Crosslinguistic Annotation of German and English Shell Noun Complexes

Fabian Simonjetz
Department of Linguistics
Ruhr-University Bochum
simonjetz@linguistics.rub.de

Adam Roussel
Department of Linguistics
Ruhr-University Bochum
roussel@linguistics.rub.de

Abstract

This contribution involves the manual annotation of shell nouns and their antecedents in a multilingual context. Shell nouns are abstract nouns, which like pronouns are semantically incomplete and derive their meanings from other parts of a text to which they refer, often anaphorically. Unlike pronouns, shell nouns also serve to characterize the content to which they refer. The annotation schema we introduce allows for the annotation of shell nouns along with their content and their translation in a parallel text. This approach should enable the production of data on shell nouns which encompasses various aspects of their behavior that have not yet been investigated in detail, including the use of multiple content phrases, nominalized content phrases, plural shell nouns or crosslinguistic behavior.

1 Introduction

Shell nouns are an open class of abstract nouns that refer to stretches of text, which complete the shell noun's semantic content and are simultaneously characterized by the shell nouns. Their name derives from the way they are said to *encapsulate* the content to which they refer (Schmid, 2000). Some typical examples are listed in (1). The shell nouns are printed in boldface and the content to which they refer is in italics.

(1) a. The **problem** was *that I had no money*. (Schmid, 2000)

b. *Pigs cannot fly.* That **fact** is well-known.

c. der **Plan** ist, *ein Auto zu kaufen* ('the plan is to buy a car')

Combining functions usually associated with pronouns (reference) and adjectives (characterization), shell nouns are a useful device for facilitating textual coherence. Yet they have received little attention so far, especially in languages other than English. The primary goal of this paper is to offer a set of annotated cross-linguistic data to serve as a basis for further exploration of shell nouns in English and German.

Previous studies on shell nouns have primarily focussed on the use of lexical patterns, such as in (2), for the discovery and analysis of shell nouns and their content phrases. Since previous work only looked at English shell nouns, this approach was more or less sufficient: though certain phenomena are systematically missed, it is thought that the bulk of relevant cases can be covered in this way, thanks to English's relatively fixed word order.

(2) Determiner + (Premodifier) + Noun + postnominal *that*-clause, *wh*-clause, or *to*-infinitive
The (deplorable) fact that I have no money (Schmid, 2000)

For this annotation task, we wanted to take advantage of the fact that we were conducting the annotation manually and annotate shell nouns in ways that are not amenable to automatic methods. Further, in order to gather data about the types of patterns in which shell nouns actually occur, we could not use Schmid's patterns for identifying shell noun instances. We also wanted to annotate in such a way so as to facilitate the crosslinguistic study of shell noun use. We were thus led to formulate three main criteria to guide our approach to shell nouns.

Incompleteness A shell noun (when used as such) is *incomplete* with regard to its semantic content. For example, a *fact* denotes a true state-of-affairs, whereas this same state-of-affairs might be cast as a *problem*, some undesirable situation

in want of a solution. An *aim* is something to be achieved in the future, a desirable situation, which has not yet come to pass. What exactly these various 'situations' or 'states-of-affairs' entail is only found in the co-text of the nouns, if it is made explicit at all. Unlike concrete nouns, shell nouns seem to possess a 'gap' or 'placeholder' for this additional information (Schmid, 2000, p. 79).

Reference A shell noun *refers* to linguistic content elsewhere in the discourse. This content could usually also occur without the shell noun itself, but the shell noun serves to describe or characterize this content and encapsulates it, allowing easier subsequent reference to it. Once a state-of-affairs has been summed up as a *problem*, a speaker can then go on to discuss this *problem* as they might do with some concrete entity. Reference can be achieved by a variety of means, for instance, with a copula verb linking the shell noun and its content (1a) or via anaphoric constructions (1b).

Abstractness The shell noun content must be *abstract*, in that it, for example, denotes entities which correspond semantically to the meanings of sentences, such as facts, states-of-affairs or propositions, i.e. *saturated* abstract objects or entities with truth values.[1]

2 Related Work

Schmid (2000) is the most extensive and detailed treatment of the topic of *shell nouns*, and it thus forms the basis of most later work on the topic. In this book, he addresses a whole range of aspects relating to shell nouns, including cognitive aspects, discourse functions, semantic categories, etc. This work is based on an extensive corpus-based study of shell noun instances. Shell noun instances are identified here primarily on the basis of lexical patterns, an approach which does not cover certain aspects, such as plurality and anaphoric shell noun complexes, but which is largely sufficient for English shell nouns, which are the sole focus of the book.

Shell nouns are in certain respects essentially a special case of *abstract anaphors*. Like abstract anaphors, they refer not to concrete entities, generally represented by NPs, but rather to propositions or proposition-like entities. The most obvious difference is that shell nouns are themselves full NPs as opposed to abstract anaphors in general, which are often pronouns, such as *this* or *it* in English and *dies* or *es* in German.

In contrast to abstract anaphors, shell nouns do not necessarily refer anaphorically to their content. Far more frequently, the content to which they refer is found in a *that*-phrase complement immediately adjoined to the shell noun. However, the similarity between the two constructions nevertheless means that annotation tasks relating to one involve techniques which are generally applicable to the other.

