What good are 'Nominalkomposita' for 'noun compounds': Multilingual Extraction and Structure Analysis of Nominal Compositions using Linguistic Restrictors

Patrick Ziering **Lonneke van der Plas**
Institute for Natural Language Processing
University of Stuttgart, Germany
{Patrick.Ziering, Lonneke.vanderPlas}@ims.uni-stuttgart.de

Abstract

Finding a definition of compoundhood that is cross-lingually valid is a non-trivial task as shown by linguistic literature. We present an iterative method for defining and extracting English noun compounds in a multilingual setting. We show how linguistic criteria can be used to extract compounds automatically and vice versa how the results of this extraction can shed new lights on linguistic theories about compounding. The extracted compound nouns and their multilingual contexts are a rich source that serves several purposes. In an additional case study we show how the database serves to predict the internal structure of tripartite noun compounds using spelling variations across languages, which leads to a precision of over 91%.

1 Introduction

Compounding is a phenomenon that is studied extensively in linguistic literature. Also in computational linguistics, compounds are enjoying more and more attention (Ó Séaghdha, 2008; Hendrickx et al., 2013). Compounding is a very productive word formation. Already 2-year-olds are able to form new words by using compounds consisting of two morphemes (Clark, 1981). As a consequence, compounds are a very common word type but many occur with a very low token count. In an analysis of the German APA corpus, Baroni et al. (2002) found that almost half (47%) of the word types were compounds. At the same time, the compounds accounted for a small portion of the overall token count (7%), which suggests that many of them are rare (83% of the compounds had a corpus frequency of 5 or lower). For English, more than half of the two-noun compounds (e.g., *car park*) in the BNC occur exactly once (Kim and Baldwin, 2006). The high productivity of compounds makes compositional approaches to automatic processing indispensable: listing all possible compounds in a dictionary would be as infeasible as listing all possible adjective-noun combinations. Even for compound nouns that occur 10 times or more in the BNC, static English dictionaries provide only 27% coverage (Tanaka and Baldwin, 2003).

Being abundant as a phenomenon but scarce in terms of individual examples (the combination of high type frequency and low token frequency) makes the analysis of these compound nouns particularly problematic for statistical techniques that need high token frequencies to make accurate predictions. Data sparsity is expected to lead to low performance. However, the correct analysis of compound nouns is important for a number of NLP tasks, for example in machine translation (Bouillon et al., 1992; Rackow et al., 1992; Johnston and Busa, 1999; Navigli et al., 2003). The accurate translation of compounds is non-trivial, because we find a large amount of variation in the way languages deal with compounding. Some languages such as German use closed compounding (i.e., they create one-word compounds, e.g., *Todesstrafe* (*death penalty*)) whereas others do not. In Romance languages, such as French, compounds are not as productive, instead postmodifying prepositional phrases (e.g., *peine de mort*) and adjectives (*peine capitale*) are used to construct complex nominals.

Another challenge in compound translation is due to the fact that the amount of underspecification in compound surface structure varies between languages. For example, whereas English leaves the compound relation (i.e., the semantic relation between two components, e.g., N_2 *made of* N_1 as in *iron door*) covert, in French we find prepositions that correlate with the relation type (Girju, 2007; Celli and Nissim, 2009). *Chocolate cake*, cake **made of** chocolate, is translated with *gateau **au** chocolat*, whereas *wedding cake*, cake **made for** a wedding, is *gateau **de** marriage*.

The first aim of this study is to extract a large database of compounds and their translations in context from a parallel corpus. This database will serve multiple purposes. For example, it will be used to study compounding across different languages, and we will exploit the cross-lingual variation for compound processing. In the second part of this paper, we will show a case study of how the extracted database can be used for analysing the structure of noun phrases, more specifically, we exploit spelling variations across languages for bracketing three-noun compounds (3NCs) such as *air traffic control*, which could be indicated as LEFT bracketing using the German phrase *Kontrolle des Luftverkehrs* (*control of air traffic*).

Compounding is an important subject of study in theoretical linguistics, because it constitutes a continuum from fully compositional to idiosyncratic word formation and is found at the boundary between words and phrases. However, there is virtually no reliable and universally accepted criterion for distinguishing compounds from phrases or other types of word formations, as stated by Lieber and Stekauer (2009). They discuss reasons for the complexity that arises when defining noun compounds, that we will review in the next section. They do, however, also describe a number of linguistic tests, each with their own advantages and drawbacks.

This brings us to the second aim of this paper. We propose an iterative method that, in absence of a clear definition, validates several linguistic tests on corpus data and continuously refines the definition. We show how we use linguistic tests to extract compounds automatically and vice versa how the results of this extraction can shed new lights on linguistic theory about compoundhood. The multilingual nature of our data (we work on parallel corpora) has the additional advantage that a cross-lingual definition can be sought by studying compounds in context and their translation across several languages.

