes close to the NLP studies which gather native speaker evaluations about the transparency of compounds. Namely, our three annotators had to evaluate how close the morphosyntactic (and semantic) relationship between the head noun and the non-head comes to the fully compositional relationship between the corresponding verb and its argument or adjunct.

The last two features *suffix* and *head_in_DC* represent two further properties of the head nouns that we considered interesting for our study. The theoretical literature does not offer much on suffixes. *-Ing* has received most attention, to the extent that Grimshaw argued that it always forms ASNs, while Borer claims that it encodes what she calls an originator (i.e., subject argument), with the effect that in compounds, the SUBJ reading is blocked for non-heads and OBJ is favored. Neither contention is true. First, *-ing* presents several examples of RNs (see *building(s)*, *writing(s)*, *reading(s)*). Second, we *do* find SUBJ-DCs headed by *ing*-nouns (see 1b). In general, the information on the suffix is independent of ASN-hood, since all suffixes allow both ASN and RN readings, but we aimed to check whether some suffixes may be more informative than others.

Feature *head_in_DC* delivers us the degree of compoundhood of a deverbal noun, i.e., how likely it is to appear within a compound. The expectation is that a noun that typically appears in compounds has undergone some meaning specialization, which requires another noun to be instantiated. One may rightly say that this makes the meaning of such head nouns less transparent than for those that freely appear both within and outside compounds. However, for deverbal nouns, to the extent that this slight meaning specialization requires a particular type of non-head, it can give us useful information about which (morpho)syntactic relationship between the base verb and one of its arguments is most likely to form a DC. If it is a non-OBJ relation, this shows that compositionality as in (6) is not a typical condition in the formation of DCs, weakening the relevance of our investigation. However, our results in Table 7 below indicate that high compoundhood correlates with an OBJ interpretation of the non-head, which supports the relevance of compositionality in the formation of DCs.

4.2 Technical support

To obtain statistics for the morphosyntactic features, we extracted counts for the selected DCs and their head nouns from the Gigaword corpus by matching patterns defined over word forms, lemmas, POS tags and dependency relations, as provided by the automatic corpus annotations. The specific patterns used for each feature are detailed in the following.

For the *inside_DC* features we extracted DCs from the Gigaword corpus by locating two adjacent nouns according to the POS tags NN for singular nouns and NNS for plural nouns, and excluding noun pairs directly preceded or succeeded by other nouns or proper nouns (POS tags NNP and NNPS). DCs were matched with the word form of the non-head and the lemma of the head, thereby extracting singular and plural occurrences. We determined the grammatical number of

a noun or compound by its POS tag or the POS tag of its head, respectively. For example, we matched *security training(s)*, but not *airport security training* and *security training instructor*, to make sure that we do not extract parts of larger compounds. Conversely, the *outside_DC* features apply to head nouns (matched by their lemma and POS tag NN or NNS) without any noun or proper noun next to them.

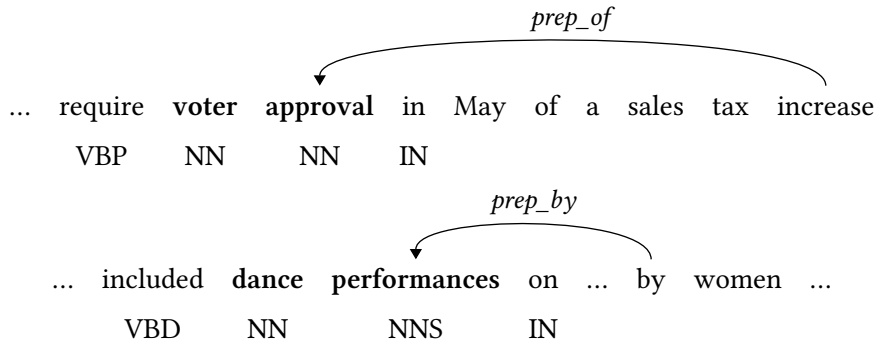


Figure 1: Illustration of morphosyntactic patterns to extract DCs heading *of-phrases* (top) and *by-phrases* (bottom)

We counted a DC (or its head noun) as being in a syntactic relation with an *of*-phrase or *by*-phrase, if it (or its head) governed a collapsed dependency labeled "prep_of"/"prep_by"[12], as in Figure 1. Since we were interested in prepositional phrases that realize internal or external arguments, but not in temporal phrases (e.g., *by Monday*) or fixed expressions (e.g., *of age*, *by chance*), we excluded phrases headed by words that typically appear in these undesired constructions. We semi-automatically compiled these lists based on a multiword expression lexicon[13] and manually added entries. To compute the feature *sum_adjectives* we counted how often each noun outside a DC governs a dependency relation labeled "amod", where the dependent is an adjective (POS tag JJ) out of the lemmas *intentional, deliberate, careful, constant*, and *frequent*.

[12]By conflating dependencies involving prepositions or conjuncts, collapsed dependencies directly link content words. This simplifies the extraction patterns, as we can obtain the complement of the prepositional phrase depending on the noun or the DC, by following a single dependency arc.

[13]http://www.cs.cmu.edu/~ark/LexSem/

4.3 Reliability of the extracted features

Our extracted features rely on the automatic corpus annotations, the manually defined extraction patterns, and, in the case of the *of*-phrases and *by*-phrases, on heuristics, to exclude undesired matches of temporal phrases or fixed expressions. The constituency parser, which was used to obtain the syntactic analyses then converted to dependency trees, obtained an average F1-score of 91.4% on a standard test set, Section 22 of the Wall Street Journal corpus (Huang et al. 2010).

