On the difficulty of making concreteness concrete

Daphne Theijssen, Hans van Halteren, Lou Boves, Nelleke Oostdijk

Centre for Language Studies, Radboud University Nijmegen

Abstract

The use of labels of semantic properties like 'concreteness' is quite common in studies in syntax, but their exact meaning is often unclear. In this article, we compare different definitions of concreteness, and use them in different implementations to annotate nouns in two data sets: (1) all nouns with word sense annotations in the SemCor corpus, and (2) nouns in a particular lexico-syntactic context, viz. the theme (e.g. *a book*) in prepositional dative (*gave a book to him*) and double object (*gave him a book*) constructions. The results show that the definition and implementation used in different approaches differ greatly, and can considerably affect the conclusions drawn in syntactic research. A follow-up crowdsourcing experiment showed that there are instances that are clearly concrete or abstract, but also many instances for which humans disagree. Therefore, results concerning concreteness in syntactic research can only be interpreted when taking into account the annotation scheme used and the type of data that is being analysed.

1 Introduction

Syntacticians commonly use labels referring to semantic properties in their research, such as animacy, imaginability, concreteness, etc. These terms have become so familiar that explicit definitions of the semantic properties implied are hardly ever provided. But when definitions are provided, these may differ between different researchers. To complicate things even further, the same definition can be instantiated in different implementations or annotation guidelines. The eventual annotations can vary with respect to the range of the values that can be assigned, and with respect to their measurement scale: e.g. binary, nominal (multiple labels that cannot be ordered on a scale from 'low' to 'high'), ordinal (multiple labels that are ordered), or interval-level scores. Finally, there is the issue of replicability. For manual annotations, it is often difficult to reach high inter-annotator agreement (e.g. Theijssen et al. 2009), casting doubt on the quality of the data. For (semi-)automatic implementations, in which one employs automatic algorithms or tools, the replicability is guaranteed, but the validity may be questionable.

Existing research directed at comparing annotation approaches often suggests ways to standardise the different labels used in different language resources (e.g. Ide and Romary 2008), or presents methods to assess the agreement between labels assigned by different annotators using the same scheme (e.g. Artstein and Poesio 2008) or the quality of automatically obtained labels, compared to gold standard labels (e.g. Kübler 2007). The comparisons are made with the ultimate goal of arriving at a standard that can be used across resources and studies. In this article, we investigate the impact of different instantiations of the definition of 'concreteness'. Instead of trying to define some standard definition of concreteness, we want to establish how the differences between the definitions and the

implementations affect the actual labels obtained (an *intrinsic* comparison). Second, we want to investigate how the outcome of syntactic research is influenced when we employ the labels obtained through the different approaches (an *extrinsic* comparison).

For tackling the first goal, we employ a data set consisting of 68,484 nouns annotated with a WordNet word sense in the SemCor corpus (Miller et al. 1993): the SEMCOR data set[1]. Using four (semi-)automatic approaches, we assign values for concreteness to these nouns and compare these quantitatively and qualitatively.

For the second goal, we take as a case study the English dative alternation, where speakers can choose between a prepositional dative construction (e.g. *I gave the book to him.*) and a double object construction (e.g. *I gave him my love.*). The logistic regression models in Theijssen (2010)[2] show that if the direct object or 'theme' is **concrete** (e.g. *the book*) speakers tend to place it *before* the recipient *him*, while it is the other way around if the theme is **abstract** (e.g. *my love*). We assign concreteness values to the themes in 619 instances (the DATIVE data set) using six different labelling approaches. The labels from the various approaches are then included in logistic regression models that predict which of the two constructions is used. We compare the models to see how the selection of an approach to annotate concreteness affects the eventual conclusions.

As will become evident in the analyses, the actual labels in the SEMCOR data and the conclusions in the syntactic study based on the DATIVE data are indeed different for the various approaches. This makes us question to what degree humans agree on the interpretation of concreteness. To address this issue, we perform a follow-up experiment in which we ask humans to rate the concreteness of noun tokens in context, without providing them with a definition.

The article is structured as follows: Section 2 presents an introduction to concreteness and a description of the (semi-)automatic labelling approaches we use. Section 3 addresses the intrinsic comparison on the SEMCOR data, Section 4 the extrinsic comparison with the DATIVE data. The follow-up experiment is presented in Section 5. A summary and conclusion can be found in Section 6.

2 Annotation approaches for concreteness

The distinction between *concrete* and *abstract* has been addressed in a broad range of research topics, e.g. semantics, anaphora resolution, probabilistic syntax, metaphors, word sense disambiguation, syntactic/semantic acquisition and image retrieval. Quite a lot of the literature is written from the viewpoint of the antonym of concreteness: *'abstractness'*. Spreen and Schulz (1966) explain that there are at least two definitions of **abstract**: (1) general, generic, not specific, and (2) lacking sense experience. We can thus interpret concreteness as either 'specificity' or 'sensory perceivability'. The definition of concreteness as specificity originated

[1]We make a typographic distinction between the name of the corpus (SemCor) and the set of sense-annotated nouns (SEMCOR).

[2]Theijssen (2010) presents six different regression models for the same data set. Only the models without a random effect for verb include a significant effect for concreteness.

in cognitive and neuro-science. In linguistics, the interpretation of concreteness as sensory perceivability is most common, for example in the line of research initiated by Lyons (1977). Since there is more and more research that combines insights from cognitive science and linguistics (e.g. cognitive sociolinguistics, cf. Geeraerts et al. 2010), we include both definitions in this article.