In Section 2, we discuss the problem of defining noun compounds (NCs) as described by Lieber and Stekauer (2009) and present an iterative method for defining and extracting English NCs starting with an initial definition based on some linguistic tests. In Section 3, we present our method for extracting English NCs and their translations to several languages from a parallel corpus using a set of extraction rules. An experimental setup and results are presented in Section 4. In Section 5 we show in a case study of bracketing three-noun compounds how our database serves for exploiting multilingual spelling variations. Section 6 describes related work and finally Section 7 concludes.

2 Iterative method for the definition and extraction of noun compounds

In this section, we outline the controversy of defining compoundhood as described in linguistic literature. We present several linguistic tests for distinguishing compounds and show how we implement some linguistic criteria that can be used for identification and extraction of noun compounds and how these constitute the initial definition.

2.1 Definition of compounds and linguistic criteria

When we seek to find a working definition of noun compounds (NCs), we have to keep in mind that not only the definition but also the existence of such an NC is controversial. Lieber and Stekauer (2009) present a discussion about this controversy sketched below. While Bauer (2003) defines a compound as a “formation of a new lexeme by adjoining two or more lexemes”, Marchand (1967) argues that there is no compounding word formation at all. Instead, he uses the word formation EXPANSION, which combines prefixed words like *reheat* with such as *steamboat* using the criterion of a free head. Lieber and Stekauer (2009) highlight two reasons for the complexity that arises when defining noun compounds. Firstly, in some languages, constituents are not free but stems or roots. For example, the Slovak term *rýchlovlak* (express train) starts with the stem of the adjective *rýchly* (as in the phrase *rýchly vlak* (fast train)). The lack of inflection in English makes compositional and phrasal structures (i.e., *fast train* as phrase or as compound (*express train*)) collapse. Secondly, sometimes phrases and derivations cannot be distinguished from compounds. While *blackboard* (in opposition to a *black board*) can be classified as compound without dissent, a *tomato bowl* that just happens to hold tomatoes might not be regarded as a single lexeme (conforming Bauer's (2003) definition).

So, the only way for getting a suitable definition is to find solid criterions. Although Lieber and Stekauer (2009) come to the conclusion that there is almost no reliable and universally accepted criterion, they mention several plausible tests. Compounds can be identified by prosody. While in the phrase *black bird*, the head (*bird*) is stressed, in the compound *blackbird* the stress is on the first syllable (*black*). A syntactic test mentioned by Lieber and Stekauer (2009) is inseparability, i.e., there must not be any element intervening a compound's components. While *black bird* can be understood as compound, *black ugly bird* is a phrase. Another promising syntactic criterion is the inability to modify the first element (i.e., the modifier) of a compound. In a phrase like *social person*, the first element (*social*) can be modified (i.e., *very social person*). This is not possible for compounds (e.g., very *social policy*). A last syntactic criterion, the inability to replace the second noun of a nominal compound with a proform such as *one* (e.g., *black bird* vs. *black one*), would need human support. A morphological criterion states that in compounds only the head is inflected. Although this assumption does not always hold (as shown in examples like *overseas investor* or *girls club*), this seems to be a promising criterion when investigating inflectional behaviour in aligned languages that show strong morphology, e.g., French. Conversely, determining compoundhood on the basis of spelling is discarded by Lieber and Stekauer (2009). English orthography is highly inconsistent: some compounds usually occur as a closed compound (e.g., *football*), some occur hyphenated and some occur as an open compound (e.g., *waiting room* or *rule of law*). For some compounds, several spellings are possible (e.g., *flowerpot*, *flower-pot*, *flower pot* or *pot of flowers*).

In our study, we focus on written language as given in a parallel corpus. Since we do not have any speech data, we cannot use any phonological features such as stress for the extraction of noun compounds. For the inability to replace the second noun of a nominal compound with a proform, we cannot assess if the meaning of a sentence would have changed (e.g., *We see blackbirds* vs. *We see black ones*).

In this paper, we focus on criteria that are most suitable with the current data. Although Lieber and Stekauer (2009) exclude spelling as a reliable criterion of compoundhood, we take it as starting point. The parallel corpus we use for the extraction includes several languages. Spelling variations between languages can be exploited to find compounds (e.g., *social policy* can be written as one word in German (*Sozialpolitik*)). We account for the English spelling variations by defining part-of-speech (PoS) patterns that cover most plausible spellings. These PoS patterns treat each noun or adjective as a compound's component and thus, this way of extraction inherently implements the criterion of inseparability. We exploit multilingual evidence in terms of cross-lingual differences in spelling to extract compounds. Diverse language families have different declinations of forming a closed compound. While languages like Danish and German prefer closed compounding, English and Romance languages like Spanish use open compounds. It is this spelling variation that we base our first set of extraction rules on with the aim of having a set of English NCs and their translations in up to 9 European languages. We will show that cross-lingual closed compounding is a promising feature for extracting English NCs.