To measure the reliability of the extracted features, more in particular the most error-prone features based on heuristics, we exemplarily conducted a manual analysis of the counts of head nouns that appear in conjunction with *of*-phrases and *by*-phrases. For this, we implemented the following pattern to extract all candidate sentences in the corpus for this feature. We selected all sentences in which one of the target head nouns outside a compound was followed by a token with lemma *of* or *by* and POS tag IN, not separated by a punctuation mark.[14] On the one hand, this was driven by the motivation to keep the number of sentences on a manageable level and focused on the feature of interest. On the other hand, we designed the pattern to maximize recall so as not to miss out on any true positives. We then randomly selected 2000 of these sentences for each preposition for a manual annotation of the target features by a single human annotator. A comparison of the annotated instances with the automatically extracted instances revealed a precision of 91.0% and recall of 90.1% for *of*-phrases, while the results for *by*-phrases were lower (85.0% precision, 73.8% recall).

5 Data exploration with machine learning techniques

Our goal is to test the features listed in Table 3 for their predictive power in determining the relation between the head and the non-head. These features are composed of numerical (1 to 7, and 9) and categorical features (8). The dependent variable is a binary feature that varies between one of the two annotation labels, OBJ and NOBJ. We trained a logistic regression classifier to model the effect of these features.[15]

We divided the data described in §3.3.1 into a test and a training set. Because the features are all head-specific, as can be seen in Table 3, the model was tested on a test set for which we ensured that neither compounds, nor heads were seen in the training data. Therefore, we randomly selected two mid-frequency heads

[14] We used the following list of punctuation characters: ".", "?", "!", ";", ":", ",".

[15] We used version 3.8 for Linux of the Weka toolkit (Hall et al. 2009) and experimented with several other classifiers that have interpretable models (decision trees), but also support vector machines and naive Bayes classifiers. All of these underperformed on our test set.

for each suffix and removed these from the training data to be put in the test data. We expect mid-frequency heads to lead to most reliable results, because high-frequency heads may show higher levels of idiosyncrasy and low-frequency heads may suffer from data sparseness.[16] This resulted in a division of roughly 90% training and 10% testing data.[17] The data set resulting from the annotation effort is skewed with OBJ being the majority class. Our selection of test instances introduces further differences in proportions of OBJ and NOBJ in the test and training set. Therefore, we balanced both the training and test set by randomly removing instances with the OBJ relation (the largest class) until both classes were equal in size.[18] The balanced training set consisted of 1248 examples, and the test set of 132 examples.

We compared our models with the random baseline, and two additional baselines to make sure that the features we are proposing are not just a by-product of the impact of simpler variables. We computed the relative[19] frequency of the head and the relative family size, i.e., how many compound types we find with a given head.[20]

We ran ablation experiments to determine the individual contribution of each feature in addition to the other features. However, because features might be interdependent and one feature could overshadow another, we first looked at the performance of each feature individually. This way, we could measure the exact predictive power of each individual feature in comparison to the baselines. Lastly, we combined the top-*n* features from ablation experiments and individual feature experiments to see the overall predictive potential of the model.

The first row in Table 4 shows that, when using all features, the classifier significantly outperforms[21] the baselines with a large margin (78.8%). This proves that the combination of features driven by linguistic theory has strong predictive power.

[16] We remind the reader that our goal is not to determine the realistic performance of our model, but to measure the contribution of the features. Therefore we believe that the bias introduced by selecting mid-frequency items for the test set is acceptable.

[17] Multiple divisions of training and test data would lead to more reliable results, but we have to leave this for future work.

[18] We also ran experiments with non-balanced data, because we reasoned that more data might result in higher performance, but the performance proved to be comparable. A balanced dataset facilitates comparisons to the random baseline of 50%.

[19] By providing relative counts, we make sure these features are on the same scale as our other features.

[20] These additional baselines were computed on a slightly different test and training set, due to the random process in balancing the data.

[21] Significance numbers for these experiments, in which training and test data are fixed, were computed with a McNemar test with $p < 0.05$, as it makes relatively few type I errors (Dietterich 1998).

Table 4: Percent accuracy for individual features. "†" indicates a statistically significant difference from the performance of all features. All results are statistically significant in comparison to the baselines.

Features	Accuracy (%)
All features	**78.8**
process-vs-result	76.5
suffix	72.0†
sg_outside_DC	68.9†
sg_inside_DC	68.9†
head_in_DC	66.7†
sum_adjectives	61.4†
of_outside_DC	59.8†
by_outside_DC	56.0†
by_inside_DC	54.5†
process-vs-result and *suffix* combined	78.0
Random baseline	50.0
Head frequency baseline	50.0
Head family size baseline	46.8

With respect to the upper bound, we cannot directly compare the numbers in Table 4 with the IAA reported in Section 3.3.1, because the data we use for testing and training includes only examples on which all annotators agree; neither can we use the 100% IAA on this selected test set as an upper bound. We expect the IAA for this high-agreement test set to lie between 100% and the 81.5% reported in §3.3.1 for the complete dataset and two annotators. The 78.8% we attain is not too far from the upper bound we can estimate from these IAA values.[22]

Furthermore, the results for the individual features in Table 4 show that each feature outperforms the baselines significantly. This means that each feature contributes significantly to the prediction of the relation. The 78.0% performance of the model that combines the top-2 features is comparable to the 78.8% of the model that includes all features. This means that although all features contribute to the quality of the prediction of the model individually, the best features overshadow the effect of the less well-performing features.

[22] A realistic upper bound for the test set could be determined by getting an independent annotator to annotate the items in the test set and measuring the agreement with the previous annotations. We leave this for future work.

Table 5 shows the results from the ablation experiments. Only the removal of features *suffix, of_outside_DC*, and *P-R* result in a significant drop in performance, which means that their contribution in addition to the other features is particularly important. Their performance together is not significantly higher than that of all features (cf. 80.3% vs. 78.8%).