Dipper and Zinsmeister (2009) present guidelines and a pilot study for an annotation task similar to our own, though this task addresses abstract anaphors as opposed to shell nouns as such. Annotators were asked to identify antecedents of 48 instances of *dies* 'this' by freely marking spans of text. These guidelines introduce the 'paraphrase test',[2] for identifying anaphoric content phrases, which we also use in our guidelines. As the authors note, this appears also to have been the first study to approach abstract anaphors in German. Interestingly, the guidelines also recommend the use of shell nouns, such as *Ereignis* 'event', *Ansicht* 'view' or *Tatsache* 'fact', in order to identify the semantic type of abstract anaphors (the 'replacement test'). In a later study, Dipper and Zinsmeister (2012) expanded this approach, annotating 643 instances and investigating correlations between the abstractness of referents and antecedents. Dipper et al. (2011) use a cross-linguistic bootstrapping approach in order to expand the set of abstract anaphors under comparison and undertake an extensive contrastive study of their realization in German vs. English. All three of these studies also employ the Europarl Corpus, focusing on English and German parallel data.

Kolhatkar (together with Zinsmeister and Hirst) has approached the topic of shell nouns in a series of publications in an explicitly computational context, with the ultimate goal of resolving shell noun instances to their content automatically. The first of these, Kolhatkar and Hirst (2012), involves an annotation task quite similar to our own: annotators were asked to mark arbitrary spans of text corresponding to the content phrases of 183 instances of the shell noun *issue*. The authors then describe an automatic resolution algorithm developed using this data. Later work (Kolhatkar et al., 2013; Kolhatkar and Hirst, 2014) expanded the annotation to other shell nouns, increased the amount of data

[1] See Asher (1993) for more on the relevant typology.

[2] Later 'namely test' (Dipper and Zinsmeister, 2012).

annotated via crowdsourcing, and improved the resolution algorithm. Kolhatkar (2015) describes, in addition to these studies, extensive work concerning the annotation and automatic resolution of shell nouns in general. Relating as it does to these topics, the annotation guidelines which appear there are of direct relevance to the current study.

Simonjetz (2015) argues that Schmid's (2000) procedure of retrieving shell nouns may not be suitable for languages with a more flexible word order than English. It recommends the use of dependency-based syntactic patterns instead of simple string-based patterns in order to identify German shell nouns. With no evaluation data being available, Simonjetz (2015) relied on a manual examination of the results, making it impossible to reach a clear conclusion as to whether or not dependency patterns are superior to linear patterns for the task of identifying shell nouns in German. The resulting data on German shell nouns proved nonetheless useful for our study and has been the basis for the selection of shell noun candidates described in the next section.

3 The Annotation Process

Our project involves the annotation of parts of the Europarl Corpus (Koehn, 2005) with information pertaining to the usage of shell nouns in the text. The data of the Europarl Corpus is divided into plenary sessions, speaker turns and sentences. In order to achieve a degree of homogeneity in the data we filtered out long turns for our annotation project, which are likely to be recitations of written documents. On the other hand we also filtered out very short turns, which are unlikely to contain any shell noun occurrences and, if they do, their content is likely to be located in another turn, thus complicating the annotation. The thresholds for filtering were based on the distribution of text lengths in the corpus, in which three prominent bumps were visible, which we presumed to correspond to each of the types of text.

A complex annotation project such as this makes special demands on its annotators, requiring not only time and patience, but also special expertise. Often, in order to present a task to naive annotators, the task must be simplified and restricted, however this simplification requires certain assumptions to be made by researchers as to the nature of the phenomenon under investigation. Wanting to make as few such assumptions as possible and to encompass the whole spectrum of shell noun phenomena, we performed the annotations ourselves. Thus the data were annotated by two linguistically-informed annotators, one native speaker of English and one native speaker of German, both fluent in their respective non-native languages. This arrangement put us in a position to produce data that, though not without its own shortcomings, would have been prohibitively difficult to produce otherwise.

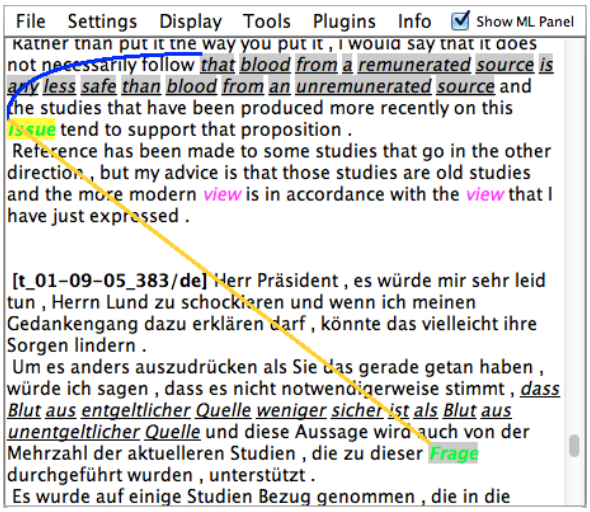

Figure 1: Screenshot of the annotation software as used in this study.