The inability to modify the first element of a compound seems to be a promising test. Since there are many linguistic factors that have to be taken into account (e.g., morphological agreement in gender, number or case), we plan to include this criterion for several languages and any combination of contextual modifier and potential noun compound. We will implement this and further morphological criteria in future work.

2.2 Initial definition for compound extraction

With a focus on multilingual validity, we adapt the definition of Bauer (2003) to our multilingual setting. Inspired by Behagel's (1909) First Law ("Elements that belong close together intellectually will also be placed close together"), we associate a closed compounding language realising an English word sequence as a closed NC with an indicator for compoundhood:

Initial definition: *A noun compound is a nominal composition of several lexemes that are represented as a one-word expression in some of the languages studied.*

This definition covers both target single words (e.g., *blackbird* translates to German as *Amsel*) or target closed compounds (e.g., *football match* translates to Dutch as *voetbalwedstrijd*).

We are aware of the fact that this definition leads to some controversial cases for English word sequences including pre-nominal adjectives. While some of them are commonly accepted such as *social policy* (German: *Sozialpolitik*), others are less accepted such as *strong wind* (German: *Starkwind*) or *small car* (German: *Kleinwagen*). This is not an unwanted side-effect. On the contrary, these controversial cases are an essential part of the iterative process we described, as they will foster linguistic discussions. Although the German *Starkwind* can be regarded as partly compositional, it is frequently used with a concrete definition (in contrast to the phrase *starker Wind*) and cannot occur in a context violating this definition, as shown in the table below.

1 a)	Als Starkwind wird meist eine Windstärke zwischen 6 und 7 Beaufort bezeichnet.
1 b)	A {strong wind} usually refers to a wind force of 6-7 Beaufort.
2 a)	Am Samstag weht ein starker Wind mit Windstärke 8 von Westen.
2 b)	On Saturday, a strong wind with wind force 8 will blow from the west.
3 a)	*Am Samstag weht ein Starkwind mit Windstärke 8 von Westen.
3 b)	On Saturday, a *{strong wind} with wind force 8 will blow from the west.

3 Multilingual extraction of NCs

This method is based on the initial definition for compound extraction described in Section 2.2 and can be adapted in succeeding iterations. English NCs are extracted from a parallel corpus that includes English and some closed compounding languages (e.g., German).

3.1 Preprocessing the parallel data

In Section 4.1, we describe the tokenization, sentence alignment, word alignment and PoS tagging we apply to the parallel data in more detail. In addition, we perform a binary compound splitter on each word that is tagged as a noun by following a variant of the methods of Stymne et al., (2013). This unsupervised splitter checks each noun for all possible segmentations into at most two components with at least two characters. All possible segmentations are scored with the geometric mean of the components' frequencies in the parallel corpus. The highest-scored segmentation (possibly with no split point) is used.

3.2 Preselection of English noun compounds using PoS patterns

As a basis for the extraction of English NCs, we use a set of possible English PoS sequences that can constitute an NC. These PoS patterns account for the various ways of composing English NCs and for the inseparability property as described in Section 2.1. Table 1 lists all plausible PoS patterns for bipartite and tripartite NCs with some examples (cf. the Penn Treebank tag set (Marcus et al., 1993)). For all examples in Table 1, we found translations to a closed compound in German, which satisfies our initial definition described in Section 2.2, e.g., *overall recovery rate* has been translated to *Gesamtrückforderungsquote*. Although the larger the number of components, the sparser the number of (correct) extractions, we create a regular expression for PoS patterns that cover English NCs with n components (where $2 \leq n \leq 10$). This regular expression combines all possible combinations of observed NC types. In the next step, we will filter noise, that occurs mostly in longer word sequences.

3.3 Noise filters

The selection of English NCs and their translations is based on automatic preprocessing, which leads to some noise due to false PoS tags or flaws in word alignment. With increasing word sequence length, the amount of noise increases. We apply several filters on each preselected NC and on their alignments to all other languages in the corpus and keep only those that pass all filters.

3.3.1 PoS filters

1. Two filters are applied to all languages: we disqualify word sequences including nouns or adjectives that (1) consist of only one character or (2) are contained in a stop list[1].