Table 5: Percent accuracy in ablation experiments. "†" indicates a statistically significant difference from the performance of all features.

Features	Accuracy (%)
All features	78.8
All features, except *sg_inside_DC*	**80.3**
All features, except *head_in_DC*	79.5
All features, except *sg_outside_DC*	78.8
All features, except *by_inside_DC*	78.8
All features, except *sum_adjectives*	78.8
All features, except *by_outside_DC*	75.0
All features, except *suffix*	73.5†
All features, except *of_outside_DC*	72.0†
All features, except *P-R*	72.0†
P-R, *of_outside_DC*, *suffix*, *by_outside_DC* combined	**80.3**

For the sake of comparison, Table 6 shows the results of a model using corpus-based features only, i.e., the data does not include the *P-R* feature that is based on human judgments. Like in Table 5, we see that the features *of_outside_DC* and *suffix* are particularly important also in this model, since their absence triggers a significant drop in performance. In this model, however, the contribution of the feature *by_outside_DC* also becomes significant, in contrast to the model in Table 5, which included the *P-R* feature.

Table 7 shows the direction of the prediction of the features in all three models (Tables 4 to 6). In other words, it shows whether higher values of a given feature are indicating higher chances of an OBJ or NOBJ relation. We gathered these directions by inspecting the coefficients of the logistic regression model.[23]

[23]We inspected the weights in the models as well, but they are not very informative, because there is a high level of collinearity in the features and the weights are calculated based on all other features staying equal. For this reason we report results on single feature models and ablation tests instead.

Table 6: Ablation experiment with corpus-based morphosynctic features (no P-R). "†" indicates a statistically significant difference from the performance of all features.

Features	Accuracy (%)
All features	72.0
All features, except *sg_outside_DC*	72.0
All features, except *sum_adjectives*	72.0
All features, except *sg_inside_DC*	72.0
All features, except *by_inside_DC*	72.0
All features, except *head_in_DC*	68.2
All features, except *suffix*	66.7†
All features, except *by_outside_DC*	59.1†
All features, except *of_outside_DC*	54.5†
of_outside_DC, *by_outside_DC*, and *suffix* combined	72.7

Table 7: Direction of prediction per feature in different models. Consistent values across studies in bold

Feature	Table 4	Table 5	Table 6
P-R	**OBJ**	**OBJ**	N/A
suffix=ment	OBJ	OBJ	OBJ
suffix=ance	OBJ	NOBJ	NOBJ
suffix=ion	NOBJ	OBJ	OBJ
suffix=al	OBJ	NOBJ	OBJ
suffix=ing	NOBJ	OBJ	NOBJ
sg_inside_DC	NOBJ	OBJ	NOBJ
by_inside_DC	OBJ	NOBJ	NOBJ
sg_outside_DC	**OBJ**	**OBJ**	**OBJ**
head_in_DC	**OBJ**	**OBJ**	**OBJ**
sum-adjectives	NOBJ	OBJ	OBJ
of_outside_DC	**OBJ**	**OBJ**	**OBJ**
by_outside_DC	NOBJ	NOBJ	NOBJ

6 Discussion

In what follows we offer a detailed discussion of our results and interpret them in view of our initial hypothesis (Section 6.1). We then show their implications for compositionality and for our starting hypothesis (Section 6.2). In the end we present the main comparison points with respect to previous NLP literature (Section 6.3).

6.1 Interpretation of results

6.1.1 *Process-vs-result* (*P-R*)

According to Table 4, the best individual feature is the *process vs. result* reading of the DC with 76.5% accuracy. The accuracy resulting from the combined model with all features (78.8%) is not significantly higher (McNemar two-tailed p-value of 0.2482), showing that this single feature is indeed very strong, and stronger than any of the morphosyntactic features on their own or in combination (cf. Table 6). This is not surprising, given that this feature encodes direct estimates for the ASN-hood of the head based on introspection.[24] In the ablation experiment in Table 5, *P-R* also proves to be very strong, since its removal yields a significantly lower result (72.0% vs. 78.8%), the lowest in this experiment. Still, the ablation study shows that removing *of-outside* is as detrimental to the model as removing *P-R*. This indicates that these two features capture characteristics that complement the rest of the morphosyntactic features to a similar extent.

Importantly, in line with our hypothesis, an increase in the *P-R* value correlates with an OBJ interpretation of the compounds in both experiments (see Table 7). To be precise, the *P-R* feature is so designed that a high value indicates that the DC is headed by an ASN, which parallels the verbal construction in (6). Given that such a compositional structure requires the object to be realized first, the fact that a high *P-R* value correlates with an OBJ reading of the DC in our models confirms our hypothesis that compositional DCs involve object non-heads.

The two columns in Table 8 illustrate pairs of DCs which, despite having the same head, reveal contrasting *P-R* values. In these examples, one can see that whenever the DC pair differs between an OBJ and a NOBJ reading, the OBJ reading receives the higher *P-R* value. This is predicted by our hypothesis and also supported by the results in Table 7. However, we also find examples with two

[24] It is interesting to see though that manual annotation was better at predicting ASN-hood than any of the features, in spite of the huge corpus we used. This suggests that we need even larger corpora to make up for the performance of (expensive) manual annotation.

considerably different *P-R* values under the same OBJ (or NOBJ) interpretation, which shows that there is no one-to-one correspondence between a (high) process reading and an OBJ interpretation of the DC.[25]