In order to assist the annotators in annotating as many instances as possible while keeping the data sets comparable across different annotators, shell noun candidates were determined beforehand and highlighted in the annotation software MMAX2 (Müller and Strube, 2006). Only highlighted instances were annotated to ease comparisons between annotators and alignment between languages. In case some shell noun candidate was highlighted in one language and its translation was not, the annotators added the translation to the annotation set.

The annotators first annotated instances of *Möglichkeit* and *possibility* using a preliminary version of the guidelines in Section 4. The data from this practice annotation are not included in the final data set. Afterwards the annotators convened and further developed the guidelines on the basis of this experience. We split the data to be annotated into three parts. For the first part, difficult cases were discussed. The last two parts were annotated completely independently, and accordingly only these two parts were used in the calculation of inter-annotator agreement measures.

Using the statistics from Schmid (2000) for English and Simonjetz (2015) for German, we chose 50 shell nouns for each language which have (a) a high ratio of shell noun vs. non–shell noun uses and (b) a high absolute frequency. Both of these factors are important since, if the nouns are too infrequent, then there may not be very much data to annotate (and it would be questionable how generalizable the data would be), and if the chosen nouns are used as shell nouns too infrequently, then it would be difficult to investigate the shell nouns' relationship to their content, since most of the instances would be negative in that case. (Note that these 50 German/English shell nouns are not necessarily translations of each other. Consequently, the shell nouns appearing in the final data slightly deviate from the original selection, as translations are added to the annotation set.)

Turns in both languages are presented to the annotators in parallel and three main levels are annotated:

- "Shell noun" – Annotators mark whether or not the given noun constitutes a usage as a shell noun. Options to mark unclear instances or instances whose content is just outside of the given turn are provided.
- "Content phrase" – Annotators may freely select spans of tokens which comprise the content to which a shell noun instance refers. This content may occur before or after its shell noun instance and may encompass multiple sentences. Each shell noun instance contains a pointer to its content phrase(s), of which there may be more than one. (Likewise, multiple shell nouns may point to the same content phrase.) Content phrases are also marked as being either 'nominal' or 'sentential'.
- "Alignments" – Annotators are then asked to associate corresponding shell noun instances in one language with their translations in the other language (insofar as a counterpart is present). The same is done for content phrases.

We hope that our manual annotation approach will mean that our data cover a greater variety of shell noun–related issues than would be possible with pattern-based methods, which could systemically miss particular unexpected properties and behaviors. For instance, patterns which do not allow for pluralized shell nouns will necessarily preclude any study of their properties as opposed to singular shell nouns, e.g. whether or not plurals tend to refer to multiple content phrases. Furthermore, pattern-based approaches are likely to be inadequate for languages, such as German, which have less strict word order. Therefore we see this project in part as an attempt to discover attributes that will be useful for studying shell nouns in languages other than English.

4 Guidelines

4.1 General features of shell nouns

Shell nouns may be identified by means of three criteria:

1. A shell noun has *incomplete* semantic content.
2. A shell noun *refers* to linguistic content elsewhere in the discourse. This content could usually also occur without the shell noun itself. The shell noun serves to describe or characterize this content.
3. The shell noun content must be *abstract*. It will generally denote entites such as facts, propositions or eventualities.

4.2 Determining shell noun content phrases

Mark the shortest possible, but complete, instance of the content to which the shell noun refers.[3] The syntactic type of the content (e.g. *that*-clause or infinitive clause) should be apparent when viewed in isolation. Content phrases should be complete constituents. Often content phrases can be deleted and the sentence will remain well-formed, as in (3). This is however not the case for all content phrase types, and annotators may need to employ other constituent tests to determine the appropriate boundaries of a content phrase.

(3) a. Die **Entscheidung**, *inwieweit die EZB die allgemeine Wirtschaftspolitik der Gemeinschaft unterstützt*, hängt also von deren Einschätzung einer möglichen Beeinträchtigung des Ziels der Preisstabilität ab. [t_98-04-01_154]

b. The **decision** *on how far the European Central Bank supports the general economic policy of the Community* thus depends on its assessment of a possible effect on the aim of price stability.

[3] After Kolhatkar (2015).

4.3 Nominalizations

Though the content phrase (owing to its propositional nature) generally has a verbal head (is either a VP or CP), there are some deverbal nouns which still take complements and thus possess a similar semantics to these verbal phrases (at least in German). These phrases can also act as the content phrase in a shell noun complex at least if they follow the shell noun in a postnominal prepositional phrase. The content of the shell noun in this example could be equivalently expressed either with a VP or an NP:

(4) a. Hier gibt es die **Möglichkeit** *zur Aktualisierung der Software*.
'Here there is the opportunity for the updating of the software.'

b. Hier gibt es die **Möglichkeit**, *die Software zu aktualisieren*.
'Here there is the opportunity to update the software.'

If such a paraphrase is not possible, as in (5), then it is unlikely that the given noun phrase's meaning is propositional and that the token in question constitutes a shell noun usage.