According to Schmid (2000) "abstract nouns are those nouns whose denotata are not part of the concrete physical world and cannot be seen or touched. Strictly speaking, what is abstract is not the nouns themselves, but what they denote" (p. 63). What a noun denotes depends on its context. Concreteness should thus not be established for different word *types* (i.e., orthographic forms), but for the individual word *tokens* in context, since words can have several senses. For instance, the noun *table* may refer to the concrete object standing in a furniture showroom, or to the tables in this article. Furthermore, words in the same or similar sense can be used figuratively: when a waitress shows you a table, she may be literally showing you a concrete object (a specific *table*), but what she means is not just this table, but a place to have dinner. As a result, this *table* could be considered less concrete than the one standing in the furniture showroom.

Most theories about concreteness agree that context is essential, but not all actual labelling approaches take context into account. Projects with human annotators have usually employed a two- or seven-point scale to establish the concreteness of nouns. In the binary distinction between concrete and abstract, concrete nouns are usually defined as referring to tangible, prototypically concrete, or real existing entities. In (semi-)automatic approaches to establish the concreteness of noun senses, the use of the lexical database WordNet (Fellbaum 1998) is quite common. One can also look up the concreteness of noun types in databases, such as the MRC Psycholinguistic Database (Coltheart 1981).

Other researchers have developed automatic approaches that use the context to assign noun tokens to noun classes (e.g. classes such as 'building' and 'event', Thelen and Riloff 2002). One can start out with a seed set of unequivocally concrete and abstract noun types, and use a bootstrapping process: Seek patterns in the context of concrete and abstract nouns, and use them to find new examples of concrete and abstract nouns for the seed set. The patterns found after the final iteration are used to establish the concreteness of the data that needs annotation.

We compare four different approaches to annotate concreteness. The resulting labels differ in their underlying definition (specificity or sensory perceivability), the 'noun level' serving as their basis (token, sense or type), the measurement scale (interval, ordinal or nominal) and the manner in which the labels are obtained (semi-automatically or automatically).

2.1 MRC: The MRC Psycholinguistic Database

In this approach, we automatically look up the sensory perceivability of a noun *type*, being an interval value between 100 and 700.

The MRC psycholinguistic database (Coltheart 1981) contains 4,004 (pro)noun types that are marked for concreteness with a value between 100 and 700, the

higher the value, the more concrete. The scores are based on ratings of words in isolation, assigned by undergraduate students. Examples of very concrete nouns are *milk*, *tomato* and *grasshopper*, very abstract nouns are *unreality*, *infinity* and *while*. The annotation instructions were taken from Spreen and Schulz (1966): "Nouns may refer to persons, places and things that can be seen, heard, felt, smelled or tasted or to more abstract concepts that cannot be experienced by our senses" (p. 460).

2.2 BOOTS: Bootstrapping the BNC

In this approach, we automatically establish the sensory perceivability with the help of a bootstrapping approach that assigns interval values.[3] In Section 3.2, we will see that the approach is mostly *type*-based, despite the fact that it aims at classifying individual noun *tokens*.

We use the 100-million-word British National Corpus (BNC Consortium 2007) and parse it with the FDG dependency parser developed at Connexor Oy (Tapanainen and Järvinen 1997). This parser has a word class (part-of-speech) accuracy of 99.3%, and it reaches a precision of 93.5% and a recall of 90.3% for linking subjects and objects.[4] From the parses, we extract all words that the parser marked as nominal head (*%NH*) and noun (*N*). We only keep those instances that have also been tagged as a noun in the BNC (*NN**), excluding for example all proper nouns. For each remaining instance, we find the lexico-syntactic patterns in which the heads appear and save them as features. The features are thus the direct dependency relations that exist between the noun phrase head and other words in the dependency parse. So, for the sentence *The major impact is yet to come*, the features for the head *impact* are *m:subj;be* (it is the subject of its 'mother' *be*), *d:det;the* and *d:attr;major* (its 'daughters' are the determiner *the* and the attribute *major*). Notably, the head noun itself is not included in the features. The resulting 'BNC set' contains 17,708,616 tokens (170,893 different noun types).

For the initial seed set, we take the examples in Garretson (2003), consisting of 27 prototypically concrete nouns (e.g. *apple*, *door* and *knee*) and 10 prototypically abstract nouns (e.g. *air*, *current* and *molecule*). We find all instances of these nouns in the BNC set and label them **concrete** or **abstract** accordingly. Next, we train a pattern discovery algorithm on this 'labeled set' inspired by the procedure in Thelen and Riloff (2002). We aim to identity features F which can act as 'concrete markers', i.e. the F with sufficient precision in suggesting concreteness and high enough frequency to be of use. We first count the number of concrete instances in the labeled set that have feature F and divide this number by the total number of concrete instances in the labeled set: $PropF_C$. The same is done for the abstract instances, yielding proportion $PropF_A$. We then check whether $PropF_C \geq w \cdot PropF_A$. The w is a strictness weight that makes the procedure stricter or more

[3] In theory, the values range from -∞ to ∞, but in practice, the range varies per data set: -0.50 to 0.85 for SEMCOR and -0.23 to 0.18 for DATIVE.

[4] These figures were established on texts from the Maastricht treaty by Connexor Oy in December 2005, after which only minor changes have been applied to the parser.

lenient, which we set to 9. Also, we demand that F occurs at least 50 times.[5] Next, the strength of a concrete marker is established with:

$$Str_C = (\frac{PropF_C}{PropF_C + PropF_A})^2 . \quad (1)$$

The division of the proportions on the right-hand side of the equation is squared in order to enhance the relative value: strong markers are made relatively even stronger and weak markers relatively even weaker. After calculating the strengths of the concrete markers, we use the same approach to find features that are 'abstract markers', assigning them strength Str_A.