Table 8: DC pairs with contrasting *P-R* values

High *P-R* > 60%			Low *P-R* < 60%		
DC	P-R (%)	Reading	DC	P-R (%)	Reading
home building	100	OBJ	police building	20.0	NOBJ
book reading	100	OBJ	temperature reading	40.0	OBJ
ship breaking	93.3	OBJ	record breaking	40.0	OBJ
science teaching	93.3	OBJ	church teaching	46.7	NOBJ
career counseling	93.3	NOBJ	telephone counseling	53.3	NOBJ
slum clearance	80.0	OBJ	safety clearance	20.0	NOBJ
body movement	80.0	OBJ	student movement	33.3	NOBJ
nicotine withdrawal	80.0	NOBJ	summer withdrawal	33.3	NOBJ
refuse disposal	80.0	OBJ	garbage disposal	46.7	OBJ
temperature tolerance	73.3	OBJ	alcohol tolerance	20.0	OBJ
cancer treatment	73.3	OBJ	spa treatment	46.7	NOBJ

The confusion matrix for the feature *P-R* in Table 9 confirms that the machine learning algorithm was not able to find a clear cut-off value for this feature above which we find only OBJ readings. The *P-R* feature misclassifies 18 OBJ-DCs as NOBJ, and 13 NOBJ-DCs as OBJ. Examples of the former case are the OBJ-DCs in the second column of Table 8, which have a low *P-R* value, because they involve RN heads (see *temperature reading*, *alcohol tolerance*). In the latter case, the errors concern the NOBJ-DCs from the first column of Table 8, which have a high *P-R* value (see *career counseling* and *nicotine withdrawal*).

Table 9: Confusion matrix for *P-R*

		Classified as		
		OBJ	NOBJ	Totals
Gold	OBJ	48	18	66
	NOBJ	13	53	66
	Totals	61	71	132

[25] NOBJ-DCs with a high *P-R* value are usually headed by simple event nominals like the nouns in (11c, d).

In our study, the *P-R* annotation feature comes closest to the transparency rating of compounds carried out in some NLP studies (cf. Section 2.2). The difference is that we correlated the rating with the semantics of the base verb in combination with its argument or adjunct, following Grimshaw's (1990) insight. At the same time, our design primarily targeted compositionality.

6.1.2 *of_outside_DC*

The next most important feature in our endeavor to capture compositionality in DCs is the realization of an *of*-phrase by the deverbal noun. This feature is intended to measure how often the deverbal noun realizes an *of*-phrase introducing the object argument, when appearing outside DCs. If the head noun of a DC shows a high tendency to realize *of*-phrases introducing objects, we expect it to also require object non-heads in DCs.

Although on its own the feature *of_outside_DC* yields a value of only 59.8% (see Table 4, insignificantly lower than the next higher value of 61.4%), the ablation study in Table 5 shows that its removal is just as detrimental for the system as the removal of the *P-R* feature: The accuracy drops from 78.8% to 72.0%. Similarly, in the model with corpus-based morphosyntactic features in Table 6, its removal triggers the largest drop, showing that in combination with the other features, the contribution of *of_outside_DC* is very important. This confirms Grimshaw's claim that the realization of the object argument is essential in identifying ASNs. Even more important for our hypothesis is the fact that *of_outside_DC* systematically correlates with an OBJ-DC in all our models (see Table 7). That is, to the extent that this feature identifies DCs with ASN heads, a high value indicates an object reading for the DC, as expected under our hypothesis.

The question is why the *of_outside_DC* feature does not score better than 59.8% on its own. First, as shown in Section 2.1.1, the presence of an *of*-phrase per se, as extracted from the corpus, is no guarantee for ASN-hood, since *of*-phrases may introduce possessive modifiers of RNs, besides the object arguments of ASNs. Second, even in their ASN reading, deverbal nouns attested in corpora do not always realize their object arguments (cf. Grimm & McNally 2013).

The samples in Table 10 show various mismatches between the realization of *of*-phrases and the formation of OBJ-DCs. For instance, *avoidance* and *preservation*, which build only OBJ-DCs in our database, have fewer occurrences with an *of*-phrase than *creation*, which forms only 72.7% OBJ-DCs. Moreover, *proposal*, which forms a high proportion of OBJ-DCs, realizes *of*-phrases in only 1.0% of its occurrences. In spite of the many OBJ-DCs like *book/contract/marriage/investment proposal*, the verbal relation is lost in this noun. It mostly functions

Table 10: Head nouns with (in)frequent *of*-phrases. Outliers in bold.

Head noun	*of_outside_DC* (%)	OBJ-reading (%)
creation	80.5	72.7
avoidance	70.4	100
obstruction	65.3	90.5
assassination	52.3	**11.8**
preservation	52.1	100
proposal	1.0	**76.2**
counseling	0.5	10.0
mongering	0	**100**

as an RN, i.e., it refers to the proposal made, and not to the process/event of proposing. In confirmation of this, these DCs received a *P-R* rating as low as 20% to 26.7%. This is an example of how our individual features complement each other.

The confusion matrix for the feature *of_outside_DC* in Table 11 shows indeed that the model based on this feature makes many false predictions, notably, it attributes 38 OBJ readings to DCs that in fact have a NOBJ reading. This means that the prediction power of *of_outside_DC* is misled by the presence of *of*-phrases with head nouns that form NOBJ-DCs (see Table 10). These DCs involve RN heads, which realize *of*-phrases as modifiers and not object arguments. The head noun *assassination* in Table 10 is one example. That this noun behaves like an RN is confirmed by the *P-R* value of the DCs it forms, which is below the average of 60%. A similar problem is posed by the DCs headed by, e.g., *creation*, which also allows RN readings and forms NOBJ-DCs, in spite of the high frequency with *of*-phrases (Table 10). In these critical cases, the results in Tables 5 and 6 show that the other morphosyntactic features compensate for the errors made by the *of_outside_DC* feature, helping the model.