(5) a. Mein Antrag *zur Geschäftsordnung* lautet wie folgt: [...] [t_99-11-16_145]

b. The point of order is as follows: ...

Such cases may look similar to conventional coreferential nouns, but the syntactic behavior of nominal shell noun complexes differs from conventional nominal coreference. For instance, in constructions like (4a), coreference appears to be only possible if the involved nouns are a shell noun and (typically) a deverbal noun.

4.4 Anaphoric shell noun complexes

Dipper and Zinsmeister (2009) introduce the 'paraphrase test', which assists annotators in locating the content of anaphoric expressions – this test may also be applied to anaphoric shell noun complexes. Upon encountering an anaphoric expression, such as *this problem*, add a '*namely* clause' along with a paraphrase which best completes the *namely* clause. The content of this paraphrase (or the most similar formulation) should be marked in the text as the shell noun content.

(6) a. Dieser Artikel in seiner jetzigen Fassung würde nämlich verhindern, *daß in Fragen des dritten Pfeilers präjudizielle Beschwerde von den Gerichten eingelegt werden könnten*. Das wäre sehr gefährlich, denn damit würde man den Gerichten eine Möglichkeit nehmen; ließe man zumindest dem höchstinstanzlichen Gericht diese **Möglichkeit**, so wäre es eine Garantie für die Bürger, denn der Gerichtshof spielte dann eine wichtigere Rolle. [t_97-05-28_93]

b. It is precisely that which, in its current version, would prevent *any presentation by the Court of Justice of appeals which would prejudice matters affecting the third pillar*. That really would be very dangerous, because it would mean cutting off a possibility which those courts have and, if at least that **possibility** were left to the Supreme Court, that would be a guarantee for citizens and would give the Court of Justice a more important function.

The 'namely' paraphrase:

(7) a. [...] ließe man dem höchstinstanzlichen Gericht diese **Möglichkeit**, *nämlich daß in Fragen des dritten Pfeilers präjudizielle Beschwerde von den Gerichten eingelegt werden könnten*, so wäre es [...]

b. [...] if at least that **possibility**, namely *(some) presentation by the Court of Justice of appeals which would prejudice matters affecting the third pillar*, were left to [...]

4.5 Cataphoric shell noun complexes

The content of cataphoric shell noun complexes can generally be found in the same sentence as the shell noun, in a subordinated phrase. However, in some cases, the content is farther away and difficult to localize, such as is more often the case with anaphoric shell noun complexes.

(8) a. Mein **Antrag** zur Geschäftsordnung lautet wie folgt: Dies ist ein so wichtiges Thema, von dem Landwirte im gesamten Vereignigten Königreich betroffen sind, *daß uns wirklich mehr Zeit für Fragen an den Kommissar*

zur Verfügung stehen sollte. [t_99-11-16_145]

b. The **point of order** is as follows: this is such an important issue which affects British farmers across the UK *that we should surely have more time to question the Commissioner*.

To help with the localization of the content for such phrases, one might pose *clarification questions*. After a shell noun (such as *Antrag* or *request*) has been identified, one might pose the question, *was wurde beantragt?* or *what did the speaker request?*. Then select as the content phrase the text which most succinctly answers this question. There may be cases where multiple phrases seem to answer the question equally well. These phrases might even be literal restatements of the same content. In such cases, the annotator should choose the statement which is located closest to the shell noun.

4.6 Content phrase types

It is possible that the content phrase for a particular shell noun usage is not to be found in the present turn, either because the speaker has intentionally left this information implicit or because the shell noun refers to content located in some other turn. Further, it is possible that, for some shell nouns, it might be unclear whether the information is indeed located elsewhere in the text or intentionally omitted. The following choices are provided to annotators:

given	The shell noun content is present in the given text (and accordingly marked).
external	Wording implies that the speaker is referring to a specific linguistic entity, located nearby, though not in the current text.
unclear	It is unclear whether the noun is used as shell noun or not.

4.7 Multivalent shell nouns

Some shell nouns (most notably *reason*) can accept multiple content phrase complements, such as distinct causes and effects. In the case of *reason*, the shell noun content consists primarily of the 'cause' complement, since this is the content being described as the 'reason' for some other state of affairs.

Example of an *attempt*-class SN:

(9) a. *[Die Ausweitung das Emissionshandelssystems (ETS) der EU auf den Luftverkehr]*$_1$ ist vielleicht die beste **Möglichkeit**, *[um diese Emissionen zu begrenzen und um dafür zu sorgen, dass der Luftverkehr so wie alle anderen Sektoren einen Beitrag zur Senkung der schädlichen Treibhausgase leistet.]*$_2$ [t_06-07-04_136]

b. *[The extension of the EU Emissions Trading Scheme (ETS) to the aviation sector]*$_1$ may be the best **way** forward *[to limiting these emissions and to ensuring that aviation, like all other sectors, contributes to reducing harmful greenhouse gases]*$_2$.

The first clause contains the content of the shell noun, but the clause which would match most conventional patterns, the second one, contains what might better be construed as a goal or result of the content in the first clause, rather than the entity to which the shell noun actually refers. Clarification questions can be helpful to identify the correct referent in such cases (Kolhatkar, 2015):

(10) The primary *reason* that the archdiocese cannot pay teachers more is that its students cannot afford higher tuition.