The instances in the unlabeled set are assigned a score for concreteness by adding all strengths Str_C for the concrete markers present, and subtracting all strengths Str_A for the abstract markers present. We next group all instances of the same noun and take the average of the scores assigned to them in order to find new seed nouns. The nouns are ranked according to this average score, and we add the top 100 nouns (with an average score above 0.001) and the bottom 100 nouns (with an average score below -0.001) to the list of concrete and abstract seed nouns, respectively. With the new seed set, we start over again, looking for new patterns and new seed nouns. After 100 iterations, we stop and use the final set of patterns and the final set of seed nouns, both including the concreteness scores they yield.

Using the FDG parser and the set of markers and seed nouns, we can assign a score to new instances. If no score can be assigned to the noun *token* because no concreteness or abstractness markers are present, we check whether the noun *type* (i.e. the lemma) is present in the final set of seed nouns. If so, we assign the average score corresponding to that seed noun.

2.3 WN-HIER: The hierarchy level in WordNet

In this approach, we semi-automatically establish the specificity of a noun *sense*, assigning an ordinal value between 0 and 16. Different tokens with the same sense always receive the same score, which is different from a true token-based approach that takes into account the context of the individual token.

We follow Changizi (2008), who established the *specificity* of a noun with a particular sense by counting the number of hypernyms above it, using the WordNet hierarchy (Fellbaum 1998). For instance, *bullock* is very concrete, having the maximum hypernym level of 16, while *entity* is very abstract, with the minimum hypernym level of 0. This approach is semi-automatic: The word sense is manually assigned on the basis of the context, and we use this sense to establish its hypernym level automatically by looking it up in the manually designed WordNet. It is the only approach in this article that uses the definition of specificity, not sensory perceivability. We thus expect this approach to differ the most from the others.

[5] Both numbers were established by trial and error, manually monitoring the selection of new seeds.

2.4 WN-PHYS: Physical entities in WordNet

In this approach, we semi-automatically establish the sensory perceivability of a noun *sense*, assigning one of the two nominal values 0 and 1.

Again, this is not at the type or token level, but at sense level. We follow the approach in Xing et al. (2010): We automatically check whether the noun sense is traced back to *physical entity* in WordNet. If so, it is labelled 1, and otherwise 0. Again, this approach is semi-automatic since the word senses are found manually, and employ the manually established WordNet.

3 Intrinsic comparison: SEMCOR

3.1 Data

The SemCor corpus contains manually assigned WordNet word senses for all 88,058 noun phrases in 186 texts taken from the Brown corpus. Since we also want to apply the BOOTS approach, we parse the sentences with the FDG dependency parser and keep only those instances that the parser marked as being a nominal head. The result is the data set SEMCOR consisting of 68,484 instances.

3.2 Method and results

We apply the four approaches for concreteness labelling to the 68,484 instances in SEMCOR. For 44,395 instances, the noun is present in the MRC Database, which means there is a missing value for MRC in 24,089 instances. These instances are not included in the evaluations with MRC. There are also missing values for BOOTS: For 8,835 instances, no score could be assigned because there was no concrete or abstract marker present, and the noun itself was not present in the seed list. All evaluations of BOOTS are thus based on the 59,649 instances for which we could assign a concreteness score. Only 5,099 of these scores are based on the presence of abstract or concrete markers in the individual *tokens*, the rest is assigned a score by looking up the concreteness score of the noun *type* in the list of seed nouns. Apparently, the set of markers is too sparse to enable the classification of the noun tokens in their lexico-syntactic context.

The concreteness scores found are compared with the help of the Spearman rank correlation coefficient, cf. Table 1. For comparisons with missing values (i.e. all comparisons except that between WN-HIER and WN-PHYS), we only use the instances without missing values (42,120 for the comparison between MRC and BOOTS). The highest correlation (0.64) is between MRC and BOOTS, which could be the result of the fact that both are mostly *type*-based. The correlation between MRC and WN-PHYS is only slightly lower (0.60). WN-HIER differs most from the other three approaches, with all correlations below 0.30.

To better understand the scores, we compared the average values of the senses in WordNet's 26 noun classes. Nouns in the classes 'animal' (e.g. *hen*), 'food' (e.g. *milk*), 'artifact' (e.g. *door*), 'body' (e.g. *arms*) and 'substance' (e.g. *water*) mostly received high concreteness scores in all four approaches, and those in the classes

Table 1: Spearman correlations between the different labelling approaches for the SEMCOR data set. The corresponding p-values are all < 0.001.

	MRC	BOOTS	WN-HIER	WN-PHYS
MRC	1.00	0.64	0.29	0.60
BOOTS		1.00	0.12	0.47
WN-HIER			1.00	0.17
WN-PHYS				1.00

'relation' (e.g. *relationship*), 'cognition' (e.g. *will*) and 'attribute' (e.g. *consequence*) mostly low scores. The approaches assigned medium or varying scores to nouns in the classes 'act' (e.g. *war*), 'group' (e.g. *people*) and 'phenomenon' (e.g. *daylight*).

BOOTS is exceptional in assigning nouns with noun class 'time' (e.g. *February*, *minute*) and 'quantity' (e.g. *ton*, *inch*) very high concreteness scores, while they are considered (relatively) abstract by the other three approaches. Apparently, somewhere in the bootstrapping process, nouns denoting time or quantity have been included as seed nouns. As a result, time and quantity specific contexts were selected as markers of concreteness, leading to the selection of even more time and quantity nouns as seeds.