Table 11: Confusion matrix for *of_outside_DC*

		Classified as		
		OBJ	NOBJ	Totals
Gold	OBJ	51	15	66
	NOBJ	38	28	66
	Totals	89	43	132

All in all, when comparing *of_outside_DC* with *P-R* in the ablation study, their contribution in combination with the other corpus features is similar. The difference is that the other features negatively affect the 76.5% individual contribution of *P-R* (cf. 72%), while they substantially improve the 59.8% contribution of *of_outside_DC* (cf. Table 4). Thus, the contribution of *of_outside_DC* greatly relies on the other ASN-features in the ablation models in Tables 5 and 6. This is not surprising, given the ambiguity of *of*-phrases, a reason for which Grimshaw (1990) used this test in combination with others (see Section 2.1.1). The contrast between *P-R* and *of_outside_DC* is also expected, since *P-R* is manually annotated and targets the underlying ASN-hood of the deverbal noun; the corpus features can only capture some aspects of it.

6.1.3 *Suffix*

Suffix is an important feature in all our models (see Tables 4, 5, and 6). It is the strongest morphosyntactic feature, as we can see from the performance of the individual features in Table 4, and has additional predictive power compared to the combination of all features (see Tables 5 and 6). However, Table 7 demonstrates a high variance in the direction of prediction of each suffix. Except for *-ment*, which correlates with OBJ readings, none of them is constant across models.

As noted in Section 4.1, the theoretical literature does not offer much on the role of suffixes in the ASN vs. RN disambiguation of deverbal nouns. Grimshaw (1990) and Borer (2013) suggest that *-ing* should form ASNs, which is disconfirmed by some data and by our models, where *-ing* oscillates between OBJ- and NOBJ-DCs. It is difficult to draw any conclusions on the role of the *suffix* feature for our compositionality hypothesis for two reasons. First, more theoretical research must be pursued to draw some definite conclusions on possible correlations between suffixes and ASN-hood, since the one suffix that was expected to show a preference did not. Second, we must also consider that the dataset of DCs for each suffix was five times smaller than for the other features in our study: i.e., the feature *suffix* subsumes five different suffix features. The small dataset may also be a reason for the inconclusiveness of the results in Table 7.[26]

The high variation between OBJ and NOBJ readings in Table 7 indicates that the valuable contribution of the *suffix* feature in the prediction task (72.0% in Table 4) comes from the complementarity between the individual suffixes. Similarly, in the ablation models in Tables 5 and 6, the contribution of the suffixes – which,

[26]To check correlations between individual suffixes and ASN-hood, one could measure how the *suffix* feature fares with respect to the *P-R* value and not the OBJ-NOBJ readings of DCs. This, however, would digress from the focus of this paper and we leave it for future research.

recall, is independent of Grimshaw's tests – is complementary to the features that diagnose ASN-hood. Thus, the *suffix* feature is not informative about the relation between compositionality and interpretation in DCs, but improves the predictive power of the models.

6.1.4 *sg_outside_DC* and *sg_inside_DC*

The frequency of the noun head in a singular form whether outside or inside a DC yields similar accuracy levels (68.9% in Table 4, 78.8% and 80.3% without a significant difference in Table 5, and 72% in Table 6). This similarity supports our assumption that within DCs the head nouns should preserve the properties from outside DCs (see Section 4.1). However, an interesting difference appears with respect to the direction of prediction, since only *sg_outside_DC* constantly predicts OBJ-DCs across all the models in Table 7, while *sg_outside_DC* is less reliable. This suggests that Grimshaw's morphosyntactic ASN-properties may be more reliable when the deverbal noun appears outside a DC than inside DCs.[27]

6.1.5 *head_in_DC* (compoundhood)

As an individual feature, the accuracy of *head_in_DC* is just above average among the other features in the present study (see Table 4). Its removal in our ablation experiments yields slight and non-significant drops in accuracy. In Section 4.1, we conjectured that an OBJ reading of DCs whose head nouns present high compoundhood would show us that a compositional construction with an object non-head is very likely to form DCs. The direction of prediction in Table 7 indicates that high values of this feature consistently correlate with OBJ-DCs, supporting this assumption. However, why does this feature not perform better? Our full database shows that its values are not informative enough: there are a few head nouns which display high compoundhood and frequently form OBJ-DCs, but the majority of DCs have very low such values. Only 5.1% of our DCs have a *head_in_DC* value above 50% and as many as 70.3% of them have one under 20%.

Table 12 illustrates the few head nouns that most often appear in DCs and the frequency of an OBJ reading among the DCs they appear as heads of. As visible there, a high frequency of a deverbal noun in DCs correlates with a high value for an OBJ reading of the compound's non-head, as predicted (cf. Section 4.1).

[27] The *inside* features do not damage our model, since removing *sg_inside_DC* and *by_inside_DC* from the ablation model yielded 77.3% accuracy – lower than 78.8% for all features together, though not significantly so.

Table 12: Head nouns with high compoundhood

Head noun	*head_in_DC* (%)	OBJ-reading (%)
laundering	94.8	95.5
mongering	91.8	100
growing	68.7	95.2
trafficking	62.0	100
enforcement	53.7	66.6

6.1.6 *sum-adjectives* and *by*-phrases

The last three features we employed in our study are *sum-adjectives*, *by_outside_DC* and *by_inside_DC*. On their own, they have some predictive power (Table 4), but their removal in Table 5 has no significant impact on the results, showing that *P-R* compensates for their absence. Interestingly, in the corpus-based morphosyntactic model in Table 6, the removal of *by_outside_DC* triggers a significant drop, indicating that in the absence of *P-R*, this feature becomes important. Yet, in spite of our expectation for this feature to identify OBJ-DCs, its direction of prediction is NOBJ in all models (see Table 7). As we saw in Section 2.1.1, *by*-phrases are ambiguous and their presence indicates ASN-hood only when the object argument is also realized (see 10). We considered using the frequency of *by*-phrases co-occurring with *of*-phrases, but the numbers were extremely low. Thus, the unexpected direction of prediction of *by*-phrases might be due to their ambiguity. The other two features do not preserve the direction of prediction (Table 7).