(11) Q. What was the reason?

A. Because *its students cannot afford higher tuition*.

4.8 Coordination

Another instance in which a shell noun can accept multiple content phrase complements is that of coordination:

(12) a. Doch die **Feststellungen**, (1) *dass Europa kein neues Wissen schafft*, (2) *dass es nicht in der Lage ist, Wissen gemeinsam zu nutzen*, und (3) *dass es Europa nicht gelingt, Wissen finanziell zu fördern*, sollten uns doch sehr zu denken geben. [t_06-07-04_200]

b. However, the **statements** (1) *that Europe does not seek to acquire new knowledge*, (2) *that it cannot share knowledge* and (3) *that it does not support knowledge financially* all have a very ominous ring to them.

Each of these individual content phrases could stand alone. Hence, they should be regarded as separate content phrases of the same type and annotated accordingly.

(13) a. Es kam die **Frage** auf *wann wir diesen Punkt besprechen und endlich abschließen*.
b. The **question** *of when we would discuss this issue and be finished* was posed.

Here the NP *diesen Punkt* complements both *besprechen* and *abschließen*, which means that here two content phrases are not being coordinated, rather two verbs (or subordinated VPs). In this case, *endlich abschließen* could not stand alone and is dependent on the rest of the phrase, therefore this is an example of just one content phrase.

Contents may also be described with multiple shell nouns, which does not result in any particular consequences for our annotation schema, i.e. multiple references are possible in both ways – a single shell noun instance can point to a number of different content phrases, while one and the same content phrase can be pointed to by multiple shell nouns:

(14) a. Es ist unser **Wunsch** und unsere **Absicht**, *ein [. . .] Wahlsystem [. . .] einzuführen*. [t_97-06-11_76]
b. It is our **wish** and **intention** *to introduce a new electoral system . . .*

In example (14) both shell nouns, *wish* and *intention*, should be annotated with a pointer to the same content phrase entity, marked here in italics.

4.9 Punctuation

Pairwise punctuation (such as quotation marks or parentheses) should be included in a shell noun phrase when one of the elements occurs within a content phrase. Other punctuation should be treated like whitespace in sentence-internal content phrases, i.e., when it occurs within the phrase it is included, but at the beginning or end, it is ignored. Punctuation at the beginning or end of sentences, however, is regarded to belong to the sentence and is thus included, which appears to be more natural than excluding them.

4.10 Alignment

Both shell noun instances and their associated content phrases should be manually aligned cross-linguistically. In many cases it is rather straightforward what elements are to be aligned, but if expressions are not formulated analogously across languages it might be difficult to decide what elements belong together. Furthermore, elements occurring in one language do not necessarily correspond to a linguistic item in the other language, i.e. there might be occurrences of shell nouns or content phrases without any alignment:

(15) a. I would also ask that the Commission take note of the **fact** *that the European people would welcome Mr Mobutu as much as they would the greatest criminal*. [t_97-05-28_22]
b. Ich fordere die Kommission auch auf, zur Kenntnis zunehmen, *daß die Bürger Europas Herrn Mobutu ebenso freundlich wie den größten Kriminellen begrüßen würden*.

Shell noun phrases can be referred to by a number of lexical items that do not belong to the class of shell nouns, e.g. pronouns or – in in the case of German – pronominal adverbs (such as *deshalb*, *daher*). Such entities should be marked as negative shell noun instances, but only if their counterpart in the other language is an actual shell noun.

5 Discussion

In total, about 2140 potential shell noun instances were annotated by both annotators. (The first third of these served as practice data, such that only two thirds of these instances are reflected in the statistics for inter-annotator agreement.) Of these, a little less than half were marked by both annotators as positive instances. Subsequently, since content phrases can only be marked along with actual shell nouns, there are approximately half as many content phrase annotations in the data. Figure 2 provides an overview of the relative frequencies of positive and negative shell noun instances.

In general, shell nouns appear to have been used more frequently in English in our data. This is likely due, at least in part, to the tendency of certain predicates in English to only accept NP complements (as opposed to sentential complements). For example, *to take note of X* requires *X* to be an NP, and this NP often takes the form of a shell noun complex, such as *the fact that [. . .]*. This stands in contrast to a number of German expressions which follow the pattern *zu Kenntnis nehmen, dass . . .*

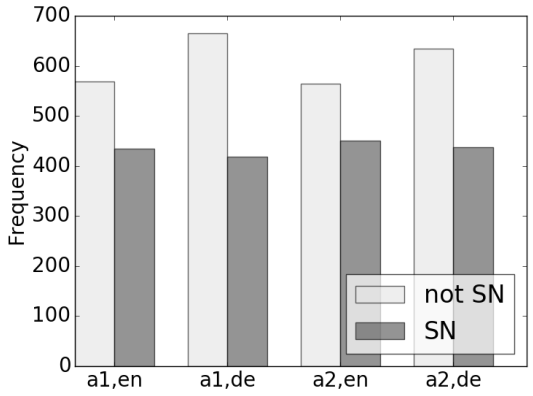

Figure 2: Comparison of shell noun vs. non–shell noun instances. ("a1" = Annotator 1, "a2" = Annotator 2)

	undef	false	true	unclear
undef	26	8	6	0
false	2	681	90	14
true	1	60	433	4
unclear	0	1	3	0

Figure 3: Confusion matrix for the two annotators ("undef" = Unannotated instance, "false" = Not a shell noun, "true" = shell noun, "unclear" = Not clear whether instance is shell noun or not).