For WN-HIER, nouns that have the noun class 'object' (e.g. *soil*, *unit*) receive relatively low scores because they are located relatively high up in the WordNet hierarchy, but relatively high concreteness scores in the other three approaches. For nouns with the noun class 'feeling' (e.g *trouble*) it is the other way around: WN-HIER considers them rather concrete because they are deeper in the WordNet hierarchy, while they are considered abstract in the other three approaches.

In order to discover how individual noun types are treated in the four approaches, we determined the 100 words with the highest scores for concreteness and abstractness.[6] For the words with multiple senses we averaged concreteness/abstractness over the senses.[7]

All four approaches agree that the noun *knowledge* is abstract. In addition, three of the four approaches have placed the following nouns in the bottom 100: *ability*, *approval*, *attitude*, *confidence*, *distinction*, *freedom*, *hatred*, *importance*, *indication*, *individualism*, *morality*, *motive*, *past*, *philosophy*, *quality*, *relationship*, *responsibility*, *security*, *sentiment*, *theory*, *understanding* and *weakness*. Eight nouns have been placed in the top 100 (very concrete) by all four approaches: *cigarette*, *coat*, *grass*, *hat*, *jacket*, *sheep*, *shirt* and *tree*.

[6]Ranking seems problematic for the binary approach WN-PHYS, but since labels are sense-based they can have different values for the same noun type. The average score is thus not necessarily 0 or 1.
[7]If number 100 is part of a tie, we include all nouns with the same value.

3.3 Discussion

Our comparison showed that the concreteness labels assigned by the four approaches vary considerably. We found no clear effect of the noun level (sense, type), the measurement scale (binary, ordinal or interval), or the annotation manner (semi-automatic or automatic). The differences we found were mostly caused by the definition used: As we would expect, 'specificity' and 'sensory perceivability' are quite different concepts. The diverging definition of WN-HIER ('specificity') made it differ greatly from the other approaches: It showed the lowest correlation with the other approaches, and the largest differences in the treatment of individual noun types and noun classes. While the other three approaches on average considered nouns denoting objects rather concrete and nouns denoting feelings rather abstract, this was the other way around for WN-HIER.

An analysis of BOOTS showed there is a risk in starting with a list of typically concrete and abstract nouns and using a bootstrapping approach to discover new concrete and abstract nouns on the basis of the lexico-syntactic context: Nouns denoting time and quantity were considered very concrete by BOOTS.[8] Moreover, the concrete and abstract markers learned from the BNC were hardly present in the SEMCOR data, which means that the fall-back option, the score of the nouns in the seed list, determined the concreteness score in most cases. As a result, most instances were not assigned a concreteness score according to their lexico-syntactic context (the individual *token*), but according to their lemma (the noun *type*).

The automatic look-up MRC has the most in common with the other three approaches. This is rather surprising, since it is the only approach that does not take into account the context at all. Apparently, the concreteness scores in the MRC database were based on the word sense that is the most frequent, minimising the effect of ignoring the intended sense. Also, different senses of the same noun can be similar with respect to concreteness. For instance, *arms* in the sense of weapons is arguably equally concrete as *arms* in the sense of the body parts. It thus seems that the lack of sense disambiguation in the MRC Database has little effect on the actual concreteness labels. The same is probably true for BOOTS, which is also mostly based on the noun types, and most similar to MRC according to the correlation coefficients. However, because the MRC Database was designed for a different purpose (providing test items for psycholinguistic experiments), its coverage is problematic: over 24,089 instances could not be annotated with MRC. The number of unclassified items is much smaller for BOOTS: 8,835.

4 Extrinsic comparison: DATIVE

In this section, we investigate how the choice for a concreteness labelling approach affects our conclusions in a syntactic study: the English dative alternation. Speakers can choose between a prepositional dative (e.g. *I gave the book to him.*) and a double object construction (e.g. *I gave him the book.*). Previous research (e.g.

[8] Similar effects occurred when we experimented with different initial seed sets.

Theijssen 2010) has indicated that over 90% of people's dative choices can be correctly predicted with a logistic regression model that combines features on different levels of description (semantic, syntactic, lexical, etc.). The features describe the two objects in the construction (e.g. *him* and *the book* in the example). The regression models show that the first object is typically (headed by) a pronoun, mentioned previously in the discourse, animate, **concrete**, definite and shorter (in number of words) than the second object. The second object generally has the opposite characteristics: non-pronominal, discourse-new, inanimate, **abstract**, indefinite and longer.

4.1 Data

We use the data set in Theijssen (2010), containing 930 instances of the dative alternation that were extracted from the one-million-word syntactically annotated ICE-GB corpus (Greenbaum 1996). The ICE-GB corpus contains written and spoken British English in various genres. The data set contains manual annotations for the explanatory features introduced in Bresnan et al. (2007). From the data set, we select those instances that were automatically parsed as an instance of the dative alternation by the FDG parser (Tapanainen and Järvinen 1997), and were manually approved by the first author in a previous study (Theijssen et al. 2011). The resulting data set (DATIVE) consists of 619 instances: 499 (80.6%) double object and 120 (19.4%) prepositional dative constructions.

For annotating the concreteness of the theme[9], we employ the four approaches in the previous section. Moreover, we include two additional approaches to establish the concreteness: the manual approach used for the annotation of the original data set (MANUAL) and an adapted bootstrapping approach (BOOTS-OBJ). Both approaches are described below.

MANUAL: Prototypically concrete

The sensory perceivability of a noun token is manually established, assigning one of the two nominal values 0 and 1. It is the concreteness feature as it is already annotated in Theijssen's (2010) data set. It follows Garretson (2003), who deems a noun concrete if it refers to a prototypically concrete object: "The rule of thumb to apply is that we want to code as 'concrete' only *good* examples of concrete things" (5.6.7). All nouns that fit this description are given value 1, all others value 0. Remember that the examples in Garretson (2003) were also used as the initial seed nouns in BOOTS. As MANUAL is established by hand, the full context has now been taken into account by the human annotator, as opposed to only the direct lexico-syntactic dependencies in BOOTS.

