

---

# Automatic Identification of Novel Metaphoric Expressions

---

Automatische Identifizierung neuartiger Metaphern

Diplomarbeit von Erik-Lân Do Dinh

June 18, 2013

---



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT



*Fachbereich*  
**Mathematik**



UBIQUITOUS  
KNOWLEDGE  
PROCESSING

---

Automatic Identification of Novel Metaphoric Expressions  
Automatische Identifizierung neuartiger Metaphern

vorgelegte Diplomarbeit von Erik-Lân Do Dinh

Supervisor: Prof. Dr. Iryna Gurevych, Prof. Dr. Michael Kohler  
Coordinator: Richard Eckart de Castilho

Tag der Einreichung:

---

# Erklärung zur Diplomarbeit

Hiermit versichere ich, die vorliegende Diplomarbeit ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 18. Juni 2013

---

(Erik-Lân Do Dinh)

---

---

---

## Abstract

---

Manually annotating novel metaphors in philosophical and historical texts is a difficult and time-consuming task, so any automated method would be welcome. In this work, we implement such a method for novel metaphor identification. The approach uses Gaussian mixture models, the parameters of which are estimated by expectation maximization. We seek to improve the results of this baseline by incorporating selectional preferences, whose violation can be indicative for metaphorical use of a term or phrase. As we intend to find only novel metaphorical expressions, we further refine the results by employing a large diachronic n-gram corpus. We find that incorporating selectional preferences and additional n-gram novelty filtering significantly improves on the baseline results. As an example application, the resulting classifications will then be presented in a web-based tool.

---

---

## Zusammenfassung

---

Die manuelle Annotation von neuartigen Metaphern in philosophischen und historischen Texten ist eine schwierige und aufwendige Aufgabe. In dieser Arbeit implementieren wir eine unüberwachte Methode zur automatischen Identifizierung von neuartigen Metaphern. Für diesen Zweck werden Gaußsche Mischverteilungen eingesetzt, deren Parameter mittels des Expectation Maximization Verfahrens ermittelt werden. Zur Verbesserung dieser Baseline werden selektionale Präferenzen ausgewertet, deren Verletzung auf mögliche metaphorische Verwendung hindeuten kann. Da speziell neuartige metaphorische Ausdrücke Gegenstand der Suche sind, verschärfen wir diese mit Hilfe eines großen diachronischen N-Gramm Korpus. Die Auswertung der implementierten Methode zeigt, dass sich die Ergebnisse der Baseline durch Einbeziehung selektionaler Präferenzen und N-Gram Filter stark verbessern. Als Beispielanwendung werden die Ergebnisse der Identifizierungsaufgabe schließlich in einem webbasierten Werkzeug präsentiert.

---

---

---

## Contents

---

<b>1</b>	<b>Introduction</b>	<b>6</b>
1.1	Metaphors . . . . .	7
1.1.1	Interaction Theory . . . . .	7
1.1.2	Conceptual Metaphors . . . . .	8
1.1.3	Operationalized definition . . . . .	9
1.2	Related work . . . . .	10
1.2.1	CorMet . . . . .	10
1.2.2	Metaphor identification through verb and noun clustering . . . . .	11
<b>2</b>	<b>Preparations</b>	<b>13</b>
2.1	Data set . . . . .	13
2.2	Candidates . . . . .	14
<b>3</b>	<b>Identification</b>	<b>16</b>
3.1	Clustering . . . . .	16
3.1.1	Gaussian mixture models . . . . .	16
3.1.2	Expectation maximization . . . . .	17
3.2	Creating a baseline . . . . .	22
3.2.1	Calculating semantic similarity . . . . .	23
3.2.2	deWaC . . . . .	24
3.2.3	Lucene . . . . .	25
3.3	Building a repository of selectional preferences . . . . .	26
3.3.1	Selectional preferences . . . . .	26
3.3.2	Resources . . . . .	27
3.3.3	Creation process . . . . .	29
3.3.4	Shortcomings . . . . .	31
3.4	Novelty . . . . .	32
3.4.1	Extracting an n-gram repository . . . . .	32
<b>4</b>	<b>Evaluation</b>	<b>33</b>
4.1	Automatic evaluation . . . . .	33
4.1.1	Inter-Annotator Agreement . . . . .	33
4.1.2	Evaluation results . . . . .	35
4.2	Export and Manual Review . . . . .	38
4.2.1	CSniper . . . . .	38
4.2.2	Integrating manual metaphor classification results into CSniper . . . . .	39

---

5	Summary	42
5.1	Outlook . . . . .	42
5.2	Summary . . . . .	43
	Appendices	45
A	Manual structural categorization of metaphors	46
B	Derivation of $\Sigma_j^{(t+1)}$	53
C	Confusion matrices for Kappa calculation	55
D	How to use the system	56
	List of Figures	58
	List of Tables	59
	Bibliography	60

---

## 1 Introduction

---

Metaphors have been studied in great detail by linguists and philosophers alike; especially in the last century there has gone much effort into showing that metaphors are not only a stylistic device, but rather a means to conceptualize our lives. Since it is argued that metaphors require not only knowledge of the language, but also an understanding of the cultural background they originate from, automatic metaphor identification — let alone interpretation — is no simple task. It becomes even more complicated when we want to distinguish between various levels of *metaphoricity*; some of which are rather clearly defined (e.g. *dead* metaphors like “table leg”) while others do not necessarily have such distinct borders (cf. conventional and novel metaphors, section 1.1.2).

In this work we build a system to identify *novel metaphors*, essentially metaphors which are not widely used. Starting with an overview over different ideas of metaphor, namely the *interaction theory* proposed by Black (1955) and the *conceptual metaphor theory* introduced by Lakoff and Johnson (2003), we then define what a novel metaphor means in the context of this work. We present some existing systems to extract metaphors from texts, and analyze if certain aspects of these approaches can contribute to finding novel metaphors as defined here.

To identify metaphors in text, we first extract *candidates* by hand-written grammar based rules, which have been created after a manual analysis of (metaphor-) annotated documents. These candidates are assessed in a later step by an unsupervised method. We use a collection of manually annotated historical and philosophical German texts from late 18th to early 20th century to evaluate our system.

The core of the metaphor identification system is a clustering algorithm which employs *Gaussian mixture models*, the parameters of which are trained by the *expectation maximization* algorithm. This method, proposed by Li and Sporleder (2010) to identify non-literal use of possible idiomatic expressions, serves as a baseline to classify the candidates. It is augmented with information from large corpora to recognize violations of selectional preferences. Such violations can indicate metaphorical use of words or phrases (cf. Wilks (1978, pp. 197-223), Manning and Schütze (1999, p. 288)). An n-gram frequency based method is employed to strengthen the system’s assessment regarding metaphor novelty.

For evaluation, we first use crossvalidation on a training set for feature selection. The performance of the system is then evaluated on a test set. As a baseline we use the Gaussian mixture model clustering without selectional preference or n-gram information. We find that this baseline performs rather poorly on both types of extracted candidates in terms of precision, but that the results can be significantly improved by using the selectional preferences as features and for filtering. For noun - genitive noun modifier candidates we improve the results by employing the n-gram based novelty filter.



---

We also show an example presentation of the identified metaphors; the found candidates along with their metaphoricity assessment are exported so they can be included in *CSniper*, a web-based tool for multi-user evaluation and assessment.

---

## 1.1 Metaphors

---

When we want to identify metaphors, we first have to agree on which *concept* of metaphor we are talking about. Since there are many different ways metaphors can be — and have been — defined in various fields in linguistics and philosophy, only a short selection of important concepts will be presented, followed by what we will call our *working definition*.

---

### 1.1.1 Interaction Theory

---

In “Metaphors” (Black, 1955), American philosopher Max Black discusses two different views of metaphors, before establishing what he calls an *interaction view* of metaphor. In the process, Black incorporates and builds upon ideas by I.A. Richards (1936), especially Richards’ understanding of an inherent *tension* between terms of a metaphor.

Starting with a rather structural description of metaphors, Black introduces the concept of *focus* and *frame* to compartmentalize a metaphor. A word or expression that is used metaphorically is labeled *focus*, while the rest of the sentence is called *frame* (or as it may be called in this work, the *context*). One important takeaway of this construction is that a metaphor consists not only of a metaphorically used component, but that the context in which an expression occurs is equally important; as Black expresses,

“[...] the presence of one frame can result in metaphorical use of the complementary word, while the presence of a different frame for the same word fails to result in metaphor.” (Black, 1955, p. 276)

The notions of metaphor Black outlines are that of a *substitution view* and of a *comparison view*. The thought of a metaphor as — simply or elaborately — substituting a (more) literal expression is criticized by Black because it ultimately reduces a metaphor to not much more than a “decoration”, to “entertain and divert” the recipient. Also under the substitution view, Black files cases of catachresis, where a metaphor is used because there is not yet an appropriate literal expression — which often results in the word that was used metaphorically gaining a new literal meaning over time, i.e. the word losing its status as metaphor in the respective contexts. The comparison view is described as treating a metaphor as some kind of analogy or simile by transforming the literal meaning of a word. In this regard Black acknowledges the comparison view as a specialization of the substitution view. Black also criticizes the comparison notion, stating that often the similarity between a metaphorical use of a word and a literal alternative is

---

often only created by the metaphor itself, instead of merely being a case of an underlying literal similarity.

Black instead proposes another analysis of metaphors, labeled *interaction view*. He quotes Richards who states that the use of a metaphor results in “two thoughts of different things [being] active together” (Richards, 1936, p. 93). For a metaphor to work, i.e. for the recipient to understand the metaphor, they would need to know a “system of associated commonplaces” about the subjects used in a metaphor to uncover its meaning. This is exemplified by Black in the sentence “Man is a wolf”, in which common — not necessarily true — ideas about “man” and “wolf” are brought together; Black’s commonplaces about wolves contain the images of something “fierce, carnivorous, treacherous”, and so these commonplaces are transferred to the concept of “man”, who, under the notion of this metaphor, “[...] preys upon other animals, is fierce, hungry, engaged in constant struggle, [...]”. These are (arguably) aspects of “man” which are highlighted through the use of the “wolf” metaphor, pushing other aspects into the background. Thus in a metaphor, views of one subject are “organized” in terms or thoughts of another subject.

---

### 1.1.2 Conceptual Metaphors

---

A popular view of metaphors which is very pervasive in scientific literature has been presented by George Lakoff and Mark Johnson in their book “Metaphors We Live By” (Lakoff and Johnson, 2003), in which they introduce the notion of a *conceptual metaphor*. This idea is, in some way, an extension to Black’s interaction view, in that his “system of associated commonplaces” is prescinded by the even more general idea of concepts organized in a *conceptual system* which Lakoff and Johnson introduce.

Lakoff and Johnson postulate that our thoughts and eventually our acts are structured in a conceptual system, which differs depending on our culture and personal background. This system, according to Lakoff and Johnson being highly metaphorical, is usually hidden from us because it permeates everyday life. Communication would be based on the same concepts that shape our thoughts, so we should be able to observe language as a means of communication to find out more about the conceptual system. In language use, they find clues and witnesses for their hypothesis of this conceptual system being metaphorical. An example is given in the metaphorical concept “argument is war”, instances of which would be e.g. “your claims are indefensible” or “he shot down all of my arguments” (Lakoff and Johnson, 2003, p. 4). This example highlights the understanding of metaphor not as a textual instance or utterance, but as a concept which can manifest itself in various forms. Lakoff and Johnson state that within these “structural metaphors”, “one concept is metaphorically structured in terms of another” (Lakoff and Johnson, 2003, p. 14). In the remainder of their work they go into much further detail — analyzing a great deal of different metaphors and studying how metaphors of mind and metaphor manifestations in language are connected, depend on and influence each other.

---

One aspect Lakoff and Johnson also discuss, and which we want to pick up, is the notion of a *new metaphor* as opposed to a *conventional metaphor*. They describe the former as being “outside our conventional conceptual system”, as “imaginative and creative”, while the latter is said to “structure the ordinary conceptual system of our culture, which is reflected in our everyday language” (Lakoff and Johnson, 2003, p. 139). Thus, new metaphors can give us insight into deviations from our conventional conceptual system and provide a deeper understanding of how metaphors can shape the culture they are rooted in. As Lakoff and Johnson put it, “Much of cultural change arises from the introduction of new metaphorical concepts and the loss of old ones.” (Lakoff and Johnson, 2003, p. 145)

---

### 1.1.3 Operationalized definition

---

When searching for metaphors in running text, we have to make concessions as to what constitutes a metaphor; both because of the ambiguity of the subject matter and the inherent imperfectness of language processing tools. Furthermore, technical limitations come into play.

We build our working definition by using Black’s interaction theory as a starting point, i.e. metaphors as an interaction between two thoughts or concepts. Since we cannot work with the thoughts behind a metaphor, we also adopt Black’s structural partition of the textual presentation of a metaphor into a metaphorically used focus and a surrounding frame, which we will call *context*. The *interaction* between context and focus is considered to manifest itself as a “break” between their literal meanings and is employed to identify metaphors. The context shall stretch to the sentence boundaries of the sentence the metaphor is located in, essentially making the sentence our unit of analysis. This amounts to “metaphor” and “sentence which contains a metaphor” meaning the same in the scope of this work.

As we intend to identify *novel* metaphors, we want to disregard dead metaphors (e.g. catachresis as Black defines it), idioms and conventional metaphors. Dead metaphors, like “falling in love”, have become so pervasive that they arguably often do not carry metaphorical meaning anymore — as Black writes, such a metaphor being “merely an expression that no longer has a pregnant metaphorical use” (Black, 2011, p. 25). Idioms and conventional metaphors (as textual instances of Lakoff’s and Johnson’s conventional metaphors) similarly, due to their common usage, have lost *metaphoricity* — much of their once metaphorical meaning has found its way into the literal meaning of their components<sup>1</sup> (Lakoff and Johnson, 2003).

---

<sup>1</sup> Note that according to Lakoff and Johnson, many of these dead metaphors (those that are instances of “systematic metaphorical expressions”, i.e. firmly grounded in metaphorical concepts) or broadly used conventional metaphors are indeed “‘alive’ in the most fundamental sense: they are metaphors we live by”. This does not, as it may seem on first glance, contradict the notion of these instances being dead metaphors; Lakoff and Johnson are merely focusing on a different aspect of “alive”, i.e. these metaphors permeating everyday live.

---

The common attribute of such dead and highly conventional metaphors is a high frequency of context and focus co-occurrence, at least higher than their original literal meanings would suggest. Thus the criterion for novel metaphors will be the frequency (i.e. low frequency) of their appearance, more precisely the frequency of their components (context words and focus words) appearing together.

---

## 1.2 Related work

---

Much work has been published on the recognition of metaphorical concept mapping (in the sense of Lakoff and Johnson) and on the distinction of literal and non-literal use of phrases. The first type of works tries to find abstract conceptual metaphors instead of just textual instances; the second is often concerned with the classification of pre-defined phrases which are potentially metaphorically (or in a broader sense, figuratively) used. While the broad theme of identifying metaphors is the same, their concrete goals (and metaphor definitions) are different from this work's purpose of finding *novel* textual instances of metaphors; still, we may profit from methods used.

We adopt the idea of searching for metaphors which show a similar structure from the approach by Shutova et al. (2010), as it gives us an angle from where to start a classification process. This means that the method used can in later steps be expanded to include further grammatical constructs. Additionally to incorporating selectional preferences as information for the clustering process, we also include the idea of filtering the results with the help of selectional preferences. Instead of using Resnik's method of deriving noun classes through clustering (Resnik, 1993) like it is implemented in *CorMet* (Mason, 2004), we will use the sense labels encoded in GermaNet (Hamp and Feldweg, 1997) directly as broad noun classes (cf. section 3.3.2).