(roughly, 'to take note that . . . '); here *zu Kenntnis nehmen* accepts a CP directly.

For the shell noun annotation level, in which annotators must mark a shell noun candidate as being an actual shell noun instance or not, raw agreement between annotators was 86%. Inter-annotator agreement, calculated according to Scott's π and Cohen's κ (Artstein and Poesio, 2008) both provide values of 0.73 for both languages taken together (though these values were minimally lower for English alone, 0.72). (Figure 3 provides more detail.)

Where a shell noun was marked as a positive instance, annotators are also asked to locate its content phrase as a span of text. When we take these spans to be sequences of token IDs, then each positive shell noun instance can be associated with a set of such spans. When comparing the sets corresponding to overlapping shell noun instances directly, approximately 65% of such sets were marked identically by both annotators. This number is comparable to Dipper and Zinsmeister's (2012, p. 47) observed agreement on exact matches, for which they report a value of 40%.[4] If, however, we require only that each annotated span overlap with some span from the other annotator, then 96% of the annotated content phrase spans could be considered matches (compared to 84% in the above-mentioned study).

Since the annotators could mark multiple, potentially discontinuous sequences of tokens for this task, determining annotator agreement is a nontrivial problem. We decided to use Krippendorff's α (Krippendorff, 2011), which was used by Kolhatkar and Hirst (2012) for a similar annotation task. This not only means that we were able to use an agreement measure appropriate to our data, but also that our values will be comparable to those resulting from a similar annotation task. We obtained a value of $\alpha = 0.84$, which is a relatively good value (by Krippendorff's standard) and a plausible one too, since it is only slightly worse than the reported agreement in Kolhatkar and Hirst's study ($\alpha = 0.86$, p. 1258).

We also analysed the distances between shell nouns and their content, for instance, in order to determine whether anaphoric shell noun complexes (in which the content precedes the shell noun instances) might be more frequent in one language or the other. In fact, as Figures 4 and 5 show, there do appear to be differences between English and German in this regard. Namely, German content phrases appear to occur more frequently at a greater distance to their shell noun, whereas English content phrases follow in the vast majority of cases the shell noun directly. The two-sample KS-test[5] confirms that the difference between these two distributions is statistically significant ($p = 1.28 \times 10^{28}$). The distribution of English content phrases shows that Schmid's (2000) pattern-based approach was appropriate for English, in that it is likely to have covered most of the data. However, our data for German show that such an approach is unlikely to suffice for the study of shell nouns in other languages.

Noting that content phrases are often headed by deverbal nouns (see Section 4.3 above), we also annotated the data with information regarding the syntactic status of the content phrases, i.e., whether they were nominal or not. There appear to be in-

[4]NB: Though that study involved the annotation of the antecedents of abstract anaphors, annotating the content phrases of shell nouns is, in many ways, the same task.

[5]The two-sample KS-test tests the null hypothesis that two independent samples come from the same distribution.

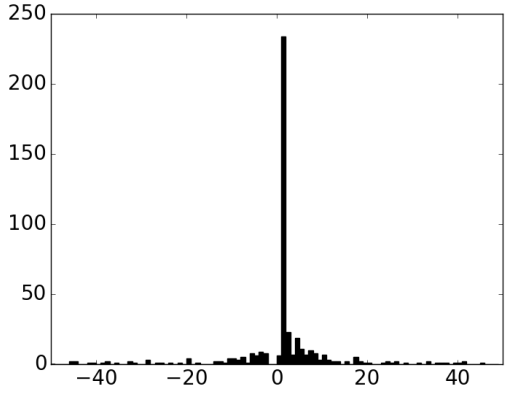

Figure 4: Distance between shell nouns and their content in tokens (English).

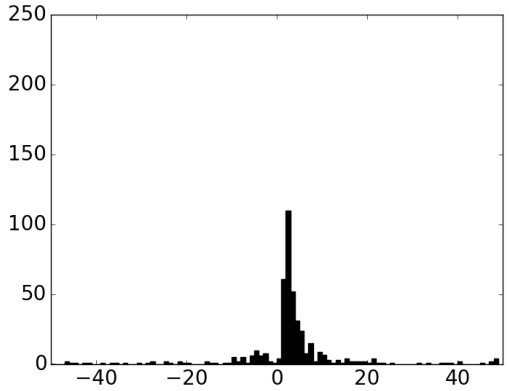

Figure 5: Distance between shell nouns and their content in tokens (German).

teresting cross-linguistic differences in this regard as well, for instance that nominal content phrases are more common in German, which could be of interest in future studies (cf. Figure 6).