[9] As in Bresnan et al. (2007), the concreteness of the recipient (*him*) is not researched here, because it is highly imbalanced (there is a strong bias towards concrete recipients).

BOOTS-OBJ**: Bootstrapping direct objects in the BNC**

In Section 3, we saw that BOOTS is mostly a type-based approach because the lexico-syntactic markers are too sparse to allow token-based classification. For this reason, we apply this bootstrapping method to a subset of the original set extracted from the BNC, containing the 837,755 noun tokens (31,345 noun types) that, according to the FDG parse, are the direct object of one of the 76 'dative verbs' in Theijssen et al. (2011).[10] In this way, the set of patterns and seed nouns obtained are more likely to occur in DATIVE. The values assigned range from -0.34 to 0.19.

Since the DATIVE data set contains pronouns, we first manually establish the antecedents of the pronouns, if possible, and replace them by the head lemmas of their antecedents. The two WordNet-based approaches, WN-HIER and WN-PHYS, require additional manual annotation: we manually assign a WordNet sense to the theme.

As also found for the SEMCOR set in Section 3, a substantial proportion of the DATIVE nouns are missing in the MRC Database: The theme head is present in the MRC Database for 436 instances, so 183 instances have a missing value. We were able to assign a concreteness score to 546 instances with BOOTS and 475 instances with BOOTS-OBJ. WN-HIER and WN-PHYS also suffer from coverage issues: We were unable to find the intended word sense in WordNet for the theme head of 43 instances.

4.2 Method

Using the feature values, we establish a regression function that predicts the logarithm of the odds that the syntactic construction S in clause j is a prepositional dative. The prepositional dative is regarded a 'success' (with value 1), while the double object construction is considered a 'failure' (0). The regression function is:

$$\ln odds(S_j = 1) = \alpha + \sum_{k=1}^{K} (\beta_k V_{jk}) \; . \tag{2}$$

The α is the intercept of the function. $\beta_k V_{jk}$ are the weights β_k and values V_{jk} of the K variables k.[11] The optimal values for the function parameters α and β_k are found with the help of Maximum Likelihood Estimation.[12]

We employ nine features suggested in previous research: the Animacy of the recipient (*Rec = anim*), the Definiteness of the recipient and theme (*Rec = defin*, *Th = defin*), the Discourse Givenness of the recipient and the theme (*Rec = given*, *Th = given*), the Pronominality of the recipient and the theme (*Rec = pron*, *Th = pron*), the Person of the recipient (*Rec = 1st/2nd*), and the Length Difference

[10] Because of the smaller size of the bootstrap set, w is set to 5, and minimum marker frequency to 10.
[11] We should note that the regression function treats the ordinal feature WN-HIER as an interval feature.
[12] We use the function `lrm()` in `R` (R Development Core Team 2008).

between the theme and the recipient (*Len dif th-rec*)[13]. We also include a feature for the Medium, indicating whether the construction appeared in spoken or written data (*Med* = *wr*), and the six features for the Concreteness of the theme, obtained with the six different labelling approaches.[14] We build six separate regression models, each with one type of concreteness and the remaining ten features.

For all approaches except MANUAL, we have to deal with the missing data. We follow the standard procedure: All instances for which the concreteness is not known are removed before building the regression model.[15] The model for MRC is thus built on 436 instances, and those for WN-HIER and WN-PHYS on 576. For BOOTS, we use the 546 instances for which the noun *token* has a lexico-syntactic marker (only 12 instances) or the noun *type* is present in the final seed list (534 instances). In the BOOTS-OBJ version there are markers for 79 instances, and the noun type is a seed in 396 instances, leading to a total of 475 instances for which the concreteness could be established.

4.3 Results and discussion

The evaluation measure C[16] is above 0.95 for all the six models, which indicates that the models fit the data well (cf. Baayen 2008). When training and testing on all available instances, the prediction accuracy is above 0.91 for all six models, which is significantly better than the baselines reached when always selecting the double object construction (approximately 0.80, depending on the missing data). The β-coefficients and significance levels of the features in the models are presented in Table 2. The Table shows that employing different instantiations of concreteness results in different regression models. The differences are not only found for *Theme concreteness* itself, but also in the other features in the model.

Concreteness in the sense of sensory perceivability seems to play a role in the dative alternation, while concreteness in the sense of specificity (WN-HIER) does not (p=0.344). It thus seems that the definition most commonly used in linguistics, sensory perceivability, is indeed more informative in our corpus linguistic study.

The implementation of sensory perceivability affects the conclusions: the Concreteness of the theme is only significant at the 0.05-level for MANUAL, WN-PHYS and MRC. This means that only the implementations with manual input resulted in a significant effect in the models. This manual step consisted either of looking up the concreteness of nouns in the manually established MRC Database (MRC), of finding the noun sense by hand in the manually designed WordNet hierarchy (WN-PHYS), or of manually assigning a concreteness value to a noun token in context (MANUAL). Neither of the two bootstrapping approaches yielding in-

[13]Length Difference is defined as the log of the number of words in the theme minus the log of the number of words in the recipient, i.e. the log of the ratio between the two lengths.

[14]We divide the MRC score by 100 to prevent that its coefficient will become extremely small. Similarly, we multiply the values for BOOTS and BOOTS-OBJ by 10 so the coefficients will not be very large.