---

### 1.2.1 CorMet

---

A system to find metaphors as defined by Lakoff and Johnson is *CorMet* (Mason, 2004). It tries to find interconcept mappings between two concrete domains (examples are "finance" and "lab"). For this task, selectional preferences for verbs are employed. The algorithms *CorMet* uses are unsupervised; resources are WordNet (Miller, 1995) and a great range of automatically acquired documents from the web.

At first, the system builds domain specific corpora; Google is queried using different combinations of a small set of user-supplied seed keywords. These keywords need to be domain specific. The websites returned are then processed to remove HTML tags and scripts. A similar search is conducted at a later stage to find domain-specific documents which contain a particular verb; additional to the keywords, different morphological forms of the verb in question are used, which

---

---

cannot be homographs of words with other part-of-speech (POS) tags (e.g. they use “attacked” but not “attack” when searching for the verb “to attack”).

In the next step, domain-characteristic verbs are collected. This is done by searching a large set of domain-specific documents for verbs. These verbs are stemmed, and their frequency in the documents is compared to their frequency in general English (with the help of a frequency dictionary). Additionally, verb stems that appear frequently in different domains are filtered out based on the assumption that they are related to internet-specific language (e.g. verbs like “send”, “mail” or “click”). The 400 stems with the best domain/general frequency ratio are used as domain-characteristic verbs.

Now, domain-specific selectional preferences are learned by CorMet. Here, they follow Resnik’s algorithm (Resnik, 1993): for words in predefined grammatical roles (e.g. direct objects of a verb-object dependency), the WordNet nodes which subsume these words are used as selectional preferences. Resulting from the observation that many verbs tend to be too specific in their assigned selection, clustering over WordNet nodes is employed to gain node clusters which act as concepts. This provides for less specific classes, while still maintaining coherence in the concepts and a finer grained system of classes than only using top nodes.

The notion of *polarity* is then introduced to measure the direction and degree of the structure that is thought to be transferred between two concepts (i.e. node clusters) A and B in different domains. First, the verbs which select most strongly for concept A are used, to compute the sum of their selectional preference strengths using the nodes of B as classes. This is repeated with verbs from the second domain which select for nodes in B, used on nodes in A. The difference of these values is then defined as the polarity between clusters/concepts A and B. Overall polarity between two domains is consequently defined as the sum over the inter-concept polarities between the most prominent concepts of each domain.

Another measure conducted by CorMet is *systematicity*, quantifying “a metaphorical mapping’s tendency to co-occur with other mappings” (Mason, 2004, p. 31). In essence, it measures the amount of “strong” metaphors co-occurring with the metaphor in question; a greater systematicity is thought to increase the credibility of a found metaphorical mapping. This is based on the assumption that metaphors which often appear in conjunction with other metaphors are both part of a cohesive metaphorical concept. At the end, CorMet calculates a *confidence* value for each found metaphor, which combines the introduced measures. Additionally to the polarity and the systematicity value, another factor which contributes to this score are the number of predicates which help induce the metaphor.

---

### 1.2.2 Metaphor identification through verb and noun clustering

---

A system for identifying metaphorical expressions is presented by Shutova et al. (2010). It employs verb and noun clustering, as well as selectional preferences to identify metaphors with a similar

---

syntactic structure as a seed set of annotated metaphors. This approach is grounded in the assumption that “target concepts that are associated with the same source concept should appear in similar lexico-syntactic environments” (Shutova et al., 2010, p. 1003).

The procedure starts with a small seed set of metaphors which all exhibit a subject - verb (e.g. “example illustrates”) or a verb - direct object (e.g. “stir excitement”) structure. To have suitable data for the clustering, all verbs from VerbNet (Schuler, 2005) are used to extract up to 10,000 occurrences (i.e. sentences containing the verb) from five corpora, amongst others the British National Corpus (BNC) and the North American News Text Corpus (NANT). Automatically obtained verb subcategorization frames are then used in conjunction with selectional preferences to create a repository of verb - subcategorization frame combination. The most frequent 2000 nouns appearing in the BNC are also clustered, into 200 different clusters; features used incorporate verb lemmas and argument heads from various grammatical verb-noun relations. The actual clustering process technique used is spectral clustering, where the data is represented as a graph (ultimately as a similarity matrix) which is cut into different components<sup>1</sup>. The resulting noun clusters serve as target classes while the verb clusters form the source classes in potential metaphorical mappings. Such mappings are created for source and target clusters which are connected via the seed metaphors.

Then, the selectional preferences are utilized again to filter out verb - subject and verb - object relations in which the verb displays selectional preference strength below an experimentally obtained threshold. This is motivated in the consideration that only verbs having strong preferences are likely to be used metaphorically (e.g. “choose” is rarely used metaphorically, because it allows for arguments from many noun classes to be used literally).

---

<sup>1</sup> The name stems from the spectrum (i.e. the set of eigenvalues) of the Laplacian of the similarity matrix, which is used for the computation of the cuts.

---

## 2 Preparations

---

In this chapter we describe in short the dataset used for the experiments, mentioning briefly the preprocessing steps which are conducted on these documents. We also describe what kind of grammatical constructions we use for clustering, and how they are found.

---

### 2.1 Data set

---

The dataset on which the experiments are run on consists of six documents written in German, manually annotated for metaphors by philosophers. These documents contain philosophical or technical thoughts and — with the exception of the text by Weber — were published in the 18th and 19th century. Table 2.1 shows basic statistics for these texts like sentence and token count, or the amount of annotated metaphors. Furthermore, the documents contain comments which can e.g. indicate that an annotated phrase may be of another type of creative language use. This requires a decision as to which degree of uncertainty such an annotated metaphor really should count as one for the gold standard. By allowing *any* such annotation to count as metaphor, we may widen the working definition, but even then only to also include ambiguous cases — which in itself may be worth investigating. In the automatic evaluation we will differentiate between these cases, essentially producing two gold standards - one including the ambiguous cases, the other without (cf. section 4.1.2).

Author	Extract from	Year	Sentences	Tokens	Lemmas	Met.	Amb.
Hegel	Grundlinien der Philosophie des Rechts, Einleitung	1820	229	10138	1526	7	9
Hegel	Grundlinien der Philosophie des Rechts, Vorrede	1820	108	5528	1195	18	6
Helmholtz	Über die Erhaltung der Kraft	1847	420	13012	2027	2	3
Kant	Was ist Aufklärung?	1784	91	3166	832	6	5
Nietzsche	Vom Nutzen und Nachteil der Historie für das Leben	1874	320	14047	2619	84	35
Weber	Wissenschaft als Beruf	1919	501	13425	2239	16	13

Table 2.1: Statistics for the used dataset. Met. stands for novel metaphor, Amb. for ambiguous.

The documents are in DOCX format for easier manual annotation; to use them for natural language processing (NLP) tasks, we first have to preprocess them. All preprocessing steps are conducted using DKPro Core<sup>1</sup>, a bundled suite of state-of-the-art components for NLP, based on the Apache UIMA<sup>2</sup> framework for managing unstructured information, such as audio, video and texts. The modular nature of UIMA allows for plugging in exactly the tools we need, which are mostly delivered by DKPro Core. A chain of such modules is called a pipeline — consisting of a

---

<sup>1</sup> <http://code.google.com/p/dkpro-core-asl/>, last accessed: 2013-06-15

<sup>2</sup> <http://uima.apache.org/>, last accessed: 2013-06-15

reader component, a number of annotators that process and annotate the imported text, and a consumer for exporting such annotations.

First, the documents are read in with a custom DocxReader which considers background markings and comments. This component is implemented using Apache POI<sup>1</sup>. The documents are then tokenized using the PTBTokenizer<sup>2</sup>. Part-of-speech tagging and lemmatization is done with TreeTagger (Schmid, 1994). At the end, the documents are parsed using the Mate parser (Bohnet, 2010).

---

## 2.2 Candidates

---

The working definition of metaphor is still too vague in technical terms as to readily implement a classifier based on it. If we assume that every metaphor consists of a focus and a frame, it is still not clear what constitutes a focus, how exactly we should find such a possible focus given an arbitrary sentence, or assess if one is present at all. Thus we define a method to find certain structures in text which can hint to the presence of a metaphor; specifying grammatical constructions that are more prevalent among metaphors than others, which could indicate where to search for a potential focus. A manual inspection of the metaphor annotations to Nietzsche’s “Vom Nutzen und Nachteil der Historie für das Leben, Vorwort” (cf. Appendix A) expectantly shows a wide variety of grammatical structures which are used in the vicinity of metaphor focuses. Although this will prevent us from finding “all” metaphors in a text, we will concentrate on finding those metaphors which incorporate one of the following two concrete constructs.

For attributive genitive noun modifiers such as in “Garten des Wissens”, “Zäunen der Vergangenheit” (cf. Figure 2.1), simple part-of-speech tag rules are employed to find occurrences of this construction, which will serve as a focus.

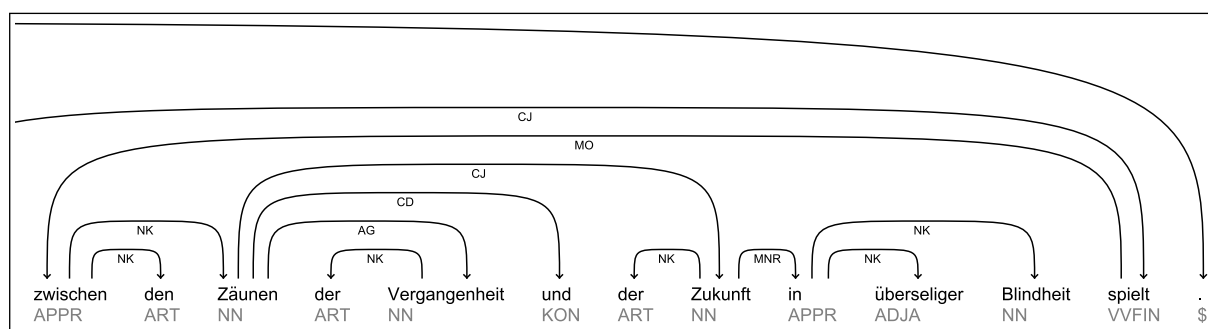


Figure 2.1: Example extract showing a noun - genitive noun modifier construction (POS tags “NN ART[der] NN”) which indicates the metaphor “Zäune der Vergangenheit”.

<sup>1</sup> <http://poi.apache.org/>, last accessed: 2013-06-15

<sup>2</sup> <http://nlp.stanford.edu/software/tokenizer.shtml>, last accessed: 2013-06-15



For verb - direct object pairings like seen in Figure 2.2, we look at the dependency trees gained by parsing the texts with the Mate parser (as the model we use for parsing is trained on the TiGer corpus (Brants et al., 2004), the relations we search for are *OA* - accusative object).

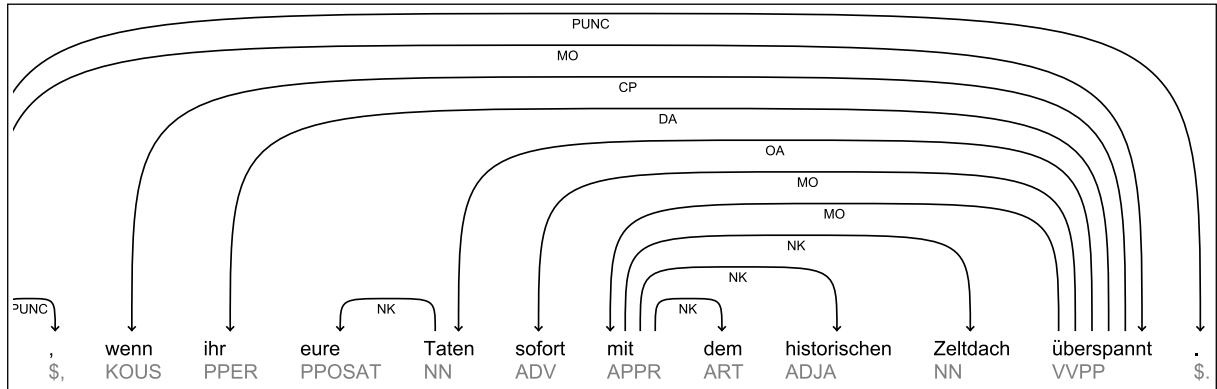


Figure 2.2: Example extract showing a verb - direct object construction (dependency “OA”) which indicates the metaphor “Taten [...] überspannen”.

Sentences which contain such grammatical constructs we will call “metaphor candidates” or simply “candidates”. Each candidate then contains a focus (the two tokens from the found construction) and a context (the remainder of the sentence). After extraction, these candidates are classified to be either metaphors or not. For future work it would be interesting to include other types of candidates, e.g. adjective - noun combinations, or subject - verb and is - a relations.

---

## 3 Identification

---

In this chapter we will describe the clustering method used to identify metaphors. Then the creation of a baseline is detailed, before describing a way to enhance the method by using selectional preferences. At the end, an n-gram based novelty filter is introduced.

---

### 3.1 Clustering

---

The system to automatically identify metaphors uses clustering as its main technique.

*“Clustering algorithms partition a set of objects into groups or clusters. [...] The goal is to place similar objects in the same group and to assign dissimilar objects to different groups.” (Manning and Schütze, 1999, p. 495)*

Applied to finding metaphors, we want to partition a set of sentences into a *novel metaphor group* and a *literal group*, the latter also containing dead and conventionalized metaphors. There are many different clustering algorithms which could be used; prominent examples include e.g. the k-means clustering algorithm, which assigns data points to clusters depending on their *distance* (which can be any distance, e.g. Euclidean) to inferred cluster centers.

Li and Sporleder (2010) propose a method to separate figurative use of idioms from literal use, employing *Gaussian mixture models* (GMM). Although a slightly different use case, we will examine if this approach can also be used to identify novel metaphors. Gaussian mixture models can be seen as a kind of *soft clustering*, where the observed data points (i.e. vectors) are not assigned a single cluster; instead a cluster membership probability is computed. We will use the *expectation maximization* algorithm to compute parameters for a GMM which best fit the given data. In our use case, the vectors which are to be clustered represent sentences. The entries in such a vector are derived from a similarity measure between words in the corresponding sentence.

---

#### 3.1.1 Gaussian mixture models

---

A *finite mixture model* is a weighted sum of  $K$  probability density functions  $p_j$

$$p(x | \theta) = \sum_{j=1}^K \pi_j p_j(x | \theta_j)$$

with weights  $\pi_j \geq 0$ ,  $\sum_{j=1}^K \pi_j = 1$ .

A *Gaussian mixture model*, then, is a mixture model consisting of multiple (multivariate) normal distributions

$$p(x | \theta) = \sum_{j=1}^K \pi_j \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} \exp\left(-\frac{1}{2} (x - \mu_j)^\top \Sigma_j^{-1} (x - \mu_j)\right)$$

with parameters  $\theta_j = (\mu_j, \Sigma_j)$ , where  $\mu_j$  is the  $d$ -dimensional mean vector and  $\Sigma_j$  the (non-singular) covariance matrix of distribution  $j$ .

When looking at mixture models from a clustering perspective, each of the underlying distributions is thought to describe a different cluster of data. The task is to find parameters  $\theta$  which best describe a dataset  $X$ ; that is, we want to estimate parameters which have a maximum likelihood given the data  $X$ .

For each data point  $x_i \in X$  we will introduce a hidden (or unobserved) variable  $z_i \in \{0, 1\}^d$ , where  $z_{ij} = 1$  iff  $x_i$  was emitted by distribution  $j$ ,  $0$  otherwise. The data  $X$  are feature vectors describing sentences (cf. section 3.2).  $K = 2$ , i.e. one distribution producing novel metaphors and one literal sentences.