6 Outlook

The data which was produced in this study and which can be produced using our annotation schema allow for the investigation of a number of questions which would be difficult to approach otherwise, such as those concerning the relative usage of shell nouns *in general* as well as the relative usage of *particular* shell nouns in German and English.

These data could furthermore serve as training data for clustering algorithms or other machine learning algorithms for categorizing content phrases or categorizing shell nouns based on the content phrases which they prefer. Such a typology of shell nouns, apart from its theoretical value, could aid in the automatic resolution of shell nouns and their content phrases.

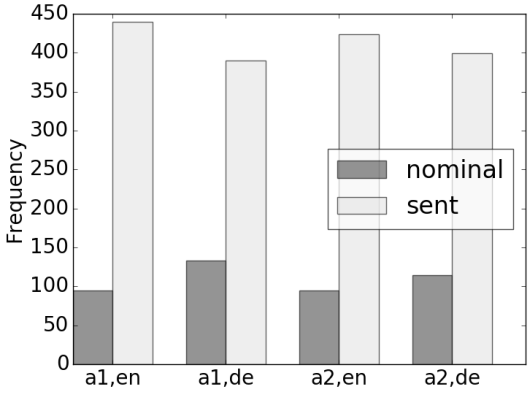

Figure 6: Shell noun content types. ("a1" = Annotator 1, "a2" = Annotator 2)

The annotated data can be found at: `https://github.com/ajroussel/shell-nouns-data`.

Acknowledgments

Many thanks to our anonymous reviewers for their helpful comments and criticism!

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

Stefanie Dipper and Heike Zinsmeister. 2009. Annotating discourse anaphora. *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP*, (August):166–169.

Stefanie Dipper and Heike Zinsmeister. 2012. Annotating abstract anaphora. *Language Resources and Evaluation*, 46(1):37–52.

Stefanie Dipper, Christine Rieger, Melanie Seiss, and Heike Zinsmeister. 2011. Abstract anaphors in German and English. In *Lecture Notes in Computer Science*, pages 96–107. Springer.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, volume 5, pages 79–86.

Varada Kolhatkar and Graeme Hirst. 2012. Resolving "this-issue" anaphora. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*, pages 1255–1265, Jeju Island, Korea.

Varada Kolhatkar and Graeme Hirst. 2014. Resolving shell nouns. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 499–510, Doha, Qatar.

Varada Kolhatkar, Heike Zinsmeister, and Graeme Hirst. 2013. Annotating anaphoric shell nouns with their antecedents. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 112–121, Sofia, Bulgaria.

Varada Kolhatkar. 2015. *Resolving shell nouns*. Ph.D. thesis, University of Toronto.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. Retrieved from `http://repository.upenn.edu/asc_papers/43`.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Hans-Jörg Schmid. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. de Gruyter, Berlin.

Fabian Simonjetz. 2015. Retrieving German shell nouns using dependency patterns. `http://www.researchgate.net/publication/306020586_Retrieving_German_Shell_Nouns_Using_Dependency_Patterns`.

A Sample of Annotations

Most common shell nouns in German and then in English. Cases in which the ratio of shell noun usages to non–shell noun usages was 1.0 have been filtered out; generally, whenever this ratio was 1.0, the shell noun only occurred once and was only added to the test set for the sake of alignment with a shell noun instance in the other language. Cases are also not listed here if the total number of shell noun usages was less than 2. Shell nouns which were not in the original list of 50 shell noun candidates (which were pre-marked for annotation) are listed here in **boldface**.

Legend:

Undef.	Number of instances left unannotated.
False	Number of instances *not* functioning as a shell noun.
True	Number of instances functioning as a shell noun.
Unclear	Annotator was unable to determine whether or not instance constitutes use as a shell noun.
%SN	Ratio of shell noun vs. non–shell noun instances.

A.1 Annotator 1

Lemma	Undef.	False	True	Unclear	%SN
Frage	5	84	73	0	0.450617
Möglichkeit	0	11	30	0	0.731707
Ziel	2	17	22	0	0.536585
Forderung	0	7	19	0	0.730769
Tatsache	2	0	19	0	0.904762
Vorschlag	1	69	19	0	0.213483
Auffassung	0	6	17	0	0.739130
Ansicht	2	23	17	0	0.404762
Recht	6	59	17	0	0.207317
Grund	0	7	14	1	0.636364
Meinung	0	18	12	0	0.400000
Entscheidung	0	41	9	0	0.180000
Plan	0	5	7	0	0.583333
Gelegenheit	1	5	7	0	0.538462
Gefahr	0	6	7	0	0.538462
Aufgabe	0	8	7	0	0.466667
Verpflichtung	0	4	7	0	0.636364
Antrag	0	17	6	0	0.260870
Überzeugung	0	2	6	0	0.750000
Versuch	0	1	6	0	0.857143
Voraussetzung	0	9	6	0	0.400000
Hoffnung	0	4	5	0	0.555556
Schlussfolgerung	0	5	4	0	0.444444
Pflicht	0	4	4	0	0.500000
Bereitschaft	0	1	3	0	0.750000
Hinweis	1	6	3	0	0.300000
Absicht	0	1	3	0	0.750000
Wunsch	0	3	3	0	0.500000
Argument	0	3	2	0	0.400000
Standpunkt	1	17	2	0	0.100000
Lage	3	31	2	0	0.055556
Argumentation	0	1	2	0	0.666667
Zielsetzung	0	3	2	0	0.400000
Wille	0	6	2	0	0.250000