[15]There are alternative ways for dealing with missing values in logistic regression, based on imputation or integrating out. However, the proportion of missing values is so high that the necessary estimates might not be reliable. Therefore, we take the safe way, and omit cases with missing values.

[16]We use the function `somers2()` created in `R` (R Development Core Team 2008).

Table 2: β-Coefficients in regression models with different types of concreteness; *** $p<0.001$ ** $p<0.01$ * $p<0.05$ · $p<0.10$.

concreteness	MRC	MANUAL	BOOTS	BOOTS-OBJ	WN-HIER	WN-PHYS
#instances	436	619	546	475	576	576
MRC	0.56 **					
MANUAL		1.77 ***				
BOOTS			-0.58			
BOOTS-OBJ				-0.19		
WN-HIER					-0.09	
WN-PHYS						0.61 *
Th = defin	0.37	0.94 *	0.79 ·	0.75	0.83 ·	0.84 ·
Th = given	1.20 ·	0.85	1.02 ·	1.18 *	1.02 ·	0.83
Th = pron	1.22 ·	0.02	2.44 **	0.97	1.36 *	1.42 *
Rec = defin	-2.39 **	-1.54 *	-1.19 ·	-1.41 ·	-1.32 ·	-1.16 ·
Rec = given	-0.79	-0.49	-0.84 ·	-0.97 ·	-0.96 *	-0.73
Rec = pron	-1.30 *	-1.21 *	-0.35	-0.68	-0.29	-0.61
Rec = anim	-0.55	-0.37	-0.19	-0.18	0.26	0.09
Rec = 1st/2nd	-0.15	-0.63	-0.82	-0.68	-0.92 ·	-0.96 ·
Len dif th-rec	-2.48 ***	-2.53 ***	-2.59 ***	-2.73 ***	-2.65 ***	-2.65 ***
Med = wr	-0.30	-0.58	-0.33	-0.24	-0.43	-0.52
(Intercept)	-1.21	0.71	0.44	1.07	1.05	0.95

terval values (BOOTS and BOOTS-OBJ) show a significant contribution to their models (p=0.327 and p=0.478, respectively). This is rather surprising given the correlation between BOOTS and MRC we found for SEMCOR.

The coefficients for the three significant types of concreteness are positive, meaning that when the theme is (more) concrete, speakers and writers are more likely to choose the prepositional dative construction (*I gave the book to him*), and if it is (more) abstract, the double object construction (*I gave him my love*). This is the same pattern as found by Theijssen (2010). The only true token-based approach, MANUAL, yields the strongest effect in the regression model, with respect to the significance as well as the regression coefficient. Still, despite the different noun levels used – tokens for MANUAL, senses for WN-PHYS and types for MRC – and the different measurement scales – binary for MANUAL and WN-PHYS, and interval for MRC – the effects found are similar. Apparently, the definition of concreteness used and the presence of human intervention have the most influence.

When we look at the effects found for the other features in the models, we see that Length Difference in Table 2 is the most stable, with a highly significant coefficient of -2.7 to -2.5 in all six models. The other effects differ in significance across the different models. The features Discourse Givenness (= *new*) and Pronominality (= *pron*) are correlated[17], which explains the variation in significance across the models: MRC and MANUAL have significant effects for the Pronominality of

[17]We decided not to solve the collinearity by (for instance) combining the features with the help of dimensionality reduction algorithms such as Principle Component Analysis. Instead, we prefer to keep in the original features, being cautious when interpreting the model.

the recipient, BOOTS-OBJ for the Discourse Givenness of the theme and BOOTS, WN-HIER and WN-PHYS for the Pronominality of the theme. WN-HIER also yields a significant effect for the Discourse Givenness of the recipient.

The missing data also seems to have an effect on the significance levels found. In the two bootstrapping and the two WordNet-based approaches, the features for Definiteness have lost significance, and in the MRC model, only the Definiteness of the recipient remains significant.

5 Follow-up experiment

Sections 3 and 4 have shown that the choice for a labelling approach influences the actual labels in the eventual data and consequently the conclusions we can draw in a syntactic study based on this data. Seeing these difference, we need to ask ourselves: To what degree do humans agree about the concreteness of words in context?

5.1 Method

To address this question, we perform a Crowdsourcing experiment on the platform Amazon Mechanical Turk.[18] We ask US-only workers to read a passage, answer a comprehension question and then indicate the concreteness of one of the nouns in the last sentence. They are not given any definition, only the following instruction:

> Each HIT consists of 4 passages of text. For each passage, you have to perform 3 actions:
>
> 1. Read it carefully.
> 2. Answer a comprehension question about the content.
> 3. Indicate how concrete a marked word is, on a scale of 1 (very abstract) to 5 (very concrete).
>
> For instance, consider the following sentence:
>
> `Consecotaleophobia, **fear** of **chopsticks**, was more of a hassle for my Japanese wife than it was for me.`
>
> The *'chopsticks'* are very concrete (5), while *'fear'* is very abstract (1).

The *fear of chopsticks* in the example is deliberately selected so that the 'chopsticks' are concrete in both definitions of concreteness (sensory perceivability and specificity), and 'fear' is abstract in both definitions. In this way, the workers can work out their own definition. Each HIT (Human Intelligence Task) consists of four text passages and is awarded by $0.10 when all four multiple-choice comprehension questions are answered correctly (to prevent cheating). Each HIT is completed by 10 different workers.