[1] ranks.nl/stopwords

PoS pattern	Example
Bipartite noun compounds	
NN	marketplace
NN NN	death penalty
JJ NN	structural policy
NN POS NN	children’s development
NN IN NN	fall in population
NN IN DT NN	concussion of the brain
Tripartite noun compounds	
NN NN NN	energy security goal
JJ NN NN	overall recovery rate
NN IN NN IN NN	income per head of population
Regular expression for 2–10 components	
NN ((IN (DT)?|POS))? NN){1,9}	greenhouse gas emission allowance trading scheme
JJ NN ((IN (DT)?|POS))? (JJ)? NN){1,8}	internal energy market package

Table 1: English PoS sequences for noun compounds

2. Then, to account for PoS tagging errors in English, we collect all words and their PoS tags in the parallel corpus. For each word, we compute the probability of being tagged as a noun or adjective as given in (1).

$$P(noun/adj \mid word) = \frac{f((noun \cup adj) \cap word)}{f(word)} \tag{1}$$

   We disqualify English word sequences, if they contain a noun or adjective w with $P(noun/adj \mid w) < \theta$. After testing several values for θ, we have decided to choose $\theta = 0.15$ because it has turned out to be a promising trade-off between coverage and precision (e.g., accepting words like *human* but rejecting words like *anywhere*).

3.3.2 Word alignment filter

Shortcomings in word alignment quality are remedied with three word alignment filters.

1. We truncate extraneous words (i.e., determiners, prepositions and (ad)verbs) from the border of the word sequence (adjectives are removed from the right border for Germanic languages and from the left border for Romance languages).

2. We disqualify the word sequence as being phrasal if it contains two consecutive nouns with verbs or adjectives in between or if the nouns are more than ϕ tokens apart from each other. When analysing many instances of Romance phrases aligned to an English noun compound, we observed that $\phi = 3$ is the maximum token distance two nominal components can be apart (usually separated by preposition or preposition+determiner). If the word sequence is qualified as phrasal, we add determiners and prepositions that occur in the context between the nouns, otherwise the word sequence remains unchanged.

3. We remove the word sequence if it does not contain at least one noun.

The resulting set of English word sequences that conform to the regular expression in Table 1 and their aligned and filtered word sequences are stored as a set of m-tuples of word sequences. Subsequently, we will refer to this set as the *basic set*. The basic set still contains English word sequences that do not comply with our initial definition for compound extraction (Section 2.2), i.e., that are not aligned to a closed compound. In the next step, we apply a restrictor to all NCs in the basic set and keep only those instances that pass the restrictor.

3.4 Closed compound restrictor

An English word sequence is considered to be an NC if it is represented as a one-word expression in some of the closed compounding languages (e.g., Dutch, German, Swedish, ...). Given a parallel corpus with $n > 1$ closed compounding languages, this definition leaves space for investigating the degree of cross-lingual closed compounding (deg_{closed}) which is necessary for optimal extraction quality (i.e., optimal precision and recall). Because the rules described in Section 3.3.2 still leave some word alignment errors (i.e., English word sequences that are aligned to only a part of the true translation), a single compounding language realising the English word sequence as one word (i.e., $deg_{closed} = 1$) might not be restrictive enough.

The closed compound restrictor with $deg_{closed} \geq i$ retains only English word sequences that are aligned to at least i one-word expressions in the aligned closed compounding languages. We will refer to this restrictor as $CCR(i)$ and to the resulting data set as $closed\ compound(i)$.

4 Experiments for NC Extraction

4.1 Setup

Data and preprocessing. We use the 7th release of the Europarl corpus[2]. Although the Europarl corpus comprises 21 European languages, the amount of common data they cover is rather small. This means, the more languages we use, the smaller the amount of common data. In order to get a good trade-off between cross-lingual coverage and language variation exploitation, we decided on a set of 10 languages: English, the closed compounding languages Danish, Dutch, German and Swedish, as well as Greek and the Romance languages French, Italian, Portuguese and Spanish. Instead of preprocessing the parallel corpus on our own, we exploit the already preprocessed Europarl resource of OPUS[3] (Tiedemann, 2012). This preprocessed resource is PoS tagged using TreeTagger (Schmid, 1995) for English, Dutch, German, French, Italian and Spanish and the Hunpos[4] tagger for Danish, Portuguese and Swedish. We additionally tagged the Greek data using the MATE[5] tagger. The sentence alignment provided by OPUS is restricted to language pairs. As we need a sentence representation that is parallel in all 10 languages, we apply the OPUS sentence aligner (with English as pivot) on our language set and extract a total of 884,164 parallel sentence representations. The word alignment information provided by OPUS was also based on language pairs. This means, the sentence-wise token indices has to be adapted to our updated sentence representation (which is different due to a larger language set). In OPUS, the word alignment tool GIZA++ (Och and Ney, 2003) has been used with the symmetrisation heuristics (grow-diag-final-and (Koehn et al., 2007)).