Lemma	Undef.	False	True	Unclear	%SN
fact	6	18	47	0	0.661972
question	0	29	46	0	0.613333
reason	1	13	30	1	0.666667
need	4	19	25	0	0.520833
opportunity	0	18	18	0	0.500000
right	0	23	17	0	0.425000
proposal	2	69	17	0	0.193182
issue	0	19	16	0	0.457143
aim	1	0	14	0	0.933333
decision	1	41	10	0	0.192308
objective	0	23	10	0	0.303030
view	0	16	8	0	0.333333
plan	0	4	7	0	0.636364
possibility	0	6	7	0	0.538462
idea	1	9	6	0	0.375000
hope	1	3	6	0	0.600000
effort	0	13	6	1	0.300000
conclusion	2	19	6	0	0.222222
requirement	0	8	5	0	0.384615
risk	0	2	5	0	0.714286
opinion	0	9	5	0	0.357143
intention	0	1	5	0	0.833333
argument	0	7	5	0	0.416667
demand	0	3	5	0	0.625000
duty	0	4	4	0	0.500000
commitment	0	3	4	0	0.571429
point	0	2	4	0	0.666667
problem	0	1	3	0	0.750000
suggestion	0	2	3	0	0.600000
attempt	0	1	3	0	0.750000
indication	0	1	3	0	0.750000
matter	0	13	2	0	0.133333
occasion	0	1	2	0	0.666667
courage	0	2	2	0	0.500000
promise	0	1	2	0	0.666667
danger	0	4	2	0	0.333333
request	0	1	2	0	0.666667
wish	0	4	2	0	0.333333
option	0	2	2	0	0.500000
doubt	0	10	2	0	0.166667

A.2 Annotator 2

Lemma	Undef.	False	True	Unclear	%SN
Frage	0	71	85	6	0.524691
Vorschlag	1	55	32	1	0.359551
Möglichkeit	0	11	30	0	0.731707
Ziel	2	15	22	2	0.536585
Tatsache	2	0	19	0	0.904762
Forderung	0	8	18	0	0.692308
Ansicht	2	23	17	0	0.404762

Auffassung	1	6	16	0	0.695652
Grund	0	6	15	1	0.681818
Meinung	0	18	12	0	0.400000
Entscheidung	0	37	12	1	0.240000
Recht	3	61	11	0	0.146667
Voraussetzung	0	7	8	0	0.533333
Plan	0	4	8	0	0.666667
Aufgabe	0	8	7	0	0.466667
Gelegenheit	1	6	6	0	0.461538
Überzeugung	0	2	6	0	0.750000
Notwendigkeit	0	1	6	0	0.857143
Lage	4	27	5	0	0.138889
Verpflichtung	0	6	5	0	0.454545
Hoffnung	0	5	4	0	0.444444
Antrag	0	19	4	0	0.173913
Gefahr	0	9	4	0	0.307692
Schlussfolgerung	0	6	4	0	0.400000
Hinweis	1	6	3	0	0.300000
Pflicht	0	5	3	0	0.375000
Absicht	0	1	3	0	0.750000
Wille	0	6	2	0	0.250000
Standpunkt	1	17	2	0	0.100000
Argument	0	2	2	1	0.400000
Wunsch	0	4	2	0	0.333333

Lemma	Undef.	False	True	Unclear	%SN
question	0	21	53	2	0.697368
fact	3	21	47	0	0.661972
reason	0	12	31	1	0.704545
need	1	20	27	0	0.562500
proposal	1	57	26	1	0.305882
opportunity	0	11	20	0	0.645161
issue	0	19	12	2	0.363636
decision	1	37	12	1	0.235294
objective	0	18	12	2	0.375000
aim	1	2	12	0	0.800000
right	0	30	11	0	0.268293
possibility	0	3	10	0	0.769231
plan	0	4	8	0	0.666667
call	0	1	7	0	0.875000
view	0	17	7	0	0.291667
argument	0	3	7	1	0.636364
idea	1	9	6	0	0.375000
effort	0	13	6	1	0.300000
conclusion	1	20	6	0	0.222222
position	0	14	5	0	0.263158
opinion	0	10	5	0	0.333333
demand	0	3	5	0	0.625000
point	0	2	5	0	0.714286
hope	0	5	4	0	0.444444
commitment	0	4	4	0	0.500000

desire	0	1	3	0	0.750000
failure	1	1	3	0	0.600000
duty	0	5	3	0	0.375000
indication	0	1	3	0	0.750000
requirement	0	7	3	0	0.300000
promise	0	1	2	0	0.666667
option	0	2	2	0	0.500000
danger	0	4	2	0	0.333333
wish	1	4	2	0	0.285714
condition	0	2	2	0	0.500000
matter	0	12	2	0	0.142857
request	0	2	2	0	0.500000
courage	0	2	2	0	0.500000