The four items in a HIT are selected from our data sets: One item is labelled relatively abstract by all approaches (an 'easy abstract' item), one item is labelled

[18] http://www.mturk.com

relatively concrete by all approaches (an 'easy concrete' item), one has different labellings in the approaches (a 'difficult' item), and one is not covered by the MRC database and/or WordNet, or requires anaphora resolution (a 'special' item). We create 20 HITS from SEMCOR and 20 HITS from DATIVE, leading to a total of 40 HITS, and thus 160 items. An example item from SEMCOR ('easy abstract'):

> Dear Julie. Thank you for your letter of 7 March. It may be difficult to give you a backstage placement during 10-12 April but you are welcome to come in on Friday 12 April to have a look around and meet our technicians. You could also stay and watch the show on Friday evening. On the Thursday, if you wanted to, you could spend a day with the Administration team who will give you a whole ***view*** of how the theatre functions.
>
> When can Julie spend a day with the Administration team?
>
> - on Friday evening
> - on 7 March
> - on the Thursday
>
> Rate the concreteness of ***view***.
>
> - 1 *very abstract*
> - 2
> - 3
> - 4
> - 5 *very concrete*

5.2 Results

For each individual item, scored by 10 different workers, we calculated the average concreteness score, together with the standard deviation. We then took the mean over the items per type and data set, as presented in Table 3.

Table 3: Mean of average score (Av) per item, and mean of standard deviation (Sd) per item. Also provided: number of items per type per data set.

	SEMCOR			DATIVE		
Type	Av	Sd	#items	Av	Sd	#items
easy abstract	1.8	0.8	20	2.0	0.9	20
easy concrete	4.6	0.6	20	4.5	0.6	20
difficult	3.1	1.0	20	3.6	1.0	20
special	2.9	1.0	20	2.6	1.2	20

Table 3 shows that the easy abstract items obtain average scores ≤ 2.0 and the easy concrete items have average ratings ≥ 4.5. The mean standard deviation of the easy abstract items is 0.8 for SEMCOR and 0.9 for DATIVE. It is lower for the easy concrete items: 0.6 for both data sets. The difficult and special items receive mean scores that are closer to the middle (i.e. 3.0), both with mean standard deviations of 1.0 or higher.

Looking at the individual items, we see that eight items received a score of 5 by all ten workers: *bottle*, *hat*, *heels*, *mirror*, *oxen*, *room* (all 'easy concrete' items), *milk* and *oil* (both 'difficult').[19] Items that were assigned average scores of maximally 1.5, and thus were considered rather abstract, were *attitude*, *delight*, *feeling*, *feelings*, *freedom*, *integrity*, *manner*, *uncertainty* (all 'easy abstract'), *heart* and *principle* (both 'difficult'). Some individual items show relatively high standard deviations (1.4 or higher), indicating that the workers disagree about their concreteness: *it*, *Judaism*, *species*, *stick*, *that* ('special' items), *room*, *bit* ('difficult') and *arms* ('easy concrete').

The instructions given to the workers did not include any definition of concreteness. In the 'difficult' category, there are six instances for which the concreteness score assigned by the 'specificity' approach WN-HIER differed greatly from the score given by the three other approaches (using the definition of 'sensory perceivability'): *ice*, *water*, *land*, *film*, *men* and *pond*. In all six cases, the words are (relatively) concrete in the definition of 'sensory perceivability', and relatively abstract in that of 'specificity'. The workers gave these cases average concreteness scores of 4.0 or higher, which means they focussed most on the definition of 'sensory perceivability'.

5.3 Discussion

The results show us two main things: First, many items are easy both for the (semi-)automatic approaches and for humans. Especially the items that are relatively concrete according to the approaches (the 'easy concrete' items) are also clearly concrete for humans (shown by the low mean standard deviation). Second, the items that lead to most disagreement among the workers are mostly also problematic for the (semi-)automatic approaches (they are mostly 'difficult' and 'special' items). In case the concreteness differs in the two definitions, humans seem to prefer the definition of 'sensory perceivability'. Note that it is impossible to say whether middle-range values in the MRC Database are due to disagreement between raters or to the fact that these words denote things that are not intrinsically concrete or abstract.

Our observations indicate that there are items that are so obviously concrete or abstract, that there is (almost) no doubt about their concreteness. There are also many instances for which the concreteness is unclear, to which different persons assign different concreteness values. Still, even if different persons have different opinions about the concreteness of some noun, the *perceived* concreteness of the individual speaker can be a factor in that speaker's syntactic choices. For these cases, averaging over the speakers may lead to loss of this potentially important information. Instead, it may be more appropriate to take into account the differences between individual language users, for example by including the speaker/writer as a random effect. Another possible solution is to treat unclear cases as missing values. In this way, the unclear cases, besides perhaps decreasing the representativeness of the models a little, will not affect the models.

[19] The workers rated the nouns in context, but we present only the nouns for the sake of readability.

6 Summary and conclusion

We have compared different approaches to establish the concreteness of nouns. The approaches differed in the definition used, in the scale of the values that can be assigned (interval, ordinal, nominal), the noun level they take as basis (token, sense or type) and the manner in which the values are assigned (manually, automatically, or semi-automatically). Our goal was two-fold: First, to find out in what way the actual labels of the concreteness of nouns change when using various definitions, or different implementations of the same definition. Second, to discover in what way the conclusions in a syntactic study change, when using these approaches.

With respect to the first goal, the scores assigned to 68,848 nouns in the SemCor Corpus showed considerable variation across the four labelling approaches we employed. The labellings by the only approach that used the definition of 'specificity' instead of 'sensory perceivability', WN-HIER, differed most from those by the other approaches. The bootstrapping approach BOOTS was problematic because at some point in the process, abstract nouns (denoting time and quantity) were included as concrete seeds. Moreover, the lexico-syntactic markers were too sparse: For most of the cases it was necessary to use the fall-back option of looking up the concreteness of the noun *types* in the list of seed nouns, because no concrete or abstract markers were present for the individual noun *token*. The use of the MRC database in the MRC approach was problematic because of its coverage (it was not designed as a tool for annotating words in arbitrary texts). The fact that BOOT and MRC mostly classify noun types, ignoring the word sense and the context, seemed to have no effect. We also failed to find an effect for the measurement scale used and the manner of annotation.