---

### 3.1.2 Expectation maximization

---

To conduct the maximum likelihood estimation for the parameters of the Gaussian mixture model, we employ the *expectation maximization* (EM) algorithm, which was proposed by Dempster et al. (1977). First we show the EM algorithm in its general form, then we derive the updates when using EM for estimating parameters for a Gaussian mixture model.

For the development of the EM algorithm we follow Manning and Schütze (1999, pp. 520–524) and Bilmes (1998). Let  $X = \{x_1, \dots, x_N\}$  be a set of observed data points, drawn from a distribution  $p$  with parameters  $\theta$ . We now want to maximize the log-likelihood  $\log L(\theta|X) = \log p(X | \theta)$  of the parameters  $\theta$  given  $X$ , that is, find

$$\theta^* = \arg \max_{\theta} \log p(X | \theta) = \arg \max_{\theta} \prod_{i=1}^N \log p(x_i | \theta).$$

The EM algorithm works not by maximizing  $\log p(X | \theta)$  directly, but by iteratively maximizing an auxiliary function  $Q(\theta | \theta^{(t)})$  which additionally to  $X$  takes into account unobserved data  $Z = \{z_1, \dots, z_N\}$  (corresponding to the  $x_i$ ). Expectation maximization consists of two steps,

**Expectation:** Compute the expected value of  $Q$ , that is, the expectation of the joint distribution  $p(X, Z | \theta)$  with respect to  $Z$  given  $X$  and parameters  $\theta^{(t)}$ :

$$Q(\theta | \theta^{(t)}) := E_{Z|X, \theta^{(t)}} [\log p(X, Z | \theta)]$$

**Maximization:** Obtain parameters  $\theta^{(t+1)}$  which maximize  $Q(\theta \mid \theta^{(t)})$ , i.e.

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta \mid \theta^{(t)}).$$

As we will show, increasing  $Q(\theta \mid \theta^{(t)})$  by repeating these steps will indeed increase the log-likelihood  $\log L(\theta \mid X)$ . We have

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= E_{Z \mid X, \theta^{(t)}} [\log p(X, Z \mid \theta)] \\ &= \sum_Z p(Z \mid X, \theta^{(t)}) \log p(X, Z \mid \theta) \\ &= \sum_Z p(Z \mid X, \theta^{(t)}) \log (p(Z \mid X, \theta) p(X \mid \theta)) \\ &= \log p(X \mid \theta) + \sum_Z p(Z \mid X, \theta^{(t)}) \log p(Z \mid X, \theta), \end{aligned}$$

which gives us

$$\log p(X \mid \theta) = Q(\theta \mid \theta^{(t)}) - \sum_Z p(Z \mid X, \theta^{(t)}) \log p(Z \mid X, \theta).$$

Using two subsequent values of  $\theta$ , we can write

$$\begin{aligned} \log p(X \mid \theta^{(t+1)}) - \log p(X \mid \theta^{(t)}) \\ = Q(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) + H(\theta^{(t)} \mid \theta^{(t)}) - H(\theta^{(t+1)} \mid \theta^{(t)}), \end{aligned}$$

where  $H(\theta \mid \theta^{(t)}) = \sum_Z p(Z \mid X, \theta^{(t)}) \log p(Z \mid X, \theta)$ . By definition (in the maximization step), we already have

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta^{(t)} \mid \theta^{(t)}).$$

In the next step, we use Jensen's inequality, which states that for a convex function  $f$  (such as  $-\log$ ) on an interval  $I$ ,  $x_1, x_2, \dots, x_n \in I$  and  $\lambda_1, \lambda_2, \dots, \lambda_n \geq 0$ ,  $\sum_{i=1}^n \lambda_i = 1$ , the following inequality holds:

$$\sum_{i=1}^n \lambda_i f(x_i) \geq f\left(\sum_{i=1}^n \lambda_i x_i\right)$$

Using this property of  $-\log$  at (\*), we then see

$$\begin{aligned} H(\theta^{(t)} \mid \theta^{(t)}) - H(\theta^{(t+1)} \mid \theta^{(t)}) &= \sum_Z p(Z \mid X, \theta^{(t)}) [\log p(Z \mid X, \theta^{(t)}) - \log p(Z \mid X, \theta^{(t+1)})] \\ &= - \sum_Z p(Z \mid X, \theta^{(t)}) \log \frac{p(Z \mid X, \theta^{(t+1)})}{p(Z \mid X, \theta^{(t)})} \\ &\stackrel{(*)}{\geq} - \log \sum_Z p(Z \mid X, \theta^{(t)}) \frac{p(Z \mid X, \theta^{(t+1)})}{p(Z \mid X, \theta^{(t)})} \\ &= - \log \sum_Z p(Z \mid X, \theta^{(t+1)}) = 0. \end{aligned}$$

Hence,  $\log p(X | \theta^{(t+1)}) - \log p(X | \theta^{(t)}) \geq 0$ , that is,

$$\log p(X | \theta^{(t+1)}) \geq \log p(X | \theta^{(t)}),$$

which shows that parameters  $\theta^{(t+1)}$  which maximize  $Q(\theta | \theta^{(t)})$  also increase  $\log p(X | \theta) = \log L(\theta | X)$ .

We now want to use expectation maximization to estimate parameters for a Gaussian mixture model to fit data  $X$ . To employ the EM algorithm we need to find update formulas for the parameters  $\theta$ . Let  $K$  be the number of mixture components and  $N$  be the amount of data points. For  $i \leq N$ ,  $j \leq K$ , we introduce the hidden variables  $Z = \{z_1, z_2, \dots, z_N\}$  with

$$z_{ij} = \begin{cases} 1 & \text{if mixture component } j \text{ is responsible for } x_i \\ 0 & \text{otherwise} \end{cases}$$

We also write  $\theta = (\theta_1, \dots, \theta_K)$ , where  $\theta_j = (\mu_j, \Sigma_j)$ . Let  $n(x | \mu, \Sigma)$  be the probability density function of a multivariate normal distribution, then

$$\begin{aligned} p(X, Z | \theta) &= \prod_{i=1}^N \sum_{j=1}^K z_{ij} \pi_j n(x_i | \mu_j, \Sigma_j) \\ \log p(X, Z | \theta) &= \sum_{i=1}^N \log \sum_{j=1}^K z_{ij} \pi_j n(x_i | \mu_j, \Sigma_j), \end{aligned}$$

and because for each  $x_i$ ,  $i \leq N$ ,  $z_{ij} = 0$  for all but one  $j \leq K$ , we have

$$\log p(X, Z | \theta) = \sum_{i=1}^N \sum_{j=1}^K z_{ij} \log(\pi_j n(x_i | \mu_j, \Sigma_j)).$$

Thus we can formulate the auxiliary function  $Q(\theta | \theta^{(t)})$  for Gaussian mixture models as

$$Q(\theta | \theta^{(t)}) = E_{Z|X, \theta^{(t)}} [\log p(X, Z | \theta)] = \sum_{i=1}^N \sum_{j=1}^K E_{Z|X, \theta^{(t)}} [z_{ij}] \log(\pi_j n(x_i | \mu_j, \Sigma_j)),$$

since the expression inside the  $\log$  is independent of  $Z$ . We introduce labels  $h_{ij}$  for  $E_{Z|X, \theta^{(t)}} [z_{ij}]$ :

$$h_{ij}^{(t)} := E_{Z|X, \theta^{(t)}} [z_{ij}] = 0 \cdot p(z_{ij} = 0 | x_i, \theta^{(t)}) + 1 \cdot p(z_{ij} = 1 | x_i, \theta^{(t)}) = p(z_{ij} = 1 | x_i, \theta^{(t)}).$$

Using Bayes theorem and acknowledging that the prior probability  $\mathbf{p}(z_{ij} = 1 \mid \theta^{(t)})$  of an  $x_i$  having been created by the Gaussian  $j$  is just  $\pi_j^{(t)}$ , we arrive at

$$\begin{aligned} h_{ij}^{(t)} &= \mathbf{p}(z_{ij} = 1 \mid x_i, \theta^{(t)}) \\ &= \frac{\mathbf{p}(x_i \mid z_{ij} = 1, \theta^{(t)}) \mathbf{p}(z_{ij} = 1 \mid \theta^{(t)})}{\mathbf{p}(x_i \mid \theta^{(t)})} \\ &= \frac{\pi_j^{(t)} \mathbf{n}(x_i \mid \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{k=1}^K \pi_k^{(t)} \mathbf{n}(x_i \mid \mu_k^{(t)}, \Sigma_k^{(t)})}. \end{aligned}$$

We now write  $Q(\theta \mid \theta^{(t)})$  as

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^N \sum_{j=1}^K h_{ij}^{(t)} \log(\pi_j \mathbf{n}(x_i \mid \mu_j, \Sigma_j)).$$

To see how to recompute the parameters for the maximization of  $Q(\theta \mid \theta^{(t)})$ , we take the partial derivative of  $Q$  with respect to  $\theta_j$ . Beginning with  $\mu_j$  we have:

$$\begin{aligned} \frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \mu_j} &= \frac{\partial}{\partial \mu_j} \sum_{i=1}^N \sum_{k=1}^K h_{ik}^{(t)} \log(\pi_k \mathbf{n}(x_i \mid \mu_k, \Sigma_k)) \\ &= \frac{\partial}{\partial \mu_j} \sum_{i=1}^N \sum_{k=1}^K \left[ h_{ik}^{(t)} \log(\pi_k (2\pi)^{-\frac{d}{2}} \left| \Sigma_k \right|^{-\frac{1}{2}}) - \frac{1}{2} h_{ik}^{(t)} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right] \\ &= \frac{\partial}{\partial \mu_j} \sum_{i=1}^N \left( -\frac{1}{2} h_{ij}^{(t)} (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) \right) \\ &= \sum_{i=1}^N h_{ij}^{(t)} \left( -\Sigma_j^{-1} (x_i - \mu_j) \right). \end{aligned}$$

Equating this with 0 yields

$$\begin{aligned} 0 &= \sum_{i=1}^N h_{ij}^{(t)} \left( -\Sigma_j^{-1} (x_i - \mu_j) \right) \\ 0 &= \sum_{i=1}^N h_{ij}^{(t)} (x_i - \mu_j) \\ \mu_j^{(t+1)} &:= \mu_j = \frac{\sum_{i=1}^N h_{ij}^{(t)} x_i}{\sum_{i=1}^N h_{ij}^{(t)}}. \end{aligned}$$

Taking the partial derivative with respect to  $\Sigma_j^{-1}$ , we end up with

$$\begin{aligned} \frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \Sigma_j^{-1}} &= \frac{\partial}{\partial \Sigma_j^{-1}} \sum_{i=1}^N \sum_{j=1}^K h_{ij}^{(t)} \log(\pi_j \mathbf{n}(x_i \mid \mu_j, \Sigma_j)) \\ &\Rightarrow \Sigma_j^{(t+1)} := \Sigma_j = \frac{\sum_{i=1}^N h_{ij}^{(t)} (x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_{i=1}^N h_{ij}^{(t)}}. \end{aligned}$$

(For the full derivation of  $\Sigma_j^{(t+1)}$ , refer to appendix B).

What is left now is the re-computation of the weights  $\pi_j$  of the mixture model. Since these are constrained by  $\sum_{j=1}^K \pi_j = 1$ , we reformulate  $Q(\theta \mid \theta^{(t)})$  by using a Lagrangian multiplier  $\lambda$ :

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^N \sum_{j=1}^K h_{ij}^{(t)} \log(\pi_j \mathbf{n}(x_i \mid \mu_j, \Sigma_j)) + \lambda \left( \sum_{j=1}^K \pi_j - 1 \right).$$

Now taking the partial derivative with respect to  $\pi_j$

$$\frac{\partial}{\partial \pi_j} Q(\theta \mid \theta^{(t)}) = \frac{\partial}{\partial \pi_j} \left[ \sum_{i=1}^N \sum_{k=1}^K h_{ik}^{(t)} \log(\pi_k \mathbf{n}(x_i \mid \mu_k, \Sigma_k)) + \lambda \left( \sum_{k=1}^K \pi_k - 1 \right) \right] = \lambda + \sum_{i=1}^N \frac{h_{ij}^{(t)}}{\pi_j}.$$

Then

$$0 = \lambda + \sum_{i=1}^N \frac{h_{ij}^{(t)}}{\pi_j} \quad \Rightarrow \quad -\lambda \pi_j = \sum_{i=1}^N h_{ij}^{(t)}.$$

Summing over the  $\pi_j$  (and incorporating  $\sum_{j=1}^K \pi_j = 1$ ) we obtain  $\lambda = -N$ , and thus arrive at

$$\pi_j^{(t+1)} := \pi_j = \frac{1}{N} \sum_{i=1}^N h_{ij}^{(t)}$$

This also implicitly secures  $\pi_j \geq 0$ . We now have update formulas for all our parameters.

Both EM steps are iterated until the log-likelihood does not increase anymore above a predefined threshold. To start the EM algorithm, initial values for  $\theta_j$  and  $\pi_j$  are needed for all component densities  $n_j$ . We use  $\pi_j = \frac{1}{K}$  as starting weights and  $\Sigma_j = I_K$ . To initialize  $\mu_j$  we use an implementation of the K-means algorithm which is provided by the machine-learning workbench Weka (Hall et al., 2009).

Also, since the EM algorithm can only find local maxima, we need to run it multiple times from different initial guesses. While the K-means algorithm often finds good starting positions, for subsequent runs we use random samples of the data  $X$  as  $\mu_j$ . Then we choose the parameters which achieve the best log-likelihood. After constructing the model we can compute the cluster membership probabilities as the posterior probability  $h_{ij}$  for each of the data points  $x_i \in X$ .

---

## 3.2 Creating a baseline

---

For creating the baseline, the approach by Li and Sporleder (2010) is implemented. They employ Gaussian mixture models to identify figurative language using its context — an approach that aligns reasonably well with our working definition of a metaphor consisting of a focus and a context.

Like in the original paper, we assume our Gaussian mixture model to consist of two Gaussians, namely a *metaphorical* and a *non-metaphorical* Gaussian. Starting with the features from Li and Sporleder (2010, p. 298), the feature vector representing a data point (i.e. metaphor candidate) is comprised of several semantic relatedness features,  $\mathbf{x} = (x_1, \dots, x_{54})$ . Let  $F$  be the list of focus lemmas and  $C$  the list of context lemmas, then

$$x_1 = \frac{2}{|F| \cdot |C|} \sum_{(w_i, c_j) \in F \times C} \text{relatedness}(w_i, c_j)$$

is the average relatedness between words in the context and words in the focus, while

$$x_2 = \frac{2}{|C| \cdot (|C| - 1)} \sum_{(c_i, c_j) \in C \times C, i \neq j} \text{relatedness}(c_i, c_j)$$

is the average semantic relatedness between words in the context.

$$x_3 = x_1 - x_2$$

is the difference between the context-focus relatedness and the inner-context relatedness, and the following feature presents a binary classification whether the context has a higher degree of cohesion with itself or the focus expression.

$$x_4 = \begin{cases} 1 & \text{if } x_3 < 0 \\ 0 & \text{otherwise} \end{cases}$$

The remaining features are defined as

$$x_{4+k} = \min_{(w_i, c_j) \in F \times C} (k, \{\text{relatedness}(w_i, c_j)\}) \quad \text{if } k < |F| \cdot |C|$$

and

$$x_{4+k} = \max_{(w_i, c_j) \in F \times C} (\{\text{relatedness}(w_i, c_j)\}) \quad \text{if } k \geq |F| \cdot |C|$$



---

for  $k = 1, \dots, 50$ , where  $\min(k, A)$  chooses the  $k$ -th lowest element in the set  $A$ . In their approach, Li and Sporleder use  $k = 1, \dots, 100$ ; after filtering out non-content words, this range of  $k$  produces many vectors which are identical regarding the entries with high indices, thus we stop at  $k = 50$ . The **relatedness** function used in the feature vector is the *Normalized Google Distance* (NGD). Note that in (Li and Sporleder, 2010)  $\max(k, A)$  is used instead of  $\min(k, A)$ . It is not discussed in detail in what way they employ the NGD, where a low value indicates a high degree of relatedness. Here we use the NGD directly as the measure for relatedness; so in order to replicate the assumed intention of the feature definition (i.e. list scores for the most related tokens first), we use  $\min(k, A)$ .