The inconclusiveness of these three features most likely resides in data sparsity. Namely, for the feature *by_outside_DC* the range of frequency in our full database is 0–6.22% with 60% of the deverbal head nouns realizing a *by*-phrase in fewer than 1% of their occurrences outside DCs. For *by_inside_DC* the range is between 0% and 4.36%, with 74% of the DCs displaying a *by*-phrase in fewer than 1% of the cases. For *sum-adjectives* the value is even lower: the frequency ranges between 0% and 1.8%, with 99% of the cases having a value under 1%.

6.1.7 Summary

In summary, *P-R*, the feature based on introspection, is the strongest. It provides a high performance individually and its removal from the model considerably

hurts the results. *Suffix* is the strongest morphosyntactic feature. It brings additional value over the combination of all features including *P-R*, but it does not reach the performance of *P-R* on its own. *Of-outside* is the next valuable feature. On its own, it is not very strong, but it is a very important addition to the other features. Its removal from the combined models hurts the performance considerably. The feature *by-outside* is valuable when only corpus-based morphosyntactic features are considered. If *P-R* is present in the model, *by-outside* is unimportant. This indicates that this feature has a considerable overlap with *P-R*. The other features all have predictive power, but their additional predictive power is not very important. They capture the same signal in a less reliable way.

The latent variable that we are trying to capture with the features presented in this study, the ASN-hood of the head, is best represented by the introspection-based feature *P-R*. The morphosyntactic features *suffix* and *of-outside* have additional value in the combined model, which includes P-R, as the ablation studies show. They seem to help the strong feature P-R to move the model in the right direction. However, although the combination of P-R and the best morphosyntactic features leads to an improvement (80.3% vs. 78.8%), we could not prove that their addition to *P-R* as a single feature model improves the results significantly.

6.2 Implications for our hypothesis

We have identified four features which are important for the interpretation of DCs: *P-R*, *of_outside_DCs*, *by_outside_DC*, and *suffix*. The first three were inspired by Grimshaw (1990) and later research in the same vein, the fourth was introduced by us. As mentioned in Section 6.1.3, the suffix does not tell us anything about ASN-hood or the compositionality of the DC. It is a morphological feature, which scores well on its own and better than most ASN-features from Grimshaw (1990); yet, in ablation studies, it is weaker than *of_outside_DCs*, which is Grimshaw's most important ASN-feature.

The other three features all give us input on ASN-hood, but in different ways. An unexpected result comes from *by_outside_DC*, whose direction of prediction is for NOBJ-DCs, instead of OBJ-DCs. In Section 6.1.6, we reasoned that this is due to the ambiguity of *by*-phrases, which we could not eliminate by measuring their co-occurrence with *of*-phrases, given data sparsity. The only way we can interpret this result is that, in combination with other ASN-features which usually point to OBJ-DCs, the input from the ambiguous *by*-phrases was used by the model for the other direction, of NOBJ-DCs.

The features *P-R* and *of_outside_DCs* are the most important for the ASN-hood of head nouns and the implicit compositional interpretation of DCs. They both

behave as predicted by our hypothesis. *P-R* represents human intuitions with respect to the ASN-hood of the head noun and scores best in our models. In addition, in line with Grimshaw's claims and our hypothesis, its direction of prediction consistently points to OBJ-DCs. *Of_outside_DCs* is not very strong on its own, but extremely important in combination with the other ASN-features. This is in fact what Grimshaw's combined use of two or three of these morphosyntactic tests (in order to circumvent ambiguity) leads us to expect (see Section 2.1.1).

These results immediately confirm two things:

1. the validity of these features as identifying ASN-hood and correlated OBJ readings in DCs (i.e., Grimshaw's theory, which is also part of our hypothesis in Section 1.5);

2. DCs are compositional and easily interpretable to the extent that their head nouns exhibit ASN-properties (i.e., the starting point of our hypothesis in Section 1).

A further implication of these observations is that, indeed, the (deverbal) head noun plays a crucial role in the compositionality and overall transparency of DCs, a conclusion that was reached by other computational studies as well (see Section 2.2.2).

For the DCs whose heads fail to exhibit ASN-properties and behave like RNs, our features cannot get very far. These DCs behave like RCs, and the relation between their two parts may even be unrelated to the base verb and its modifiers. For these DCs, the addition of other features, especially some designed for non-heads, should improve the results. In this case, it would be worth including features from previous NLP work, which deals with noun-noun compounds in general, especially that reported in Section 2.2.2. We leave such a study for future research, since it departs from our focus here.

6.3 Comparison to other NLP approaches

We mentioned in Section 2.2.1 that the aims of previous work on predicting the relation between heads and non-heads in DCs are different from ours. Whereas this work focuses on building classifiers that reach state-of-the-art performance on the task of predicting the relation between the head and the non-head of deverbal compounds, our interest lies in uncovering in how far the behavior of the derived nominals (as ASNs or RNs) can help in predicting the (compositional) relation between head and non-head. As a result, the datasets are very different.

However, we present here some meaningful comparisons with this work. In the two-class prediction task, Lapata (2002) reaches an accuracy of 86.1% compared to a baseline of 61.5%, i.e, 24.6% above the baseline. The accuracy we achieve is 80.3%, i.e., 30.3% above the 50% baseline of our balanced test set. Relative improvements are comparable. Note that the data set of Lapata (2002) included DCs ending in suffixes such as *-er* and *-ee* which are biased in the relation they select. Including them in our dataset could have resulted in better accuracy overall and a stronger predictive power for the *suffix* feature.