4.2 Evaluation procedure and scoring

In order to compare the added value in terms of recall and precision of each closed compound restrictor (i.e., $CCR(1)$ to $CCR(4)$), we randomly select 50 accepted and 50 rejected English word sequences for each restrictor. We rate the correctness of acceptance and rejection and compute precision and recall as given in (2) and (3). F-Score is defined as harmonic mean of precision and recall.

$$Precision = \frac{accepted \cap correct}{accepted} \tag{2}$$

$$Recall = \frac{accepted \cap correct}{(accepted \cap correct) \cup (rejected \cap incorrect)} \tag{3}$$

The precision of the basic set is measured as the accuracy of a 50 sample subset. We do not compute recall and F-Score for the basic set.

[2] statmt.org/europarl

[3] opus.lingfil.uu.se

[4] code.google.com/p/hunpos/downloads/list

[5] code.google.com/p/mate-tools

We measure the amount of closed NCs in a given closed compounding language (ccl) and for a given set of NCs (Set) by using the frequency of closed NCs relative to the number of all word sequences ($N_{Set,cll}$) (word sequences removed in Section 3.3 are excluded). Since the alignment to single words is still somewhat noisy (i.e., our compound splitter does not work error-free and there are still deficiencies in the word alignment), we select a set of 50 closed noun compound samples and rate the accuracy. The final amount of closed NCs is the product of relative frequency and accuracy, as given in (4).

$$p_{ccl}(Set) = \frac{f_{Set}(closed\ NC)}{N_{Set,cll}} \cdot acc_{ccl}(Set) \tag{4}$$

4.3 Results

Set	Size	Precision	Recall	F-Score	p_{en}
Basic set	3,178,661	38.0%	—	—	1.5%
closed compound (1)	795,518	84.0%	71.2%	77.1%	4.7%
closed compound (2)	495,837	92.0%	74.2%	82.1%	6.6%
closed compound (3)	316,330	98.0%	65.3%	78.4%	9.2%
closed compound (4)	143,121	98.0%	63.6%	77.2%	10.4%

Table 2: Extraction quality of the basic set after restrictor application

Table 2 shows the results when applying the four different degrees of the closed compound restrictor to the basic set. The first result is that using only a PoS-based method leads to a very poor extraction accuracy (38%). For the applications of the closed compound restrictors, the result is that increasing deg_{closed} means increasing precision but decreasing recall in NC extraction. The reason for this is that an aligned closed NC is generally a sufficient condition for an English NC (except for controversial cases such as *strong wind*) but not a necessary condition (i.e., a true English NC may be aligned to only periphrastic constructions). The highest F-Score (82.1%) is achieved using $CCR(2)$. We can conclude that the closed compound restrictor is a reliable method for extracting English NCs. In future work, we will use a large set of human annotators with different backgrounds in order to get a widely distributed sense of compoundhood. Moreover, instead of a binary rating, we will consider compoundhood as a continuum and compare rating scores with the amount of aligned closed compounding languages realising a closed compound in a larger parallel corpus.

The last column in Table 2 shows the amount of closed English NCs in each respective set. Since deg_{closed} correlates with the amount of closed English NCs, we can conclude that, despite the cross-lingual differences in spelling conventions attested in linguistic literature, there is a bias for a universal consensus in closed compounding.

Language	p_{ccl}
German	71.2%
Danish	63.3%
Swedish	62.2%
Dutch	58.7%

Table 3: The amounts of closed noun compounds

Table 3 shows the amounts of closed noun compounds in the closed compounding languages Danish, Dutch, German and Swedish, extracted from the closed compound (1) set. Our result shows that German is the most productive language in closed compounding (71.2%), while the other languages have a similar productivity (58-63%).

The result of our extraction method is a database of English NCs and their translations in up to 9 European languages. As described in the introduction, this database will serve several purposes. One is to study cross-lingual variation. Table 4 shows some examples of multilingual noun compound extractions from closed compound (2).

English	German	Dutch	French	Italian
automotive sector	Automobilmarkt	automobielsector	secteur automobile	mercato dell' automobile
fishing techniques	Fischfangtechniken	visserijmethoden	techniques de pêche	tecniche di pesca
timetable	Zeitplan	tijdschema	calendrier	calendario
highways	Autobahnen	snelwegen	autoroutes	autostrade
trading system	Handelssystem	handelsbestel	système commercial	sistema di scambi

Table 4: Examples of multilingual noun compounds

The examples show that English noun compounds have various realisations in European languages. Although French and Italian are open compounding languages, we do find closed compounding (e.g., *autoroutes*). Compounds such as *timetable* can also be aligned to single nouns such as *calendrier* (calendar). We found three common word formation types in Romance languages for bipartite noun compounds: (1) two nouns and a preposition in between, (2) one noun and a post-nominal adjective and (3) a single (possibly compounding) noun. Although Romance languages usually agree with respect to the word formation type, they may disagree as is the case for French and Italian for the example concerning *trading system*. One interesting observation is that while the head of *highways* (*ways*) is translated fairly literally, the modifier (*high*) is replaced by alternative aspects. On highways, cars (***Auto****bahnen* (*car-ways*)) usually drive fast (***snel****wegen* (*fast-ways*)). In future work, we will use this database for researching the nature of compoundhood in a cross-lingual perspective. The resource is publicly available for future research[6].