---

### 3.2.1 Calculating semantic similarity

---

*Normalized Google Distance* is a semantic distance measure introduced by Cilibrasi and Vitanyi (2004). It is based on the theoretical Normalized Information Distance, which is a non-computable similarity distance measure between binary strings. It needs to be mentioned that because of this heritage, the *Normalized Google Distance* does not disambiguate between senses of input words; thus we are not measuring the maximum similarity over all senses of the input words, but rather the combined similarity. The Normalized Google Distance incorporates page counts from a search engine (while they use Google as an example, any search engine could be used), so a great advantage over semantic similarity measures which work on manually crafted resources such as WordNet (Miller, 1995) is its coverage (given a large enough corpus) – we would be hard pressed to find a word which is not contained in any web page indexed by Google. Let  $x$  and  $y$  be input words whose similarity we want to assess. With a sufficiently large number  $N$  (having  $N \geq M$ , where  $M$  is the number of web pages indexed by the search engine), the Normalized Google Distance is formally defined as

$$\text{NGD}(x, y) = \frac{\max(\log f(x), \log f(y)) - \log f(x, y)}{\log N - \log \min(\log f(x), \log f(y))}$$

$f(x)$  is the number of web pages in the search engine index which contain the word  $x$ , while  $f(x, y)$  is the number of web pages which contain both  $x$  and  $y$ . This means that the Normalized Google Distance of a word with itself is zero,  $\text{NGD}(x, x) = 0$ . Furthermore, if  $f(x) = 0$  for a word  $x$ ,  $\text{NGD}(x, y)$  is defined as 1, which is sensible. To see this, consider  $\lim_{f(x) \searrow 0} f(x, y) \rightarrow 0$  and l'Hôpital's rule (\*):

$$\lim_{f(x) \searrow 0} \text{NGD}(x, y) = \lim_{f(x) \searrow 0} \frac{\log f(y) - \log f(x, y)}{\log N - \log f(x)} \stackrel{(*)}{=} \lim_{f(x) \searrow 0} \frac{-f(x)^{-1}}{-f(x)^{-1}} = 1, \quad (3.1)$$

assuming w.l.o.g.  $f(x) < f(y)$ . In the case that we have  $f(x) > 0$  and  $f(y) > 0$ , but  $f(x, y) = 0$ , we manually set  $f(x, y) = 1$ .

---

The purpose of using the Normalized Google Distance values to create the feature vectors prevents us from using the Google search engine directly. The first reason is stated by Cilibiasi and Vitány: Google result numbers are not reliable, as they change with every search (for the same word) and are an estimate rather than a concrete count. It is also too time consuming to search for every word, even if the results are cached. And importantly, Google has since restricted API access to its search engine service to 100 requests per day<sup>1</sup>, way less than would be necessary to compute the feature vectors. Other search engines like Yahoo<sup>2</sup> or Bing<sup>3</sup> impose similar restrictions on their API access.

---

### 3.2.2 deWaC

---

To solve the resource problem, instead of a search engine index, *deWaC* (Baroni et al., 2009) is used as a corpus. *deWaC* was built by crawling the web (confined to the .de domain), thus is a corpus comprised of web sites, consisting of over 1.2 billion tokens in 1.75 million web pages. Together with its English and Italian counterparts, *ukWaC* and *itWaC* respectively, *deWaC* was created by the WaCky project, “an informal consortium of researchers interested in the exploration of the web as a source of linguistic data” (Baroni et al., 2009, p. 1).

The corpus is already POS tagged and lemmatized using TreeTagger (Schmid, 1994). *deWaC* can be obtained online for free<sup>4</sup> and the files are delivered in the IMS Open Corpus Workbench format (Evert and Hardie, 2011). Because of its nature as a web corpus — which, given a broad enough seed, implies a wide diversity — and its size, *deWaC*’s coverage of our dataset is sufficient, covering 93.66 % of all content words (i.e. adjectives/adverbs, nouns and verbs) of our experiment dataset (cf. Table 3.1). A manual inspection shows that an overwhelming majority of the non-covered words are either wrongly lemmatized, i.e. their lemma is identical to their respective type when that should clearly be not the case, or represent composite nouns like “Hundert-Männer-Schar” or “Kollektivwillensbildung”. The main causes for wrong lemmatization apparently include wrong spelling (“Mseitwärts”), obsolete spelling (“controlirenden”) or wrong tagging (in some instances, non-word-characters like “§” or “-” are not recognized correctly and are assigned adjective, noun or verb tags). Some texts also include descriptions and meta-tags (“pages187-229”, “titleUeber”).

---

<sup>1</sup> <https://developers.google.com/custom-search/v1/overview>, last accessed: 2013-06-15

<sup>2</sup> <http://developer.yahoo.com/search/rate.html>, last accessed: 2013-06-15

<sup>3</sup> <http://datamarket.azure.com/dataset/bing/search>, last accessed: 2013-06-15

<sup>4</sup> <http://wacky.sslmit.unibo.it/doku.php?id=download>, last accessed: 2013-06-15

---

Part of speech	covered / existing	coverage
Adjectives / Adverbs	1708 / 1809	94.42 %
Nouns	2923 / 3194	91.52 %
Verbs	1209 / 1232	98.13 %

---

Figure 3.1: deWaC coverage of lemmas in dataset documents

---

### 3.2.3 Lucene

---

Because the search for lemmas in the raw deWaC XML files would take too long, *Lucene*<sup>1</sup> is employed to speed up the lookup process. Lucene is a “text search engine library”, which allows for indexing texts and fast lookup. Indexing deWaC in Lucene allows us to use the resulting *inverted index*. Instead of a document being an index for its words, we now have words which point to all documents they are used in. This makes it trivial to obtain the number of documents a word is part of and — e.g. via an intersection of their document sets — also the number of documents which include two different words.

For indexing we set up a simple UIMA pipeline. Since deWaC comes in the IMS CWB format, we can use the corresponding DKPro reader to import the deWaC information into CASes. We now have to supply a custom consumer which creates the Lucene index. This consumer iterates over all adjective/adverb, noun and verb tokens and writes their lemmas along with their respective document id into the index. We also filter lemmas which do not begin with a word-character, in which cases we assume incorrect tagging or lemmatization.

Searching is straight forward, with one exception: in some cases, TreeTagger produces multi-lemmas for tokens which have to be handled specially. Taking an example

*“Die Frage aber, bis zu welchem Grade das Leben den Dienst der Historie überhaupt brauche, ist eine der höchsten Fragen und Sorgen betreff der Gesundheit eines Menschen, eines Volkes, einer Kultur.” (Nietzsche, 1874)<sup>2</sup>*

Here, “Sorgen” is lemmatized as “Sorge|Sorgen” (a combination of the noun “Sorge” — “worry” and the nominalization of “to worry” — “Sorgen”); in deWaC, although also lemmatized with TreeTagger, there is no such lemma, and also no lemma “Sorgen”; “Sorge” is therefore used exclusively. This problem is solved by taking these instances of ambiguity and searching for every lemma of the multi-lemma expression. Then we use the one which yields the best results (i.e. the least distance) when computing the Normalized Google Distance to another given lemma.

---

<sup>1</sup> <http://lucene.apache.org/>, last accessed: 2013-06-15

<sup>2</sup> (Nietzsche, *Unzeitgemässe Betrachtungen, Zweites Stück: Vom Nutzen und Nachtheil der Historie für das Leben. Vorwort., 1874, ch. 1*)

---

The search itself is conducted similarly as outlined above. For the single term frequencies we query Lucene for the amount of documents (web pages) a lemma is situated in. For obtaining the combined frequency, directly querying for documents containing both terms proved to be a faster alternative to intersecting the respective document sets of the tokens.

---

### 3.3 Building a repository of selectional preferences

---

To improve on the baseline results we want to add explicit semantic information to the vectors. One way of doing this is to find out if there are strong semantic ties between a verb and its object in the candidate sentence — likewise for genitive noun modifier candidates. Thus, we are employing *selectional preferences*.

---

#### 3.3.1 Selectional preferences

---

In “Foundations of Statistical Natural Language Processing” (Manning and Schütze, 1999), Christopher Manning and Hinrich Schütze define selectional preferences as follows:

*“Most verbs prefer arguments of a particular type. Such regularities are called selectional preferences or selectional restrictions. Examples are that the objects of the verb eat tend to be food items, the subjects of think tend to be people, and the subjects of bark tend to be dogs.” (Manning and Schütze, 1999, p. 288)*

They further state that

*“We use the term preferences as opposed to rules because the preferences can be overridden in metaphors and other extended meanings.” (ibid.)*

When overriding such a preference we also speak of a *violation of selectional preferences*. As cited, such a violation may indicate the presence of a metaphor.

To measure the strength of a selectional preference, Manning and Schütze describe the notion of *selectional preference strength* and *selection association*, following a model first proposed by Resnik (1993). Going by the example of a verb - direct object relationship, they define selectional preference strength as a measure of “how strongly the verb constrains its direct object” (Manning and Schütze, 1999, p. 289). Formally, it is introduced as the Kullback-Leibler divergence (Kullback and Leibler, 1951) between the prior probability of a class of a direct object — specifically noun classes, such as “animate”, “consumable” — and the posterior probability of this class given the verb as head. The Kullback-Leibler divergence  $D(P_1 \parallel P_2)$  itself is an information theoretical

---

measure to quantify the difference between two probability distributions  $P_1, P_2$ . For a verb  $v$  we write its selectional preference strength as

$$S(v) = D(P(C | v) || P(C)) = \sum_c P(c | v) \log \frac{P(c | v)}{P(c)},$$

where  $C$  denotes the set of noun classes; this amounts to  $P(C | v)$  being the probability distribution of noun classes as direct objects of the verb  $v$ , and  $P(C)$  the distribution of noun classes in direct object position of any verb;  $P(c)$  is the prior probability of encountering a noun with class  $c$  as a direct object of an arbitrary verb, and  $P(c | v)$  is the posterior probability of a noun with class  $c$  being the direct object of verb  $v$ . Selectional association (between a verb and a noun class) is a derived measure which quantifies the proportion a noun class takes of the preference strength:

$$A(v, c) = \frac{P(c | v) \log \frac{P(c | v)}{P(c)}}{S(v)}$$

The last definition we want to use is that of selectional association between a verb and a specific noun, which is simply defined as the maximum over the noun classes of all senses the noun possesses:

$$A(v, n) = \max_{c \in \text{classes}(n)} A(v, c)$$

In order to compute and employ the selectional association, we now need the prior and posterior probabilities  $P(c)$  and  $P(c | v)$  for all noun classes of the nouns we are interested in, and for all verbs we are interested in. Note that because selectional preference is not constrained to verb - direct object relations, the  $v$  in above formulas generally stands for the head word in any discussed dependency relation.

---

### 3.3.2 Resources

---

Because resources for selectional preferences in German are scarce, a limited repository needs to be built; we will use dependency and n-gram frequencies from large corpora. “Limited” meaning here, that e.g. for noun - genitive noun modifier relations we are only covering the nouns which occur in the documents we want to evaluate in this work (cf. section 2.1) as head nouns. This limit is imposed because of time and resource constraints — it would not be feasible to create a more “complete” repository for this task, nor is it necessary. Although for generalizing this metaphor detection method to arbitrary documents it would be necessary to have a larger database covering as many selectional preferences as possible, as we do not want to rebuild the database every time a new document is analyzed.

In contrast to a curated database of selectional preferences the relations encoded in a frequency based repository are expected to be much less clear-cut. This does not have to be a disadvantage, as it allows for a more fine-grained computation of selectional association. From a manually crafted repository we can also not expect the scope and variety presented by one created automatically. Especially for the task of finding novel metaphors, such a repository built on frequencies in a large corpus can have the advantage of capturing some conventional metaphoric relations (because they are, by definition, used widely) as “legal”, i.e. non-violating their respective selectional preferences — an expert created database would not necessarily reflect this treatment of conventional metaphors.

We first describe the creation for a selectional preference repository for nouns as head words, followed by a slightly different approach for verbs. For the creation of the repository, we start by examining the candidates we found by noun - genitive noun modifier relations. For these, we employ the dependency relations encoded in the *Google Books Ngram Corpus* (Lin et al., 2012). The Google Books Ngram Corpus (GBNC) is a corpus which comprises n-grams (up to  $n = 5$ ) of books from eight languages, for a total of over eight million different books (as we deal with German documents, we employ only the German part of the GBNC). In addition, the books have been parsed using a dependency parser (Michel, J.-B. et al., 2011). Figure 3.2 shows an excerpt of such a GBNC dependency file. Dependency relations are grouped into files by the first letter of the head word of a relation.

Jahrzehnte=>Weltlage_NOUN	1982	4	2
Jahrzehnte=>Weltlage_NOUN	1992	1	1
Jahrzehnte=>Weltlage_NOUN	1994	1	1
Jahrzehnte=>Wirkens_NOUN	1862	1	1
Jahrzehnte=>Wirkens_NOUN	1897	1	1
Jahrzehnte=>Wirkens_NOUN	1904	4	4

Figure 3.2: Excerpt from a GBNC dependency file (j), showing head=>dependant, year, match-count and volumecount.

We also need information about noun classes, more precisely what semantic category a given noun belongs to. Since many nouns can have different senses depending on the context they are used in, consequently they can belong to different categories. So more precisely, we speak about semantic classes for senses which a given noun possesses.

For this task of extracting semantic categories for nouns, *GermaNet* (Hamp and Feldweg, 1997) is queried<sup>1</sup>. *GermaNet* is a “lexical-semantic net” of the German language and is also described as a “light-weight ontology”. More importantly, *GermaNet* also includes *semantic fields* for word senses, similar to the major nodes of the *GermaNet* taxonomy. There are 24 different

<sup>1</sup> This work uses version 6.0, although at the time of writing 8.0 is current

semantic fields for noun senses encoded in GermaNet (cf. Table 3.1), which produces a rather coarse classification. Querying for “Wasser” (water) e.g. yields the two *semantic fields* “Nahrung” (food) and “Substanz” (substance) — there is no semantic field just for liquids. For the remainder of this work, we will use the terms *noun class* and *semantic field* synonymously, because we are only employing semantic fields for nouns (the term used in the UBY API (see next paragraph), *sense label*, is also used synonymously here).

Semantic noun field	Relative frequency	Semantic noun field	Relative frequency
Artefakt	10.47 %	Motiv	0.94 %
Attribut	6.14 %	Nahrung	0.97 %
Besitz	2.40 %	Ort	6.86 %
Form	1.54 %	Pflanze	0.28 %
Gefuehl	1.54 %	Relation	3.22 %
Geschehen	15.46 %	Substanz	1.51 %
Gruppe	8.35 %	Tier	0.27 %
Koerper	2.19 %	Tops	0.46 %
Kognition	9.99 %	Zeit	4.14 %
Kommunikation	10.82 %	artificial	0.00 %
Menge	3.62 %	natGegenstand	0.77 %
Mensch	6.80 %	natPhaenomen	1.27 %

Table 3.1: GermaNet semantic fields and their relative frequency occurring as dependant of a noun in the GBNC.