Apart from the differences in the data set, we also see large differences in the type of features selected. In this paper we exclusively tested the predictive power of morphosyntactic features of the deverbal noun for determining the covert relation. In the future, it would be interesting to compare these to the encyclopedic/pragmatic features prevalent in the CL literature, by incorporating the latter into our models.

Schulte im Walde, Hätty & Bott (2016) evaluate the influence of several properties of the constituents (frequency, productivity and ambiguity) on the performance of the model in its predictions on transparency. Just as they attribute the influence of these properties to the underlying property of ambiguity, so do we attribute the non-compositionality in the relation between head and compounds (in RNs) to the greater underspecification of RNs in comparison to ASNs. Although we do not have access to transparency ratings for our DCs, we have gathered annotations on their process vs. result interpretation (see Section 3.3.2). This information can be seen as a proxy for the transparency of the head, because by default the more result-like the DC is, the less transparent it will be.

Furthermore, Schulte im Walde, Hätty & Bott (2016) emphasize the importance of properties of the head and the compound, and to a lesser extent of the modifier (i.e., non-head) for the prediction of the transparency of the compound. The authors stress the need to carefully balance datasets according to the empirical and semantic properties of the compounds, as well as of their heads. We have balanced our data set for corpus frequency of the head and measured the family size of the heads. We have not measured other properties that they have used, but will consider these in future work.

7 Conclusions

In this paper we have presented a study on the (syntactic) compositionality of DCs, as predictable from the morphosyntactic properties of their head nouns. We have employed theoretical insights on the behavior of deverbal nominals, on the basis of which we collected corpus data, as well as manual annotations. We used this data collection in the form of indicative features in a logistic regression model, by means of which we evaluated the prediction power of each feature for the OBJ (vs. NOBJ) interpretation of the compounds.

Our approach to compositionality comes from the theoretical linguistic perspective according to which the compositionality of a complex expression (here, the DC) depends on the meanings of its parts, as well as the syntactic relationship between them. To the extent that DCs are headed by deverbal nouns, the fully compositional ones encode the syntactic-semantic relationship between the base verb and its object, while the less compositional ones are underspecified/ambiguous. This difference is traced back to the ambiguity of deverbal nouns between ASN and RN uses from Grimshaw (1990). ASNs preserve the compositional requirements of the base verb, while RNs do not.

Our results confirm our hypothesis that DCs with ASN-heads are compositional and receive an OBJ reading. This study, however, raises a few questions for future research. It especially highlights the need for more study on the role of individual suffixes in the interpretation of the deverbal noun, since previous claims on *-ing* as primarily building OBJ-DCs have not been confirmed. In addition, some tests which are popular in the theoretical literature (e.g., *in/for*-adverbials, agentive and aspectual adjectives, as well as *by*-phrases) could not be used or were not reliable enough as features, probably due to data sparsity. On the one hand, their low attestation in corpora throws doubts on their authenticity, requiring further empirical study. On the other hand, this is also an alarm signal for the need of even larger corpora in order to reliably test theoretical insights, which human intuitions are considerably better at, as proven by our *P-R* feature.

By comparison to the previous NLP work on the transparency of (root) compounds, we did not consider both constituents to evaluate the mapping with the compound; we focused on the head noun, which has a crucial influence on the relationship that it establishes with the non-head in DCs. In future work, we will consider including some predictive features of the non-head. We expect that the encyclopedic features exploited in the NLP literature such as in Nicholson & Baldwin (2006), Lapata (2002), and Grover et al. (2005) will benefit the disambiguation of RNs and the DCs headed by these.

Abbreviations

ASN	argument structure nominal	*P-R*	*process-vs-result* feature
CL	computational linguistics	PoC	principle of compositionality
DC	deverbal compound	POS	part of speech
DS	distributional semantics	RC	root compound
IAA	inter-annotator agreement	RN	result nominal
NLP	natural language processing	TL	theoretical linguistics

Acknowledgements

We are grateful to Katherine Fraser, Bethany Lochbihler and Whitney Frazier Peterson for annotating our database, to Kerstin Eckart and the INF project in the SFB 732 for important technical support, and to Alla Abrosimova for help with further technical details. This research has been funded by the German Research Foundation (DFG) via grants offered to the projects B1 *The form and interpretation of derived nominals* and D11 *A crosslingual approach to the analysis of compound nouns*, as part of the SFB 732 *Incremental specification in context*, as well as the project IO 91/1-1, all hosted at the University of Stuttgart.

References

Ackema, Peter & Ad Neeleman. 2004. *Beyond morphology*. Oxford: Oxford University Press.

Alexiadou, Artemis & Jane Grimshaw. 2008. Verbs, nouns, and affixation. In Florian Schäfer (ed.), *Working Papers of the SFB 732 Incremental Specification in Context*, vol. 1, 1–16. Universität Stuttgart.

Alexiadou, Artemis, Gianina Iordăchioaia & Elena Soare. 2010. Number/aspect interactions in the syntax of nominalizations. *Journal of Linguistics* 46:3. 537–574.

Baroni, Marco & Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1183–1193. Boston.

Borer, Hagit. 2013. *Taking form*. Oxford: Oxford University Press.

Chomsky, Noam. 1970. Remarks on nominalization. In Roderick A. Jacobs & Peter S. Rosenbaum (eds.), *Readings in English transformational grammar*, 184–221. Waltham, MA: Ginn.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge, MA: MIT Press.

de Marneffe, Marie-Catherine & Christopher D. Manning. 2008. *Stanford typed dependencies manual*. Tech. rep. Stanford University.

Di Sciullo, Anna Maria. 1992. Deverbal compounds and the external argument. In Iggy M. Roca (ed.), *Thematic structure: Its role in grammar*, 65–78. Berlin: Foris.