5 Bracketing three-noun compounds

In this section, we show a case study of how our extracted database can be used to predict the structure of NPs, more specifically to bracket tripartite noun compounds (3NCs), i.e., a composition of three bare nouns that function as one unit. Given a 3NC, we can either have RIGHT bracketing, as in *baby [bicycle seat]*, or LEFT bracketing, as in *[human rights] abuses*.

5.1 The cross-lingual bracketing method

We first start with six phrase patterns that correspond to foreign phrases that are aligned to an English 3NC, as shown in Table 5, where SN refers to a single (non-compounding) noun, FC refers to a functional context (i.e., a sequence of functional words), ADJ refers to an adjective and CNC refers to a closed (bipartite) NC (based on the splitter described in Section 3.1). Each phrase pattern contains a complex unit that is separated from the rest, e.g., a closed NC or a combination of adjective and single noun. For each pattern, we know what is the head and what is the modifier: the first phrase pattern contains only one nominal component, that can be identified as head. For the other patterns, the order is: head, FC, modifier. Based on the assumption that the aligned head corresponds to the English head, we can infer the English bracketing from the complexity of the aligned head. If the aligned head is the complex unit, the English bracketing label is RIGHT, otherwise LEFT. The third column in Table 5 shows the inferred labels for the English 3NC based on the foreign phrase pattern. For an English 3NC, we check all aligned languages for a matching phrase pattern and collect, in the case of a match, the inferred label. The majority label determines the final bracketing label.

The examples below illustrate instances for each phrase pattern, where the indices correspond to those in Table 5.

[6] www.ims.uni-stuttgart.de

	Phrase pattern in foreign language	Label for English 3NC
(1)	ADJ CNC	RIGHT
(2)	CNC FC SN	RIGHT
(3)	SN FC CNC	LEFT
(4)	SN FC ADJ SN	LEFT
(5)	ADJ SN FC SN	RIGHT
(6)	SN ADJ FC SN	RIGHT

Table 5: Phrase pattern and inferred label

(1) de: *staatliche Steueraufsichtsbehörden*
state {tax inspectorates}
”state tax inspectorates”

(2) de: *Absatzmarkt für Fahrzeuge*
{sales market} for vehicles
”car sales market”

(3) nl: *methode voor geboortebeperking*
method for {birth control}
”birth control method”

(4) sv: *brottet mot mänskliga rättigheterna*
abuses of {human rights}
”human rights abuses”

(5) da: *gennemsnitlige overførsel af data*
{average transfer} of data
”data transfer rate”

(6) es: *consumo final de energía*
{consumption final} of energy
”energy end consumption”

We observed that the initial assumption (saying that the aligned head corresponds to the English head) is not always true. Sometimes the English head and modifier are swapped in aligned languages, as illustrated in example (7).

(7) nl: *stabiele wisselkoersen*
stable {exchange rate}
”exchange rate stability”

To solve this problem, we inspect the word alignment from the phrase pattern of language l_j to the English nouns N_1, N_2 and N_3 in a 3NC. If the complex unit is aligned to $\{N_2, N_3\}$ or to $\{N_1, N_3\}$, l_j provides the label RIGHT. If the complex unit is aligned to $\{N_1, N_2\}$, l_j votes for LEFT. If the complex unit is aligned to all three nouns, this is an indicator for a word alignment error. In this case, l_j will not perform any prediction. In all other cases, the inferred label from the phrase pattern is used.