To access the senses and their semantic fields for a given noun, *UBY* (Gurevych and Eckle-Kohler, 2012) is employed. *UBY* is a unified lexical-semantic resource which combines information from different expert or collaboratively constructed sources in English and German like WordNet, Wikipedia, OmegaWiki, and also GermaNet. The different original resources are linked in the *UBY* database on a sense level. Access is provided through the *UBY* Java API. Searching for English metaphors we could employ the selectional restrictions encoded in *VerbNet* (Schuler, 2005) and *FrameNet* (Baker et al., 1998) through the use of *UBY*, but unfortunately the links between the same senses in different languages only exist for nouns.

---

### 3.3.3 Creation process

---

We use the preprocessed documents from our dataset (cf. section 2.1) and use the candidates of type noun - genitive noun modifier. The head words of these relations pose as the index entries for our repository of posterior probabilities. These lemmatized nouns are grouped by their first letter, as the entries in the GBNC are also grouped into files by their first letter. For each group,

we scan through the respective GBNC file and extract the token and the match count for each entry having one of the group’s nouns as head, essentially building a mapping

$$\textit{head noun} \rightarrow (\textit{dependant} \rightarrow \textit{count}).$$

We then look up the semantic field in GermaNet for each sense of every dependant with the help of UBY. The counts for a dependant noun are split between its sense labels, and the counts for each sense label are accumulated for every noun head, gaining another mapping

$$\textit{head noun} \rightarrow (\textit{label} \rightarrow \textit{count}).$$

These final mappings are saved to a database and serve as a base repository for the posterior probabilities  $P(c | v)$  (here,  $v$  stands for the head word in a noun - genitive noun modifier relation) we can use to calculate selectional preference strength.

The prior probabilities  $P(c)$  for a noun class  $c$  are calculated again using the Google Books Ngram Corpus, specifically the POS tag dependency files. Entries in these files are similar to the “normal” dependency files, but each entry starts with a POS tag instead of a token (cf. Figure 3.3).

_NOUN_=>Luftschiffhafen	2007	15	5	
_NOUN_=>Luftschiffhafen	2008	3	3	
_NOUN_=>Luftschiffhafen	2009	7	3	
_NOUN_=>Luftverkehrsaufkommens	1954		1	1
_NOUN_=>Luftverkehrsaufkommens	1956		1	1
_NOUN_=>Luftverkehrsaufkommens	1958		3	3

Figure 3.3: Excerpt from a GBNC dependency file (noun), showing `_POS_=>`dependant, year, matchcount and volumecount.

Since German nouns begin with a capital letter, we scan those files for such tokens<sup>1</sup>, lemmatizing the findings. For each lemma, we extract the corresponding noun classes from UBY/GermaNet and split the token count among those classes. After normalizing the counts by the total count we arrive with the prior probabilities presented in Table 3.1.

Following the formula for selectional association above, we can now easily compute  $A(v, n)$  for any noun - genitive noun modifier pair from our dataset. To be able to also calculate the selectional association between verbs and direct objects, we have to go a different route.

One problem with the Google Books Ngram Corpus dependency files is the missing type information. While we ignored that problem under the assumption that most of the dependencies

<sup>1</sup> This essentially means that we also search for non-noun tokens which, in their original location, had been at the beginning of a sentence or are misspelled to have an uppercase first letter. These are mostly filtered out at the sense label lookup step, because there is no noun class encoded for them in GermaNet.



---

encoded in the GBNC between nouns are indeed noun - genitive noun modifier relations, the situation for verbs is more complicated, as we cannot predict whether a noun is used as a subject or object. Thus we need to follow a different strategy for verbs - direct object relations.

We parse the first 60,000 documents from deWaC and extract all direct object dependencies (TiGer label *OA* - accusative object). Analogous to the process for extracting noun - genitive noun relations, now the verbs serve as our index entries, building a mapping

$$verb \rightarrow (direct\ object \rightarrow count).$$

Again, the semantic fields for each dependant are queried from GermaNet, distributing the count evenly across the labels for the dependant, resulting in a mapping

$$verb \rightarrow (label \rightarrow count).$$

The prior probability for each label is then just the sum over the label counts for all verbs and the label in question.

---

### 3.3.4 Shortcomings

---

There are some shortcomings in the repository creation process, some of which have already been mentioned.

The noun detection method for the Google Books Ngram Corpus (i.e. by uppercase first letter) can misdetect tokens as nouns which are not — e.g. tokens at the beginning of a sentence or misspelled tokens. This is mostly resolved by the noun class search in GermaNet, as we add a noun part-of-speech tag constraint to the query.

Another issue exists because the words in the GBNC are not lemmatized. Since lemmatization of the nouns prior to the GermaNet lookup would be too time consuming, it is foregone under the assumption that each noun class suffers from missing counts proportionately to their quantity. If so, this would change the absolute counts, but keep the percentages (i.e. prior probabilities  $P(c)$ ) relatively stable.

The missing lemmatization in the GBNC files is also a problem for the dependency lookup. We would have to conjugate each noun we want to search a dependency for, to all possible forms. As implemented, the counts only reflect a search for the corresponding lemma — again it is assumed that this does not alter the percentages (i.e. posterior probabilities  $P(c | v)$ ) significantly.

On the issue of using unlabeled dependency information from the GBNC for the noun - genitive noun modifier relations, it would be desirable to instead harvest this information from the n-gram (e.g. 3-gram) part of the corpus, using token-based rules. This could probably result in slightly higher quality selectional preferences, but is also much more time-consuming.

---

## 3.4 Novelty

---

As outlined in section 1.1.3 we want to be able to distinguish between metaphors which are novel and metaphors which have been largely integrated into our common use of language; the latter shall be dismissed like any non-metaphor. We assume that — to an extent — this is already taken into account by the way the feature vectors for the candidates are created (cf. section 3.2). If we look at e.g. the first feature where we use the semantic distances between context words and focus words, it is more likely to obtain a low distance if the focus words are often used in conjunction with the context. In a sense this treats dead or conventional metaphors in the same way non-metaphors are treated. This, however, also means that we cannot distinguish clearly between non-metaphors and often used conventional metaphors, because the similarity distance measure we use (i.e. Normalized Google Distance, but true also for any similarity measure based solely on corpus frequencies) does not operate on a directly semantic but only statistical level. Since the task specifically requires identification of *novel* metaphors, this does not pose a big problem - though it has to be mentioned that for the general detection of metaphors, using only frequency based methods will likely not suffice.

---

### 3.4.1 Extracting an n-gram repository

---

That being said, we want to further assess the novelty of the metaphor candidates, to see if improvements can be made nonetheless; for this goal an n-gram based approach is evaluated. Again the Google Books Ngram Corpus is employed, but this time we use the 3-gram files to assess novelty of the candidates which were found by the genitive noun modifier method. This approach could also be applicable to other candidate types like adjective - noun; we cannot use it for verb - direct object candidates, because in such a relation the verb and its object are often not occurring directly side by side, but in a much larger token window.

First, the n-gram information from the GBNC files are extracted for all candidate focus expressions. We also supply the year a document was published manually; this information is then used to extract the volume counts (i.e. the number of books an n-gram is found in) for each candidate n-gram in a time frame of +/- 30 years from the publication year of the document the n-gram is situated in. This cumulative volume count is then normalized by the total amount of books published in that time frame (and indexed by Google), which can be easily harvested from a “total counts” file also supplied by Google. These relative occurrence counts for each n-gram are later employed by filtering out those candidates with a *novelty* above a certain threshold, e.g. average or above 1%.

---

## 4 Evaluation

---

We explain the procedure and results for automatic evaluation of the built system, after first discussing the inter-annotator agreement between three annotators for the dataset. In the subsequent section we present a way to export the found metaphors for review into *CSniper*, a web-based tool for assessment, evaluation and review.

---

### 4.1 Automatic evaluation

---

In order to evaluate our system automatically, we disregard the “fuzziness” the Gaussian mixture model grants us, deriving a binary decision from the cluster membership probabilities. We essentially distribute the rated sentences into two groups which correspond to the clusters, assigning a sentence the cluster it most probable belongs to. It is reasonable to assume our set of candidates contain more non-metaphorical sentences than metaphors, so the automatic labeling process is designed to label the candidates in the smaller group as metaphors.

One problem that has not been mentioned so far is the existence of more than one candidate finding for the same sentence. Suppose we have the sentence as shown in Figure 4.1, which contains two attribute genitive noun modifiers.

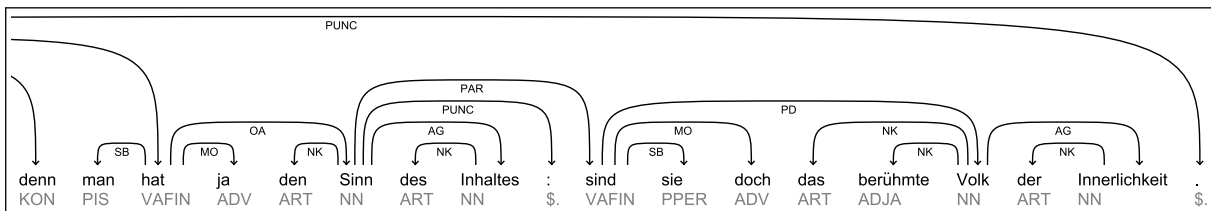


Figure 4.1: Example extract showing two noun - genitive noun modifier constructions in one sentence: “Sinn des Inhaltes” and “Volk der Innerlichkeit”.

This sentence “produces” two candidates, as each construction would have to be investigated on its own. For clustering, we use all candidates, as they represent different focus/context combinations. However, when classifying candidates, we only allow for one candidate to represent a sentence. The simple selection process employed chooses the candidate with the strongest association to one of the clusters.

---

#### 4.1.1 Inter-Annotator Agreement

---

To evaluate the inter-annotator agreement we look at the different annotations which were made by three annotators A1, A2, A3, before the annotations were consolidated into one final gold standard

document. Annotations were automatically divided into three groups: *metaphor*, *suspicion* and *other*, based on comments in the document. *Metaphor* is self-explanatory; *suspicion* means that an annotation has been marked as *Verdacht*, i.e. that the annotated phrase might not be a novel metaphor but e.g. an idiom or a trope<sup>1</sup> other than metaphor. Last, *other* means that the annotator believes that the phrase in question is indeed not a novel metaphor but another kind of figurative language use. In the cases where there are more than one annotation in a sentence, the strongest classification is used - metaphor, suspicion and other, in this order.

For calculating Cohen’s kappa, we merge the classes for suspicion and other. Also added to this group are the sentences without an annotation. This group is labeled *N*, while the group of confident novel metaphor annotations form a second group, labeled *M*. In Tables 4.1 to 4.4 we see kappa scores for four documents.

Annotator combination	Kappa
A1 and A2	0.3567
A1 and A3	0.2647
A2 and A3	0.3115

Table 4.1: Kappa scores for Nietzsche: “Vom Nutzen und Nachteil der Historie ...”

Annotator combination	Kappa
A1 and A2	0.7446
A1 and A3	0.6170
A2 and A3	0.7545

Table 4.2: Kappa scores for Hegel: “Grundlinien der Philosophie des Rechts, Einleitung”

Annotator combination	Kappa
A1 and A2	0.5581
A1 and A3	0.2875
A2 and A3	0.3797

Table 4.3: Kappa scores for Hegel: “Grundlinien der Philosophie des Rechts, Vorrede”

Annotator combination	Kappa
A1 and A2	0.7988
A1 and A3	0.1599
A2 and A3	0.1444

Table 4.4: Kappa scores for Helmholtz: “Über die Erhaltung der Kraft”

Overall, the kappa values are relatively low and again highlight the complexity of annotating metaphors. The values for the Helmholtz text seem strange, but are explained looking at the confusion matrices (cf. Table 4.5). We see that two of the annotators only annotated 2 resp. 3 metaphors, while the third annotator marked 10.<sup>2</sup> Appendix C shows the confusion matrices for the remaining documents.

<sup>1</sup> A trope is a hypernym of various forms of figurative language, e.g. metonymy.

<sup>2</sup> Note that in the manually consolidated gold standard file there are 2 metaphor annotations.

	N	M									
N	449	1	450	N	441	9	450	N	440	9	449
M	0	2	2	M	1	1	2	M	2	1	3
	449	3	452		442	10	452		442	10	452

Table 4.5: Confusion matrices for the Helmholtz text.

---

#### 4.1.2 Evaluation results

---

We use a three stage evaluation process: first we show the results of the baseline, i.e. the original method used by Sporleder and Li, then results incorporating selectional preferences and at the end we analyze whether additional novelty filtering yields better results. For each of the steps, we show results for two derived gold standards - a “strict” gold standard (*METAPHER*), and a more “relaxed” set where also the ambiguous phrases which were labeled “Verdacht” are included in the gold standard (*VERDACHT*).

To conduct an adequate evaluation, we use the clustering to build a classifier: we divide the set of candidates into a 70% training set and a 30% test set. On the training set we conduct a 5-fold crossvalidation for all feature combinations. From the results, we choose the features which result in the best precision and use the whole training set to train the classifier using these features, obtaining model parameters for a “final” Gaussian mixture model. This model is then evaluated on the test set. To have comparable results we also use the 70/30 split to obtain the baseline results. Below we will show the results obtained on the test set<sup>1</sup>

As we use filtering (i.e. candidate extraction) before the clustering, we consequently only include those metaphors in the gold standard which can be found by the respective candidate type. Thus our (purposely limited) goal to only try to find a subset of metaphors is reflected in the recall, considering only those gold metaphors that can be found by the candidate type. A metaphor however is not exclusive to a certain construction (since we use sentences as our base unit of evaluation, some metaphors may also be found by filtering through another construction). The final goal, then, would be to use many different candidate constructions, which all achieve high precision, to find a large enough set of novel metaphors in a given text. The use of precision instead of recall or  $F_1$ -measure as a dimension on which to choose the features is not only motivated in this goal, but also by our approach of narrowing down the candidates after the clustering step; we employ a filtering for selectional preferences and novelty, which can only lead to a worse recall.

The implementation is done using DKPro Lab (Eckart de Castilho and Gurevych, 2011), an experimentation framework which allows us to easily test different configurations of features and

---

<sup>1</sup> Because of the amount of data, the intermediate results on which feature selection was performed are located on the supplied DVD, at DVD://crossvalidation.

parameters, and also provides methods for conducting crossvalidation experiments. It was also used to provide the example export pipeline used in section 4.2.2.

We describe briefly the features implemented, and explain the keys of the result tables. The first four features are the ones used by Li and Sporleder (2010) (cf. section 3.2): *fc* stands for the feature which uses focus-context similarity, *cc* denominates the inner-context similarity, *s* is the difference of the two preceding features and *n* is the feature which assigns 1 for  $s < 0$  and 0 otherwise. New features are *sa*, which is the selectional association (cf. section 3.3.1) of the two terms which represent the candidate (either verb - object relation or noun - genitive noun modifier construction) and *ff*, the inner-focus similarity. Also new is *n10*, which is a “relaxed” *n* feature. If the difference between focus-context similarity and context-context similarity is below ten percent of the context-context similarity, this feature assigns 0; if it is above and *s* positive, it assigns 1,  $-1$  if *s* negative.

The thresholds for the selectional association filtering are *median* and *average*; the former deems all assessed candidates with a selectional association greater than the median of those clustered candidates as not being novel metaphors, the latter uses the average for this threshold. Novelty thresholds tested were *average* and *above001*, using the average of the novelty scores and 0.01 as thresholds respectively.