Dietterich, Thomas G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10(7). 1895–1923.

Dowty, David. 2007. Compositionality as an empirical problem. In Chris Barker & Pauline I. Jacobson (eds.), *Direct compositionality*, 14–23. Oxford: Oxford University Press.

Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fokkens, Antske Sibelle. 2007. *A hybrid approach to compound noun disambiguation*. Universität des Saarladens, Saarbrücken. (MA thesis).

Gillick, Dan. 2009. Sentence boundary detection and the problem with the U. S. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 241–244. Boulder, Colorado.

Grimm, Scott & Louise McNally. 2013. No ordered arguments needed for nouns. In Maria Aloni, Michael Franke & Floris Roelofsen (eds.), *Proceedings of the 19th Amsterdam Colloquium*, 123–130. Institute for Logic, Language & Computation, University of Amsterdam.

Grimshaw, Jane. 1990. *Argument structure*. Cambridge, MA: MIT Press.

Grover, Claire, Mirella Lapata & Alex Lascarides. 2005. A comparison of parsing technologies for the biomedical domain. *Journal of Natural Language Engineering* 11(1). 27–65.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann & Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11(1). 10–18.

Hamp, Birgit & Helmut Feldweg. 1997. GermaNet - A lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 9–15. Madrid, Spain.

Henrich, Verena & Erhard Hinrichs. 2010. GernEdiT - The GermaNet editing tool. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta, Malta.

Huang, Zhongqiang, Mary Harper & Slav Petrov. 2010. Self-training with products of latent variable grammars. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 12–22. Cambridge, Massachusetts.

Iordăchioaia, Gianina. to appear. English deverbal compounds with and without arguments. In *Proceedings of the 54th Annual Meeting of the Chicago Linguistic Society*. Chicago.

Iordăchioaia, Gianina, Artemis Alexiadou & Andreas Pairamidis. 2017. Morphosyntactic sources for nominal synthetic compounds in English and Greek. *Zeitschrift für Wortbildung/Journal of Word-formation* 1. 47–72.

Iordăchioaia, Gianina, Lonneke van der Plas & Glorianna Jagfeld. 2016. The grammar of English deverbal compounds and their meaning. In Eva Hajicova & Igor Boguslavsky (eds.), *Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces*, 81–91. Osaka, Japan.

Juhasz, Barbara J., Yun-Hsuan Lai & Michelle L. Woodcock. 2015. A database of 629 English compound words: Ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods* 47. 1004–1019.

Kratzer, Angelika. 1996. Severing the external argument from its verb. In Johan Rooryck & Laurie Zaring (eds.), *Phrase structure and the lexicon*, 109–137. Dordrecht: Kluwer Academic Publishers.

Lapata, Maria. 2002. The disambiguation of nominalizations. *Computational Linguistics* 28(3). 357–388.

Larson, Richard K. 1988. On the double object construction. *Linguistic Inquiry* 19:3. 335–391.

Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

Libben, Gary, Martha Gibson, Yeo Bom Yoon & Dominiek Sandra. 1997. *Semantic transparency and compound fracture*. Tech. rep. 9. CLASNET Working Papers.

Libben, Gary, Martha Gibson, Yeo Bom Yoon & Dominiek Sandra. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language* 84(1). 50–64.

Lieber, Rochelle. 2004. *Morphology and lexical semantics*. Cambridge: Cambridge University Press.

Lieber, Rochelle. 2016. *English nouns. The ecology of nominalizations*. Cambridge: Cambridge University Press.

Macleod, Catherine, Ralph Grishman, Adam Meyers, Leslie Barrett & Ruth Reeves. 1998. NOMLEX: A lexicon of nominalizations. In *Proceedings of EURALEX*.

Marelli, Marco & Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122(3). 485–515.

Mitchell, Jeff & Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34(8). 1388–1429.

Montague, Richard. 1970. Universal grammar. *Theoria* 36(3). 373–398.

Napoles, Courtney, Matthew Gormley & Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, 95–100. Montreal, Canada.

Nicholson, Jeremy & Timothy Baldwin. 2006. Interpretation of compound nominalisations using corpus and web statistics. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 54–61. Montreal, Canada.

Ó Séaghdha, Diarmuid. 2008. *Learning compound noun semantics*. Tech. rep. UCAM-CL-TR-735. University of Cambridge, Computer Laboratory.

Partee, Barbara H. 1984. Compositionality. In Fred Landman & Frank Veltman (eds.), *Varieties of formal semantics: Proceedings of the 4th Amsterdam Colloquium*, 281–311. Dordrecht: Foris Publications.

Reddy, Siva, Diana McCarthy & Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 210–218. Chiang Mai, Thailand.

Roeper, Thomas & Muffy Siegel. 1978. A lexical transformation for verbal compounds. *Linguistic Inquiry* 9. 199–260.

Santorini, Beatrice. 1990. *Part-Of-Speech tagging guidelines for the Penn Treebank project (3rd revision, 2nd printing)*. Tech. rep. Philadelphia, PA, USA: Department of Linguistics, University of Pennsylvania.

Schulte im Walde, Sabine, Anna Hätty & Stefan Bott. 2016. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, 148–158. Berlin, Germany.

Schulte im Walde, Sabine, Anna Hätty, Stefan Bott & Nana Khvtisavrishvili. 2016. $G_h$ost-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 2285–2292. Portoroz, Slovenia.

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

van Heuven, Walter J. B., Pawel Mandera, Emmanuel Keuleers & Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *Journal of Experimental Psychology* 67(6). 1176–1190.

Zwitserlood, Pienie. 1994. The role of semantic transparency in the processing and representation of Dutch compounds. *Language and Cognitive Processes* 9(3). 341–368.