5.2 Evaluation for cross-lingual bracketing

As there are only two possible structures for 3NCs, namely LEFT or RIGHT branching, we regard this task as a binary classification and score the accuracy of class agreement. As basis, we use the basic set created in Section 3, because alignments to closed compounds are not of interest for the bracketing task. Two trained human annotators (of which one is one of the authors) individually bracket a sample of 100 randomly selected 3NCs in context. Contextual cues can help the annotator to disambiguate the structure of the English NC, so the accompanying sentences are shown to the annotator. The annotators are no domain experts and since terms in Europarl can be quite domain specific, they are allowed to look up the meaning of the constituents in a dictionary or check Google. Annotators are asked to label 3NCs as LEFT or RIGHT, or UNDECIDED if they are unclear. Furthermore, the annotators are asked to mark extraction errors. When inspecting the inter-annotator agreement for the bracketing classes (LEFT/RIGHT; i.e., 76 of 100 samples), we achieved an agreement rate of 89% and $\kappa = 0.693$ (Cohen, 1960), which means substantial agreement (Landis and Koch, 1977). Afterwards, the annotators discuss disagreements and revise their annotations. This has led to a perfect agreement in our setting. The 8 UNDECIDED labellings show that in some cases the bracketing remains ambiguous even in context. In future work, we would like to investigate if larger contexts or domain knowledge is necessary for the disambiguation process or if the NCs are inherently flat (i.e., if LEFT or RIGHT bracketing does not make any difference in meaning). We evaluate our cross-lingual bracketing system for (1) inferred label of a phrase pattern and (2) word alignment information for phrase pattern with inferred label as back-off. We compare the bracketing performance against the LEFT class baseline.

5.3 Results

Method	Accuracy
LEFT baseline	71.1 %
Inferred phrase pattern labels	$89.0^{\dagger}$ %
Word alignment for phrase patterns	$91.6^{\dagger}$ %

Table 6: Bracketing performance; † indicates significantly higher than the LEFT baseline

Table 6 shows the results of our system compared to the LEFT class baseline. The first result is that both inferred label and word alignment information for phrase pattern outperform the LEFT class baseline significantly[7]. Bracketing with word alignment information for phrase pattern outperforms bracketing based on the inferred labels.

6 Related Work

Our methods for extracting and structuring English NCs rely on the spelling of various aligned languages. Previous work on multilingual extraction include Morin and Daille (2010) and Weller and Heid (2012). These type-based approaches focus on bilingual terminology extraction using comparable corpora. Our token-based extraction method includes 10 languages and we extract both the NCs and their context. While the aforementioned work serves as resource for improving machine translation (MT) systems, we focus on NC research and how multilingual evidence can help analysing and interpreting English NCs.

This multilingual perspective on a considerable number of languages has been adopted as well by Macherey et al., (2011), who present a multilingual language-independent approach to compound splitting. Moreover, they learned morphological operations on compounding automatically. Here, Macherey et al., (2011) extract training instances using a method related to Garera and Yarowsky (2008): select a single word f in a language l translated to several English words e_i . If there is a translation for each e_i to a word g_i that shows a (partial) substring match with f, $(f; e_1, \ldots, e_n; g_1, \ldots, g_n)$ is extracted. While Macherey et al., (2011) extract training instances type-based in a bilingual setting, we directly extract NC instances with a set of four closed compounding languages. This token-based perspective has the advantage that we can process English NCs for which there is no literal translation to the target language (e.g., *health insurance* aligned to *Krankenversicherung* (lit. invalid insurance)).

In cross-lingual annotation transfer (Yarowsky and Ngai, 2001; Padó, 2007; Van der Plas et al., 2011) human annotations are transferred from one language to the other in parallel data. In this paper, we use the structural differences between languages as found in parallel corpora to generate annotations on the target language and do not rely on annotations on the source language.

Bracketing methods for both three-noun compounds and complete base NPs have been designed both supervised and unsupervised. Vadas and Curran (2007) used a supervised bracketing method on manually annotated data. Pitler et al. (2010) used the data from Vadas and Curran (2007) for a parser applicable on base NPs of any length including coordinations. Their supervised classifier exploited web-scale N-grams. Although supervised methods outperform unsupervised methods by far, the need for annotated data is a drawback of supervised approaches. Bergsma et al. (2011) used crosslingual data as additional supervision to make the need for manual annotations less pressing. Unsupervised methods use N-gram statistics (Marcus, 1980; Lauer, 1995; Nakov and Hearst, 2005) or semantic information (Kim and Baldwin, 2013).

7 Conclusion

In this paper, we discussed the complexity related to the definition of compoundhood and presented an iterative method that tries to refine existing definitions by tentatively demonstrating the efficacy of

[7] Approximate randomization test (Yeh, 2000), $p < 5\%$

linguistic criteria on corpus data. The initial implementation of two linguistic criteria, based on cross-lingual spelling conventions and the inseparability of a compound's components, achieved an F-Score of 82.1% on the task of extracting English compounds.

The extracted multilingual database of compounds in contexts serves multiple purposes. For example, it can be used to study cross-lingual variations in compounding. We showed in an additional experiment how the cross-lingual evidence found in the multilingual database can be used to bracket English three-noun compounds using cross-lingual spelling variation with a set of six phrase patterns. We achieved a bracketing accuracy of 91.6% that is very close to human performance.