The baseline implementation uses the approach by Li and Sporleder, i.e. a classifier using Gaussian mixture model estimated by expectation maximization and the features as explained in section 3.2.

gold standard	sa threshold	fc	cc	s	n	sa	ff	n10	TP	FP	TN	FN	precision	recall
METAPHER	none	x	x	x	x	-	-	-	21	90	121	6	0.1892	0.7778
VERDACHT	none	x	x	x	x	-	-	-	30	81	116	11	0.2703	0.7317

Table 4.6: Baseline results for noun - genitive noun modifier candidates

gold standard	sa threshold	fc	cc	s	n	sa	ff	n10	TP	FP	TN	FN	precision	recall
METAPHER	none	x	x	x	x	-	-	-	31	90	218	11	0.2562	0.7381
VERDACHT	none	x	x	x	x	-	-	-	42	79	213	16	0.3471	0.7241

Table 4.7: Baseline results for verb - object candidates

As can be seen in the baseline result tables (cf. Tables 4.6 and 4.7), the precision is low, for noun - genitive noun modifier constructions below 20% using only confident metaphor manual annotations as the gold standard, and 27% if we also include ambiguous (the *suspicion* cases) annotations in the gold standard. The verb - direct object relations tend to be identified easier,

---



---

gold standard	sa	threshold	fc	cc	s	n	sa	ff	n10	TP	FP	TN	FN	precision	recall
METAPHER	median		x	x	x	x	x	-	-	10	32	179	17	0.2381	0.3704
VERDACHT	median		x	x	x	x	x	-	-	15	27	170	26	0.3571	0.3659

---

Table 4.8: Results of the classifier with the best feature set for noun - genitive noun modifier candidates, but without novelty filter

---

gold standard	sa	threshold	fc	cc	s	n	sa	ff	n10	TP	FP	TN	FN	precision	recall
METAPHER	median		x	x	x	-	-	x	x	17	36	272	25	0.3208	0.4048
VERDACHT	median		x	x	x	-	-	x	x	24	29	263	34	0.4528	0.4138

---

Table 4.9: Results of the classifier with the best feature set for verb - object candidates

with 25% resp. 34%. With their GMM approach, Li and Sporleder achieve 40.71% precision on a randomly selected set of 500 verb - noun phrase constructions from the Gigaword corpus (Li and Sporleder, 2010, p. 299); but since they do not distinguish between different kinds of figurative language like idioms and metaphors, these figures are difficult to compare.

If we set the results incorporating selectional association (cf. Tables 4.8 and 4.9) in contrast to our baseline findings, we achieve quite a gain in precision, countered by a hit in recall. As explained earlier, this is to be expected because of the filtering we employ.

We also see that the feature set differs between candidate types — interestingly, the best feature set for verb - direct object candidates determined by the crossvalidation does not include selectional association as a feature, contrary to the noun - genitive noun modifier constructions. This could be the case because we do not take into account the selectional preference strength (SPS) itself, as opposed to the selectional association (SA). As a reminder, in verb - object relations the SPS measures how strong a verb constraints its arguments, while the SA measures how great the proportion of a noun class of a concrete object is, in comparison to the other noun classes.

Also as stated before (cf. section 3.4), we suspect that novelty of found metaphors is to a degree implicitly ensured through the use of frequency based measures for the feature vector creation and for creating the selectional preference repositories. Nonetheless, we want to see if our novelty filter can improve our results. Since it is only implemented for the first type of candidates (noun - genitive noun modifier), we show these results in tables 4.10.

We can again observe an increase in precision, while the recall drops are not nearly as big compared to the changes we see from the baseline to the best-feature version. In fact, considering the absolute numbers, we find that e.g. in the *METAPHER* gold standard only one formerly correctly found metaphor now is not found anymore. This, in conjunction with the increased

---



---

gold std.	sa threshold	novelty	fc	cc	s	n	sa	ff	n10	TP	FP	TN	FN	precision	recall
METAPHER	median	average	x	x	x	x	x	-	-	9	24	187	18	0.2727	0.3333
VERDACHT	median	average	x	x	x	x	x	-	-	13	20	177	28	0.3939	0.3171

---

Table 4.10: Results of the classifier with the best feature set for noun - genitive noun modifier candidates, but with novelty filter

precision, shows that the novelty method can indeed be quite helpful for enhancing the assessment, contrary to our earlier assumption.

Overall we can observe that the precision for verb - direct object candidates is seven to ten percentage points higher than for the noun - genitive noun modifier candidates, depending on the gold standard which is employed. This is true even for the baseline, so can not be purely ascribed to the different methods of harvesting selectional preferences for the different candidate types. Nonetheless, verbs being more assertive regarding their context than nouns could be one reason for this disparity.

---

## 4.2 Export and Manual Review

---

To review the results of the evaluation (i.e. the sentences which were classified as containing metaphors) we need to export the candidates along with their assessment. We include an example pipeline for *CSniper* export, which is limited to the documents used in the evaluation task (cf. section 5.1).

---

### 4.2.1 CSniper

---

*CSniper* (Eckart de Castilho et al., 2012) is a query-driven web-based annotation tool, with focus on assessment and evaluation. It has been created as part of the LOEWE (Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz) program *Digital Humanities* at the UKP Lab TU Darmstadt and is available as open source software<sup>1</sup>. The main goal of *CSniper* is to facilitate multi-user assessment and evaluation of certain grammatical structures, namely non-canonical constructions. It also provides a rich statistics overview, e.g. showing inter-annotator agreement with different options for filtering, and also enables analysis of the separate evaluation items.

To explain the extensions added to *CSniper*, we describe in short some of the concepts and terms used. The data on which search shall be made possible on is organized by *CSniper* in corpora, containing documents. For each corpus exists a folder, containing *engine* folders. Such

---

<sup>1</sup> <http://code.google.com/p/csniper/>, last accessed: 2013-06-15



an engine folder then contains the documents of the corpus in the format the query engine requires (an example for such a query engine would be the *Corpus Query Processor* or *CQP*, which is used in the Open Corpus Workbench (Evert and Hardie, 2011)). When querying a corpus using the assessment page, CSniper produces *EvaluationItems*. Those consist of a sentence which matches the query, offset information, identifiers for corpus and document in which the sentence is situated, and the type of annotation that the user selected (from a user-created list) as an expected result of the query. Such an *AnnotationType* can be e.g. “It-Cleft” (the original use case) or for our usage, “novel metaphor”. An *EvaluationResult* encapsulates an *EvaluationItem* and provides an assessment field.

#### 4.2.2 Integrating manual metaphor classification results into CSniper

Although CSniper has initially been developed to simplify classification of non-canonical constructions, it can also be extended to evaluate other text phenomena, including metaphors.

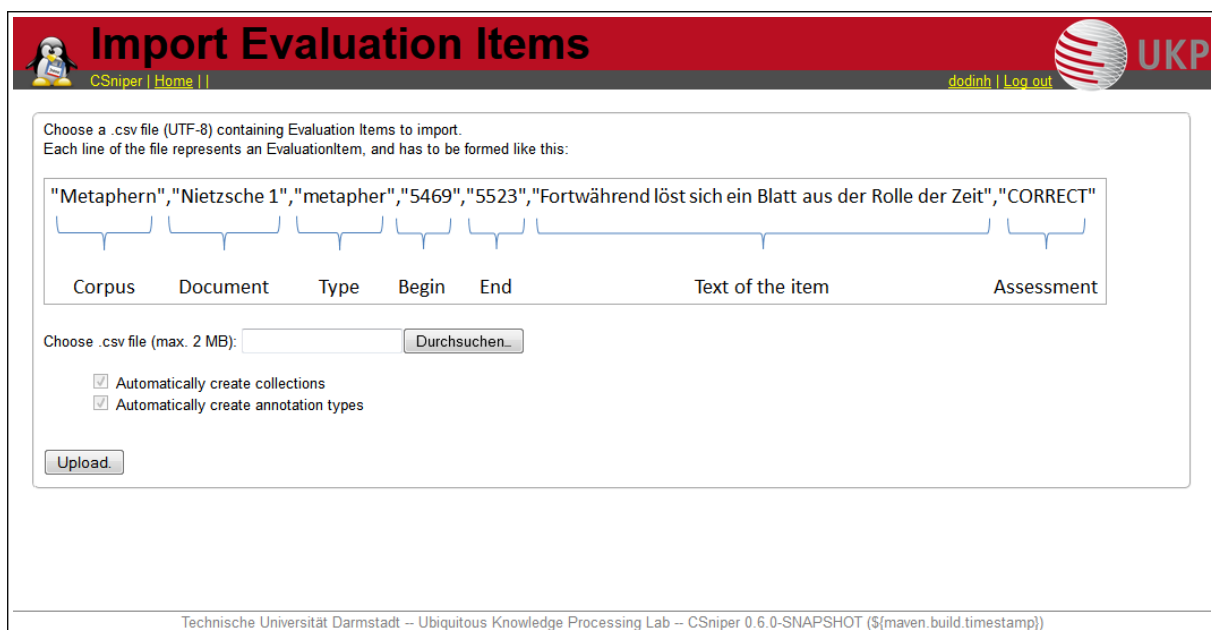


Figure 4.2: The import page in CSniper.

The ultimate goal of directly running the Gaussian mixture model classifier in CSniper is not feasible at this time because the creation of the feature vectors is too time consuming — for long sentences of over 40 tokens the lookups and calculations involved in determining the semantic similarities can take up more than 30 seconds. Such long processing times are not practical for a web-based application. Instead, a slightly more intricate approach is taken: the vectors are constructed locally, and the Gaussian mixture model classifier is used in conjunction with the selectional preference and novelty filter to cluster them. Then we can export the classified sentences

along with their assessment to a CSV (Comma Separated Value) file, which we eventually import into CSniper. This requires us to extend CSniper, more precisely to add an import feature. Such a feature would have the additional benefit of not being restrained to metaphors, but useful for any assessment which takes place outside of CSniper. Importing such results into CSniper then makes it possible to evaluate them on a multi-user basis.

On the implementation side we created a new page where CSV files can be uploaded (cf. Figure 4.2). Such a file consists of multiple EvaluationItems, one per line, containing: collectionId, documentId, type, beginOffset, endOffset, coveredText, assessment. The last column (assessment) is optional.

The screenshot shows the 'Query and Assessment' interface. At the top, there's a navigation bar with 'CSniper | Home | Help | Macros' and 'dodinh | Log out' next to the 'UKP' logo. Below this, there are dropdown menus for 'Corpus: METAPHORS' and 'Type: metaphor'. A red bar contains buttons for 'Query', 'Review', 'Complete', 'Sample sets', and 'Find'. The main form includes fields for 'Query engine: database', 'Query history: Choose One', 'Query: regex.', and 'Query comment:'. A 'Submit query' button is at the bottom right of the form. Below the form is a 'Filter' section with radio buttons for 'all items', 'not assessed items', 'assessed items', and 'items to check'. A status bar shows 'Showing 61 to 70 of 459' and pagination controls. The main content is a table with columns 'Doc', 'Match', 'Label', and 'Comment'. The table lists several entries with their respective matches and labels (WRONG, Correct, CORRECT, WRONG, WRONG). At the bottom, there's a footer with 'Technische Universität Darmstadt - Ubiquitous Knowledge Processing Lab - CSniper 0.6.0-SNAPSHOT (\$[maven.build.timestamp])'.

Doc	Match	Label	Comment
Nietzsche-Vom-Nutzen-und-Nachteil-1-5-Final.docx	Mag unsre Schätzung des Historischen nur ein okzidentalisches Vorurteil sein; wenn wir nur wenigstens innerhalb dieser Vorurteile fortschreiten und nicht stillstehn!	WRONG	...
Nietzsche-Vom-Nutzen-und-Nachteil-1-5-Final.docx	Fortwährend löst sich ein Blatt aus der Rolle der Zeit, fällt heraus, flattert fort – und flattert plötzlich wieder zurück, dem Menschen in den Schoß.	Correct	...
Nietzsche-Vom-Nutzen-und-Nachteil-1-5-Final.docx	1 [211] Betrachte die Herde, die an dir vorüberweidet: sie weiß nicht, was Gestern, was Heute ist, springt umher, frißt, ruht, verdaut, springt wieder, und so vom Morgen bis zur Nacht und von Tage zu Tage, kurz angebunden mit ihrer Lust und Unlust, nämlich an den Pflock des Augenblicks, und deshalb weder schwermütig noch überdrüssig.	CORRECT	...
Nietzsche-Vom-Nutzen-und-Nachteil-1-5-Final.docx	Zumeist winkt ihm kein Lohn, wenn nicht der Ruhm, das heißt die Anwartschaft auf einen Ehrenplatz im Tempel der Historie, wo er selbst wieder den Späterkommenden Lehrer, Tröster und Warner sein kann.	WRONG	...
Nietzsche-Vom-Nutzen-und-Nachteil-1-5-Final.docx	Indem man zum Natürlichen zurückzuflihen glaubte, erwählte man nur das Sichgehenlassen, die Bequemlichkeit und das möglichst kleine Maß von Selbstüberwindung.	WRONG	...

Figure 4.3: CSniper assessment page showing imported metaphor candidates with assessment.

The EvaluationItems are parsed and saved in the underlying database. If an assessment is specified for an EvaluationItem, an EvaluationResult is created containing the item and the assessment. Missing AnnotationTypes are created automatically, as well as missing corpus folders. In the affected corpus directories, an “artificial” engine directory is set up, named “database”. Contrary to the regular engine directories, “database” is empty and only serves as an indicator to CSniper to enable the newly implemented DatabaseEngine for the corpus in question.

Users are now able to use the “Search” and “Assessment” features to review the imported EvaluationItems (in our use case, the sentences which have been classified as metaphors), as can be seen in Figure 4.3. The “query” field on these pages acts as a filter on the covered text of the

---

EvaluationItem. It supports regular expressions (Java style) when preceded by “regex:”. So to see all imported items the user can query for “regex:”.

The changes made to CSniper do not change how the underlying system deals with data, they merely add to it; since they are unobtrusive, these changes can be readily incorporated to facilitate the inclusion of different kinds of experiments into CSniper with the goal of manual review and evaluation.

---

## 5 Summary

---

This work is concerned with the construction and subsequent evaluation of a method for novel metaphor identification. To use it in a real world setting, additional steps have to be undertaken; those will be discussed in the next section. Finally, we will present a summary of the work.

---

### 5.1 Outlook

---

To use the presented combination of methods for novel metaphor identification on arbitrary texts (instead of only using the evaluation dataset), several points need to be addressed.

The first issue is the small scale of the selectional preference inventory for nouns as head (only covering the dataset documents) and the relatively small sample size used for the creation of the repository for verb - object selectional preferences (cf. section 3.3). Using the full deWaC instead of the first 60.000 documents should present a more mature repository for both types of candidates, which could improve the selectional preference filtering. Although this would be very time-intensive because of the dependency parsing of 1.75 million documents, it is a one-time task, which — given enough time and resources — should not pose a big problem. When we have such a large resource readily parsed, we can also think about including other types of candidates in the identification process, such as adjective - noun pairs, or subject - verb relations.

By using Lucene for document count lookup of words for the calculation of similarity, we already employ a state-of-the-art retrieval solution; nonetheless, lookup is still somewhat slow. Possible alternatives could be for example NoSQL databases like MongoDB<sup>1</sup> or key-value stores like Tokyo Cabinet<sup>2</sup> or BerkeleyDB (Olson et al., 1999), although all those examples proved to be slower than a non-optimized Lucene index for indexing — and more importantly querying — the complete deWaC corpus.