We approached the second goal by taking as a case study the English dative alternation. Using a data set of 619 instances extracted from the ICE-GB corpus, we built several regression models to predict the construction used, each using a different type of concreteness as a feature. The effects of the different types of concreteness varied considerably. Concreteness defined as 'specificity' did not seem to play a role in the choice. When defined as 'sensory perceivability', concreteness only seemed to play a role when the approach included manual input, either making use of the manually established MRC Database (MRC), manually performing word sense disambiguation with the help of the manually designed WordNet hierarchy (WN-PHYS), or manually assigning a value to the noun token itself (MANUAL). Again, we saw that the noun level and the measurement scale used have no clear effect, although the strongest effect was found for the only true token-based approach in the present research: MANUAL.

The results made us wonder to what degree humans agree about the concreteness of words in context. To investigate this, we employed a crowdsourcing experiment in which we asked workers to rate the concreteness of nouns presented in context. The human ratings showed that (also) for humans, there are instances that are clearly concrete or abstract, but also many instances for which humans disagree about the concreteness. In cases where the concreteness differed in the two definitions, people seem to focus most on the definition of 'sensory perceivability'.

Our general conclusion is that results concerning the concreteness in syntactic research can only be interpreted when taking into account two factors: (1) the annotation scheme used and (2) the type of data that is being analysed. With respect to the annotation scheme (factor 1), we saw that the definition used and the presence of human intervention have the strongest effect. The type of data being analysed (factor 2) is relevant mostly because of the coverage issues of the resources we employed (MRC and WordNet), and because of the differences in the concreteness ratings of individual language users.

References

Artstein, Ron and Massimo Poesio (2008), An empirically based system for processing definite descriptions, *Computational Linguistics* **34** (4), pp. 555–596, MIT Press.

Baayen, R. Harald (2008), *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*, Cambridge University Press.

BNC Consortium (2007), *The British National Corpus, version 3 (BNC XML Edition)*, Oxford University Computing Services. http://www.natcorp.ox.ac.uk/.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen (2007), Predicting the Dative Alternation, *in* Bouma, Gerlof, Irene Kraemer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, Royal Netherlands Academy of Science, pp. 69–94.

Changizi, Mark A. (2008), Economically organized hierarchies in WordNet and the Oxford English Dictionary, *Cognitive Systems Research* **9** (3), pp. 214–228, Elsevier.

Coltheart, Max (1981), *MRC Psycholinguistic Database user manual: Version 1*, Birkbeck College.

Fellbaum, Christiane (1998), *WordNet: An Electronic Lexical Database*, MIT.

Garretson, Gregory (2003), Coding manual for the project "optimal typology of determiner phrases", Unpublished manuscript, Boston University.

Geeraerts, Dirk, Gitte Kristiansen, and Yves Peirsman (2010), *Advances in Cognitive Sociolinguistics*, Walter de Gruyter, Berlin, Germany.

Greenbaum, Sidney (1996), *Comparing English Worldwide: The International Corpus of English*, Clarendon Press.

Ide, Nancy and Laurent Romary (2008), Towards international standards for language resources, *in* Dybkjær, Laila, Wolfgang Minker, and Holmer Hemsen, editors, *Evaluation of Text and Speech Systems*, Springer, pp. 69–94.

Kübler, Sandra (2007), How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges, *in* Nicolov, Nicolas, Kalina Boncheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, John Benjamins, pp. 79–88.

Lyons, John (1977), *Semantics*, Vol. 2, Cambridge University Press.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross T. Bunker (1993), A

semantic concordance, *Proceedings of the 3 DARPA Workshop on Human Language Technology*, pp. 303–308.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing.

Schmid, Hans-Jörg (2000), *English abstract nouns as conceptual shells: from corpus to cognition*, Topics in English Linguistics, Mouton de Gruyter.

Spreen, Otfried and Rudolph W. Schulz (1966), Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns, *Journal of Verbal Learning and Verbal Behavior* **5**, pp. 459–468.

Tapanainen, Pasi and Timo Järvinen (1997), A non-projective dependency parser, *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 64–71.

Theijssen, Daphne (2010), Variable selection in Logistic Regression: the British English dative alternation, *in* Icard, Thomas and Reinhard Muskens, editors, *Interfaces: Explorations in Logic, Language and Computation*, Vol. 6211 of *Lecture Notes in Computer Science*, Springer.

Theijssen, Daphne, Hans van Halteren, Karin Fikkers, Frederike Groothoff, Lian van Hoof, Eva van de Sande, Jorieke Tiems, V'eronique Verhagen, and Patrick van der Zande (2009), A regression model for the English benefactive alternation, number, 14 14 in *LOT Occasional Series*, LOT, Utrecht, The Netherlands, pp. 115–130.

Theijssen, Daphne, Lou Boves, Hans van Halteren, and Nelleke Oostdijk (2011), Evaluating automatic annotation: Automatically detecting and enriching instances of the dative alternation, *Language Resources and Evaluation*, Springer.

Thelen, Michael and Ellen Riloff (2002), A bootstrapping method for learning semantic lexicons using extraction pattern contexts, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 214–221.

Xing, Xing, Yi Zhang, and Mei Han (2010), Query difficulty prediction for Contextual Image Retrieval, *Proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010)*.