In future work, we plan to continue refining the definition of compoundhood in a cross-lingual setting. We will experiment with additional linguistic criteria defined over multiple languages. This way, we hope to improve the quality of the multilingual database that we will further explore for compound analysis and translation.

Acknowledgements

This research was funded and supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) as part of the SFB 732. We thank the anonymous reviewers for their comments. We also thank Gianina Iordachioaia for her helpful input and interesting discussion.

References

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Predicting the components of German nominal compounds. In *Proceedings of ECAI*, pages 470–474, Lyon. IOS Press.

L. Bauer. 2003. *Introducing Linguistic Morphology*. Introducing Linguistic Morphology. Edinburgh University Press.

Otto Behaghel. 1909. Beziehungen zwischen umfang und reihenfolge von satzgliedern. *Indogermanische Forschungen*, page 110142.

Shane Bergsma, David Yarowsky, and Kenneth Church. 2011. Using large monolingual and bilingual corpora to improve coordination disambiguation. In *ACL-HLT 2011*, pages 1346–1355.

Pierrette Bouillon, Katharina Boesefeldt, and Graham Russell. 1992. Compound Nouns in a Unification-Based MT System. In *ANLP 1992*, pages 209–215, Trento.

Fabio Celli and Malvina Nissim. 2009. Automatic identification of semantic relations in Italian complex nominals. In *IWCS 2009*, pages 45–60.

Eve V. Clark. 1981. Lexical innovations: How children learn to create new words. In Werner Deutsch, editor, *The Child's Construction of Language*, pages 299–328. Academic Press, New York.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1).

Nikesh Garera and David Yarowsky. 2008. Translating compounds by learning component gloss translation models via multiple languages. In *IJCNLP*, pages 403–410.

Roxana Girju. 2007. Improving the Interpretation of Noun Phrases with Cross-linguistic Information. In *ACL 2007*, pages 568–575.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. Semeval-2013 task 4: Free paraphrases of noun compounds. In *Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143.

Michael Johnston and Frederica Busa. 1999. Qualia structure and the compositional interpretation of compounds. In *E. Viegas (ed.), Breadth and depth of semantics lexicons*, pages 167–187. Dordrecht: Kluwer Academic.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *ACL 2006*, pages 491–498.

Su Nam Kim and Timothy Baldwin. 2013. A lexical semantic approach to interpreting and bracketing english noun compounds. *Natural Language Engineering*, 19(3):385–407.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL - Interactive Poster and Demonstration Sessions 2007*, pages 177–180.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Macquarie University.

R. Lieber and P. Stekauer. 2009. *The Oxford Handbook of Compounding*. Oxford Handbooks in Linguistics. OUP Oxford.

Klaus Macherey, Andrew M. Dai, David Talbot, Ashok C. Popat, and Franz Och. 2011. Language-independent compound splitting with morphological operations. In *ACL-HLT 2011*.

Hans Marchand. 1967. Expansion, transposition, and derivation. *La Linguistique*, pages 13–26.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June.

Mitchell Marcus. 1980. *A Theory of Syntactic Recognition for Natural Language*. MIT Press.

Emmanuel Morin and Batrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44:79–95.

Preslav Nakov and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CoNLL 2005, pages 17–24.

Roberto Navigli, Paola Velardi, and Aldo Gangemi. 2003. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1):22–31.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*

Diarmuid Ó Séaghdha. 2008. *Learning compound noun semantics*. Ph.D. thesis, Computer Laboratory, University of Cambridge.

S. Padó. 2007. *Cross-lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Saarland University.

Emily Pitler, Shane Bergsma, Dekang Lin, and Kenneth Ward Church. 2010. Using web-scale n-grams to improve base np parsing performance. In *COLING 2010*, pages 886–894.

Ulrike Rackow, Ido Dagan, and Ulrike Schwall. 1992. Automatic translation of noun compounds. In *COLING 1992*, pages 1249–1253.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *ACL SIGDAT-Workshop 1995*, pages 47–50.

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation: A feasibility study on shallow processing. In *ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, page 1724.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC 2012*.

David Vadas and James R. Curran. 2007. Large-scale supervised models for noun phrase bracketing. In *PACLING 2007*, pages 104–112.

L. Van der Plas, P. Merlo, and J. Henderson. 2011. Scaling up cross-lingual semantic annotation transfer. In *ACL-HLT 2011*.

Marion Weller and Ulrich Heid. 2012. Analyzing and aligning german compound nouns. In *LREC 2012*, Istanbul, Turkey.

D. Yarowsky and G. Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL 2001*, pages 1–8.

A. Yeh. 2000. More accurate tests for the statistical significance of result differences. In *COLING 2000*.