Speed is also of concern when it comes to the estimation of the parameters for the GMM. Especially costly in terms of processing time are the calculations involved in inverting the covariance matrix  $\Sigma_j$  in the expectation maximization algorithm. While quite slow with the current Java implementation, this could probably be sped up by an implementation in native code.

Retrieving the n-gram information from Google Books Ngram Corpus also takes time; it could however also be useful for other types of candidates such as adjective - noun pairs. To speed up the n-gram extraction, one possibility is to recompress the GBNC so that the files are internally sorted, which would allow for an adapted binary search (Grabowski and Swacha, 2012). For files with over 100 million lines, the time saved moving from the implemented  $O(n)$  solution to a method in  $O(\log n)$  should be substantial.

---

<sup>1</sup> <http://www.mongodb.org/>, last accessed: 2013-06-15

<sup>2</sup> <http://fallabs.com/tokyocabinet/>, last accessed: 2013-06-15

---

---

Finally, when reproducibility is not the main focus (as opposed to the evaluation task followed in this work), we can forego the intermediate steps of saving and retrieving candidates and vectors from the database. Instead, a document can then be processed in one simple pipeline that handles reading, preprocessing, candidate extraction, clustering, selectional preference and novelty filtering and export. We showed an example for export in section 4.2.2, but the modular nature of UIMA allows for convenient implementation of a writer for many different export formats, e.g. any kind of custom XML format.

---

## 5.2 Summary

---

We presented a combined approach for identifying novel metaphors in German texts, incorporating Gaussian mixture model clustering, selectional preferences and n-gram frequency filtering.

We first discussed two important ideas of metaphor, the *interaction theory* by Black (1955) and the theory of *conceptual metaphor* by Lakoff and Johnson (2003). While not fundamentally different, the former provides us with a structural description of (a textual instance of) metaphors as consisting of a *context* and a metaphorically used *focus*. It is emphasized that such a focus can have a literal meaning when used together with another context; to create *metaphoricity*, focus and context have to *interact* in a special way, invoking “system[s] of associated commonplaces” in the reader. The theory by Lakoff and Johnson somewhat extends and abstracts this idea of a system of associated commonplaces to propose a *system of conceptual metaphors* which every human possesses. They distinguish between *conventional* and *new* metaphors — the former being well integrated in our conceptual system, the latter presenting a new line of association between two concepts. In formulating our *working definition* of metaphor we incorporated the structural break between context and focus, as well as the notion of novelty, which we translated to low frequency of textual instances.

We described our dataset, six technical and philosophical historical documents in German, from the 18th to early 20th century, which were annotated by philosophers for novel metaphors. Pre-processing is briefly discussed, using Apache UIMA as a framework in conjunction with different modules from DKPro Core, a selection of state-of-the-art NLP processing tools. We described our method and the motivation behind *candidate* extraction, choosing noun - genitive noun modifier constructions and verb - direct object relations as candidates.

Following an approach by Li and Sporleder (2010) for identifying figurative language in text, we employed clustering with *Gaussian mixture models* to utilize the semantic break between focus and context. We described the *expectation maximization* algorithm used to estimate parameters for this clustering process. This process also involves semantic similarity features based on the *Normalized Google Distance*, a semantic distance measure between words. For computation of these similarity we used lemma counts from *deWaC*, a large German web page corpus. To improve on this baseline we used selectional preferences, both as features in the Gaussian mixture model

---

clustering and as a filter afterwards, to improve precision. For this task, we built selectional preference repositories for German noun - genitive noun modifier relations and verb - direct object dependencies, using the *Google Books Ngram Corpus* (a large diachronic n-gram and dependency corpus) and deWaC. For noun - genitive noun modifier constructs we also implemented an n-gram filter which considers the time period in which a document was written, also employing the Google Books Ngram Corpus.

We briefly discussed kappa scores from three annotators on four of the documents in the dataset, finding a mixed inter-annotator agreement with an average of 0.45. We also explained the two gold standards used, one containing unambiguous cases of novel metaphors, the other including uncertain annotations. Evaluation was done using a semi-supervised approach, in that we conducted a 5-fold crossvalidation on 70% of the dataset, the results of which were used for feature selection. The best performing features in terms of precision were then used to train on the whole training set, using the resulting model as a classifier. This classifier was subsequently tested on the remaining 30% of the dataset. For noun - genitive noun modifier constructions we found that the baseline delivers low results in terms of precision; only 27% of the candidates classified as novel metaphors were annotated at least as ambiguously novel. This baseline could be significantly improved to a precision of 35% by employing selectional preferences as features and for filtering, and further improved using n-gram based novelty filters, resulting in a precision of 39%. For verb - direct object relations we observed a similar improvement when incorporating selectional preferences, although figures were generally higher; starting at 34% precision obtained using the baseline and the ambiguous gold standard, we achieved a precision of 45% incorporating selectional preferences and additional features.

We then showed an export example for displaying and reviewing the annotated novel metaphor candidates in CSniper, a web-based tool for multi-user assessment and evaluation. An import feature was implemented, along with a basic search facility. This allows users to display and review the automatically annotated metaphors.

In chapter 5.1 we described what work has to be conducted for the implemented method to be used in practice, but also left some notes on how the quality of the assessment could possibly be further improved. This includes building a potentially more robust selectional preference repository using more data, or tweaking the n-gram novelty detection method, as well as searching for candidates of other type.

Because the task of identifying novel metaphors is rather special compared to finding general figurative language, it cannot be easily compared to other methods; thus for future research, it would be interesting to see how other approaches — e.g. using supervised techniques such as support vector machines instead of clustering — handle the identification of novel metaphors.



# Appendices

---

## A Manual structural categorization of metaphors

---

These are coarse grained manual structural categorizations of the novel metaphors annotated in “Nietzsche-Vom-Nutzen-und-Nachteil-1-5-Final.docx”, which can be found on the supplied DVD at DVD://documents/Nietzsche-Vom-Nutzen-und-Nachteil-1-5-Final.docx.

---

Page	Marked phrase	Criterion
1	wir alle an einem verzehrenden historischen <b>Fieber</b> leiden	ADJ N
2	kurz angebunden mit ihrer Lust und Unlust, nämlich an den Pflock <b>des Augenblicks</b>	innerhalb NP
2	das Vergessen nicht lernen zu können und immerfort am Vergangenen zu hängen: mag er noch so weit, noch so schnell laufen, <b>die Kette</b> läuft mit	VP subj
2	Fortwährend löst sich <b>ein Blatt aus der Rolle</b> der Zeit, fällt heraus, <b>flattert fort – und flattert</b> plötzlich wieder zurück, dem Menschen in den Schoß	innerhalb NP
2	es geht auf in der Gegenwart, wie <b>eine Zahl</b> , ohne daß <b>ein wunderlicher Bruch</b> übrigbleibt	Vergleich
2	Der Mensch hingegen [212] <b>stemmt sich gegen</b> die große und immer größere <b>Last des Vergangenen</b> : diese <b>drückt ihn nieder oder beugt ihn seitwärts</b> , diese <b>beschwert seinen Gang</b> als <b>eine unsichtbare und dunkle Bürde</b> , welche er zum Scheine einmal verleugnen kann	innerhalb NP
2	zwischen <b>den Zäunen</b> der Vergangenheit und der Zukunft in überseliger Blindheit spielt	innerhalb NP
2	<b>drückt damit das Siegel</b> auf jene Erkenntnis – daß Dasein nur ein ununterbrochenes Gewesensein ist	VP subj vs obj
3	Es gibt Menschen, die diese Kraft so wenig besitzen, daß sie an einem einzigen Erlebnis, an einem einzigen Schmerz, oft zumal an einem einzigen zarten Unrecht, wie an einem ganz kleinen blutigen Risse <b>unheilbar verbluten</b>	
3	alles Vergangene, eigenes und [214] fremdestes, würde sie an sich heran-, in sich hineinziehen und gleichsam <b>zu Blut umschaffen</b>	VP obj
3	innerhalb eines fremden den eigenen Blick <b>einzuschließen</b>	VP obj
4	aus dem viel <b>zarteren Netze</b> seiner Gerechtigkeiten und Wahrheiten nicht wieder zum derben Wollen und Begehren <b>herauswinden</b>	innerhalb NP

---



Page	Marked phrase	Criterion
4	erst dadurch, daß der Mensch denkend, überdenkend, vergleichend, trennend, zusammenschließend jenes unhistorische Element einschränkt, erst dadurch, daß innerhalb jener <b>umschließenden Dunstwolke ein heller blitzender Lichtschein</b> entsteht	
4	ohne <b>jene Hülle</b> des Unhistorischen würde er nie angefangen haben	innerhalb NP
4	Wo finden sich Taten, die der Mensch zu tun vermöchte, ohne vorher in <b>jene Dunstsicht</b> des Unhistorischen eingegangen zu sein	innerhalb NP
4	Es ist der ungerechteste Zustand von der Welt, eng, undankbar gegen das Vergangene, blind gegen Gefahren, taub gegen Warnungen, <b>ein kleiner lebendiger Wirbel in einem toten Meere von Nacht und Vergessen</b> : und doch ist dieser Zustand – unhistorisch, widerhistorisch durch und durch – <b>der Geburtsschoß</b> nicht nur einer ungerechten, sondern vielmehr jeder rechten Tat	innerhalb NP
5	the <b>dregs</b> of life	innerhalb NP
5	so erleuchtet sich der überhistorische Denker [218] alle Geschichte der Völker und der einzelnen von innen heraus, hellseherisch den Ursinn der verschiedenen <b>Hieroglyphen</b> erratend und allmählich sogar der immer neu <b>hinzuströmenden Zeichenschrift</b> ermüdet ausweichend	ADJ N
6	wenn wir nur wenigstens innerhalb dieser Vorurteile <b>fortschreiten und nicht stillstehn</b>	
6	bei einem gewissen Übermaß derselben <b>zerbröckelt</b> und entartet das Leben	VP subj
7	Daß der Tätige mitten unter den schwächlichen und hoffnungslosen Müßiggängern, mitten unter den scheinbar tätigen, in Wahrheit nur aufgeregten und zappelnden Genossen nicht verzage und Ekel empfinde, blickt er hinter sich und <b>unterbricht den Lauf</b> zu seinem Ziele, um einmal aufzuatmen.	VP subj vs obj
7	in ihnen ein <b>Höhenzug</b> der Menschheit durch Jahrtausende hin sich verbinde	innerhalb NP
7	Die dumpfe Gewöhnung, das Kleine und Niedrige, alle Winkel der Welt erfüllend, als <b>schwere Erdenluft</b> um alles Große <b>qualmend, wirft sich hemmend, täuschend, dämpfend, erstickend</b> in den Weg, den das Große zur Unsterblichkeit zu gehen hat.	Vergleich
7	Wer möchte bei ihnen jenen <b>schwierigen Fackel-Wettlauf</b> der monumentalen Historie vermuten, durch den allein das Große weiterlebt!	innerhalb NP

Page	Marked phrase	Criterion
7	Aber eines wird leben, das <b>Monogramm</b> ihres eigensten Wesens, ein Werk, eine Tat, eine seltene Erleuchtung, eine Schöpfung	innerhalb NP
8	wie müßte es ihn bestärken wahrzunehmen, daß die Kultur der Renaissance sich <b>auf den Schultern</b> einer solchen Hundert-Männer-Schar heraushob	VP subj vs obj
8	wie gewaltsam muß die Individualität des Vergangnen in eine allgemeine Form hineingezwängt und an allen <b>scharfen Ecken und Linien</b> zugunsten der Übereinstimmung zerbrochen werden!	VP subj vs obj
8	wenn es feststünde, daß dieselbe <b>Verknötung</b> von Motiven, derselbe deus ex machina, dieselbe Katastrophe in bestimmten Zwischenräumen wiederkehrten	innerhalb NP
8	ein solcher »Effekt an sich«: er ist es, der die Ehrgeizigen nicht schlafen läßt, der den Unternehmenden <b>wie ein Amulett</b> [223] am Herzen liegt	Vergleich
8	Solange <b>die Seele</b> der Geschichtsschreibung in den großen Antrieben liegt	innerhalb NP
8	ganze große Teile derselben werden vergessen, verachtet, und <b>fließen fort</b> wie <b>eine graue ununterbrochene Flut</b> , und nur einzelne geschmückte Fakta heben sich als <b>Inseln</b> heraus	Vergleich, VP subj
9	so gebärden sie sich als <b>Ärzte</b> , während sie es im Grunde <b>auf Giftmischerei</b> abgesehen haben	Vergleich
9	so bilden sie ihre Zunge und ihren Geschmack aus, um aus ihrer Verwöhntheit zu erklären, warum sie alles das, was ihnen von <b>nahrhafter Kunstspeise</b> angeboten wird, so beharrlich ablehnen.	zusammengesetztes N
9	Die monumentalische Historie ist das [225] <b>Maskenkleid</b> , in dem sich ihr Haß gegen die Mächtigen und Großen ihrer Zeit für gesättigte Bewunderung der Mächtigen und Großen vergangner Zeiten ausgibt	Is-a
10	Der Besitz von <b>Urväter-Hausrat</b> verändert in einer solchen Seele seinen Begriff	zusammengesetztes N
10	die bewahrende und verehrende Seele des antiquarischen Menschen in diese Dinge <b>übersiedelt</b> und sich darin ein heimisches <b>Nest</b> bereitet	VP subj
10	Mitunter grüßt er selbst über weite <b>verdunkelnde und verwirrende</b> Jahrhunderte hinweg die Seele seines Volkes	ADJ N
10	ein Hindurchfühlen und Herausahnen, ein <b>Wittern</b> auf fast verlöschten Spuren	Vergleich

Page	Marked phrase	Criterion
10	ein instinktives Richtig-Lesen der noch so überschriebenen Vergangenheit, ein rasches Verstehen der <b>Palimpseste, ja Polypseste</b> – das sind seine Gaben und Tugenden.	
10	zerriß der historische, <b>zwischen ihnen ausgebreitete Wolkenschleier</b>	ADJ N
10	Mitunter sieht es wie Eigensinn und Unverstand aus, was den einzelnen an diese Gesellen und Umgebungen, an diese mühselige Gewohnheit, an diesen kahlen Bergrücken <b>gleichsam festschraubt</b>	
11	wenn der historische Sinn das Leben nicht mehr konserviert, sondern <b>mumisiert</b>	VP subj
11	Dann erblickt man wohl das <b>widrige Schauspiel</b> einer <b>blinden Sammelwut</b>	innerhalb NP
11	Der Mensch <b>hüllt sich in Moderduft</b>	VP obj
11	mit jeder Kost zufrieden ist und mit Lust selbst den Staub bibliographischer Quisquilien <b>frißt</b>	VP obj
11	wenn die antiquarische Historie das Fundament, auf dem sie allein zum Heile des Lebens <b>wurzeln</b> kann, nicht verliert	VP subj
11	falls sie nämlich allzu mächtig wird und die andren Arten, die Vergangenheit zu betrachten, <b>überwuchert</b>	VP subj
12	Er muß die Kraft haben und von Zeit zu Zeit anwenden, eine Vergangenheit <b>zu zerbrechen und aufzulösen</b>	VP obj
12	dies erreicht er dadurch, daß er sie <b>vor Gericht zieht, peinlich inquiriert</b> und endlich <b>verurteilt</b>	VP obj
12	Dann wird seine Vergangenheit kritisch betrachtet, dann <b>greift</b> man <b>mit dem Messer an seine Wurzeln</b> dann <b>schreitet</b> man grausam über alle Pietäten hinweg.	VP obj
12	Menschen oder Zeiten, die auf diese Weise dem Leben dienen, daß sie eine Vergangenheit <b>richten</b> und vernichten, sind immer gefährliche und gefährdete Menschen und Zeiten.	VP obj
12	Denn da wir nun einmal die [230] Resultate früherer Geschlechter sind, sind wir auch die Resultate ihrer Verirrungen, Leidenschaften und Irrtümer, ja Verbrechen; es ist nicht möglich, sich ganz von dieser <b>Kette</b> zu lösen.	
12	wir pflanzen eine neue Gewöhnung, einen neuen Instinkt, eine zweite Natur an, so daß die erste Natur <b>abdorrt</b> .	Is-a
13	nicht zur Schwächung der Gegenwart, nicht zur <b>Entwurzelung</b> einer lebenskräftigen Zukunft	innerhalb NP

Page	Marked phrase	Criterion
13	hat sich wirklich die Konstellation von Leben und Historie verändert, dadurch, daß ein mächtig <b>feindseliges Gestirn</b> zwischen sie getreten ist?	ADJ N
13	Es ist allerdings ein solches Gestirn, ein <b>leuchtendes und herrliches Gestirn</b> dazwischengetreten, die Konstellation ist wirklich verändert – durch die Wissenschaft	
13	das Gedächtnis öffnet alle seine <b>Tore</b> und ist doch nicht weit genug geöffnet	VP subj vs obj
13	die Natur bemüht sich aufs höchste, diese <b>fremden Gäste</b> zu empfangen, zu <b>ordnen</b> und zu ehren	VP subj
13	Die Gewöhnung an ein solches unordentliches, stürmisches und kämpfendes <b>Hauswesen</b> wird allmählich zu einer zweiten Natur	zusammengesetztes N
13	Der moderne Mensch schleppt zuletzt eine ungeheuere Menge von unverdaulichen <b>Wissenssteinen</b> mit sich herum, die dann bei Gelegenheit auch ordentlich im Leibe <b>rumpeln</b> , wie es im Märchen heißt.	zusammengesetztes N
13	Das dagegen, was wirklich Motiv ist und was als Tat sichtbar nach außen tritt, bedeutet dann oft nicht viel mehr als eine gleichgültige Konvention, eine klägliche Nachahmung oder selbst eine <b>rohe Fratze</b> .	Vergleich
14	wenn nur immer neue wissenschaftliche Dinge hinzuströmen, die sauberlich in den <b>Kästen</b> jenes Gedächtnisses <b>aufgestellt</b> werden können	innerhalb NP
14	Die Form gilt uns Deutschen gemeinhin als eine Konvention, als <b>Verkleidung und Verstellung</b>	Vergleich
14	aus der <b>Schule</b> der Konvention <b>entlaufen</b> , ließ er sich nun gehen	innerhalb NP
15	Indem man zum Natürlichen <b>zurückzuflihen</b> glaubte, erwählte man nur das Sichgehenlassen	VP obj
15	aber als Ganzes bleibt es schwach, weil alle die <b>schönen Fasern</b> nicht in einen <b>kräftigen Knoten</b> geschlungen sind	
15	ein schwächerer oder roher Versuch <b>irgendeiner Faser</b> , zum Schein einmal für das Ganze gelten zu wollen	innerhalb NP
15	Wenn nur nicht gerade diese Bücher neuerdings mehr als je einen Zweifel darüber erweckten, ob die berühmte Innerlichkeit wirklich noch in ihrem unzugänglichen <b>Tempelchen</b> sitze	VP subj

Page	Marked phrase	Criterion
15	Fast ebenso schrecklich, als wenn jene Innerlichkeit, ohne daß man es sehen könnte, gefälscht, gefärbt, übermalt darinsäße und zur <b>Schauspielerin</b> , wenn nicht zu Schlimmerem geworden wäre	Is-a
15	wir wissen kaum mehr, wie sich die Empfindung bei unseren Zeitgenossen äußert; wir lassen sie [236] <b>Sprünge</b> machen	VP subj vs obj
15	Was soll noch gehofft, noch geglaubt werden, wenn der Quell des Glaubens und Hoffens getrübt ist, wenn die Innerlichkeit gelernt hat, <b>Sprünge</b> zu machen, <b>zu tanzen, sich zu schminken</b> , mit Abstraktion und Berechnung sich zu äußern und sich selbst allgemach zu verlieren!	VP subj vs obj
16	Vielleicht <b>vergräbt</b> er seinen <b>Schatz</b> jetzt lieber, weil er Ekel empfindet, von einer Sekte anspruchsvoll patronisiert zu werden, während sein Herz voll von Mitleid mit allen ist	
16	so tauscht er die tiefe [237] Einsicht seines Schicksals gegen die göttliche Lust des Schaffenden und Helfenden ein und endet als einsamer Wissender, als <b>übersatter</b> Weiser.	ADJ N
16	jener <b>Riß</b> zwischen dem Innen und dem Außen muß unter den <b>Hammerschlägen</b> der Not wieder verschwinden.	innerhalb NP
16	mit vollen Händen <b>ausstreuend</b> , hofft er ein Bedürfnis zu <b>pflanzen</b>	VP obj
16	wie er sich selbst unter dem einströmenden Fremden verlor und bei dem kosmopolitischen <b>Götter-, Sitten- und Künste-Karneval</b> entartete, so muß es dem modernen Menschen ergehen, der sich fortwährend das Fest einer <b>Weltausstellung</b> durch seine historischen Künstler bereiten läßt	zusammengesetztes N
17	eure Taten sind plötzliche <b>Schläge</b> , keine <b>rollenden Donner</b>	Is-a
17	Denn die Kunst flieht, wenn ihr eure Taten sofort mit dem historischen <b>Zeltdach</b> überspannt	VP obj
17	trotz der vielen <b>schlaunen Fältchen</b> seiner pergamentnen Züge und der virtuosens Übung seiner Finger, <b>das Verwickelte aufzuwickeln</b>	ADJ N
17	er kann nun nicht mehr, dem »göttlichen Tiere« vertrauend, die Zügel hängen lassen, wenn sein Verstand <b>schwankt</b> und sein <b>Weg durch Wüsten führt</b> .	VP subj
17	Sieht man einmal aufs Äußerliche, so bemerkt[239] man, wie die Austreibung der Instinkte durch Historie die Menschen fast zu lauter abstractis und <b>Schatten</b> umgeschaffen hat	

Page	Marked phrase	Criterion
17	Greift man solche <b>Masken</b> an, weil man glaubt, es sei ihnen ernst und nicht bloß um ein Puppenspiel zu tun – da sie allesamt den Ernst affichieren –, so hat man plötzlich nur Lumpen und bunte Flicker in den Händen.	VP obj
17	Oder sollte als Wächter des großen geschichtlichen <b>Welt-Harems</b> ein Geschlecht von Eunuchen nötig sein?	zusammengesetztes N
17	an die Stelle jener ängstlich versteckenden Konvention und <b>Maskerade</b> können dann, als wahre Helferinnen, Kunst und Religion treten, um gemeinsam eine Kultur <b>anzupflanzen</b> , die wahren Bedürfnissen entspricht	ADJ N
17	die wahrhaftigste aller Wissenschaften, die ehrliche <b>nackte Göttin</b> Philosophie	innerhalb NP
18	wäre der moderne Mensch überhaupt nur mutig und entschlossen, wäre er nicht selbst in seinen Feindschaften nur ein innerliches Wesen: er würde sie verbannen; so begnügt er sich, ihre <b>Nudität</b> schamhaft zu verkleiden.	
18	weil sie keine Menschen sind, sondern nur <b>eingefleischte Kompendien</b>	ADJ N
18	Und da euch das Ewig-Weibliche nie hinanzieht wird, so zieht ihr es zu euch herab und nehmt, als Neutra, auch die Geschichte [242] als <b>ein Neutrum</b> .	Vergleich
19	sind sie doch selbst weder Mann noch Weib, nicht einmal Kommunia, sondern <b>immer nur Neutra</b> oder, gebildeter ausgedrückt, eben nur die Ewig-Objektiven.	
19	sofort sieht der <b>ausgehöhlte</b> Bildungsmensch über das Werk hinweg und fragt nach der Historie des Autors.	ADJ N

---

## B Derivation of $\Sigma_j^{(t+1)}$

---

In this appendix section we show the intermediate steps for the derivation of  $\Sigma_j^{(t+1)}$ . We start with

$$\frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \Sigma_j^{-1}} = \frac{\partial}{\partial \Sigma_j^{-1}} \sum_{i=1}^N \sum_{k=1}^K h_{ik}^{(t)} \log(\pi_k \mathbf{n}(x_i \mid \mu_k, \Sigma_k)).$$

Re-arranging and losing terms independent of  $\Sigma_j^{-1}$ , we see that

$$\begin{aligned} & \frac{\partial}{\partial \Sigma_j^{-1}} \sum_{i=1}^N \sum_{k=1}^K h_{ik}^{(t)} \log(\pi_k \mathbf{n}(x_i \mid \mu_k, \Sigma_k)) \\ &= \frac{\partial}{\partial \Sigma_j^{-1}} \sum_{k=1}^K \sum_{i=1}^N h_{ik}^{(t)} \left( \log(\pi_k) + \frac{1}{2} \log((2\pi)^{-d}) + \frac{1}{2} \log(|\Sigma_k^{-1}|) - \frac{1}{2} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right) \\ &= \frac{\partial}{\partial \Sigma_j^{-1}} \sum_{i=1}^N h_{ij}^{(t)} \left( \frac{1}{2} \log(|\Sigma_j^{-1}|) - \frac{1}{2} (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) \right). \end{aligned}$$

For maximization, equating the derivative with 0 (and multiplying by 2) yields:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \Sigma_j^{-1}} \sum_{i=1}^N h_{ij}^{(t)} \left( \log(|\Sigma_j^{-1}|) - (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) \right) \\ 0 &= \frac{\partial}{\partial \Sigma_j^{-1}} \left[ \sum_{i=1}^N h_{ij}^{(t)} \log(|\Sigma_j^{-1}|) - \sum_{i=1}^N h_{ij}^{(t)} (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) \right] \\ 0 &= \frac{\partial}{\partial \Sigma_j^{-1}} \left[ \sum_{i=1}^N h_{ij}^{(t)} \log(|\Sigma_j^{-1}|) - \sum_{i=1}^N h_{ij}^{(t)} \text{tr} \left( (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) \right) \right] \\ 0 &= \frac{\partial}{\partial \Sigma_j^{-1}} \left[ \sum_{i=1}^N h_{ij}^{(t)} \log(|\Sigma_j^{-1}|) - \sum_{i=1}^N h_{ij}^{(t)} \text{tr} \left( \Sigma_j^{-1} (x_i - \mu_j) (x_i - \mu_j)^\top \right) \right]. \end{aligned}$$

Here we used  $\text{tr}(c) = c$  for  $c \in \mathbb{R}$ , and  $\text{tr}(AB) = \text{tr}(BA)$  for  $A$  being an  $m \times n$  matrix and  $B$  an  $n \times m$  matrix.

From Searle (1982, pp. 336–337) we also have for a symmetric matrix  $X$  and a quadratic matrix  $A$  with matching dimensions:

$$\begin{aligned} \frac{\partial}{\partial X} \text{tr}(XA) &= A + A^\top - \text{diag}(A) \\ \frac{\partial}{\partial X} \log(|X|) &= 2X^{-1} - \text{diag}(X^{-1}). \end{aligned}$$

Using  $M_{ij} = (x_i - \mu_j)(x_i - \mu_j)^\top$  for readability and keeping in mind the symmetry of  $M_{ij}$ , taking the derivative yields:

$$\begin{aligned}
0 &= \sum_{i=1}^N h_{ij}^{(t)} (2\Sigma_j - \text{diag}(\Sigma_j)) - \sum_{i=1}^N h_{ij}^{(t)} (2M_{ij} - \text{diag}(M_{ij})) \\
0 &= \sum_{i=1}^N h_{ij}^{(t)} (2(\Sigma_j - M_{ij}) - \text{diag}(\Sigma_j - M_{ij})) \\
0 &= 2 \left( \sum_{i=1}^N h_{ij}^{(t)} (\Sigma_j - M_{ij}) \right) - \text{diag} \left( \sum_{i=1}^N h_{ij}^{(t)} (\Sigma_j - M_{ij}) \right).
\end{aligned}$$

From the last equation we can see that

$$\sum_{i=1}^N h_{ij}^{(t)} (\Sigma_j - M_{ij}) = 0,$$

and after re-substituting  $(x_i - \mu_j)(x_i - \mu_j)^\top$  for  $M_{ij}$  we arrive at

$$\begin{aligned}
\sum_{i=1}^N h_{ij}^{(t)} \Sigma_j &= \sum_{i=1}^N h_{ij}^{(t)} (x_i - \mu_j)(x_i - \mu_j)^\top \\
\Sigma_j^{(t+1)} := \Sigma_j &= \frac{\sum_{i=1}^N h_{ij}^{(t)} (x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_{i=1}^N h_{ij}^{(t)}},
\end{aligned}$$

which concludes the derivation of  $\Sigma_j^{(t+1)}$ .



---

## C Confusion matrices for Kappa calculation

---

Confusion matrices for three documents.  $M$  encompasses novel metaphor annotations,  $N$  the rest of the sentences. The matrices on the left show results for annotators A1 and A2, in the middle for annotators A1 and A3 and on the right for annotators A2 and A3.

A note on the differing total numbers for the Nietzsche document: One part of the text from annotator A1 is missing, so less sentences can be compared.

	N	M	
N	174	34	208
M	21	22	43
	195	56	251

	N	M	
N	163	44	207
M	20	23	43
	183	67	250

	N	M	
N	207	44	251
M	30	38	68
	237	82	319

Table C.1: Confusion matrices for Nietzsche: “Vom Nutzen und Nachteil der Historie ...”.

	N	M	
N	84	4	88
M	4	16	20
	88	20	108

	N	M	
N	85	3	88
M	5	15	20
	90	18	108

	N	M	
N	83	5	88
M	7	13	20
	90	18	108

Table C.2: Confusion matrices for Hegel: “Grundlinien der Philosophie des Rechts, Vorrede”.

	N	M	
N	218	5	223
M	4	2	6
	222	7	229

	N	M	
N	219	4	223
M	2	4	6
	221	8	229

	N	M	
N	217	5	222
M	4	3	7
	221	8	229

Table C.3: Confusion matrices for Hegel: “Grundlinien der Philosophie des Rechts, Einleitung”.

---

## D How to use the system

---

In this section we will briefly describe how to preprocess the gold standard documents, create the resources, run the evaluation pipeline, and cluster candidates for exporting to CSniper. It requires understanding of Apache UIMA<sup>1</sup>, eclipse<sup>2</sup>, Apache Maven<sup>3</sup> and related concepts. Note that in this “how to” we omitted the upper package qualifiers `de.wangtang.diplom` in some places for lack of space.

The code is located in `DVD://code/de.wangtang.diplom` on the supplied DVD. Import the project as a Maven project into eclipse.

First, customize the paths in `de.wangtang.diplom.Config` to suit your setup.

Then you have to create the database. There are two SQL files supplied, one containing the already built vectors, candidates and other resources (`DVD://data/sql/data.sql`), the other one only containing the structure and the manually looked up year of origin for each document (`DVD://data/sql/structure.sql`). If you chose `data.sql`, you can skip the next steps and continue with the evaluation step.

Otherwise, you have to move the data from the DVD to your specified folders. You also have to obtain Google Books Ngram Corpus files<sup>4</sup> and the deWaC corpus<sup>5</sup>. The UBY database has to be created<sup>6</sup> and after obtaining GermaNet files<sup>7</sup>, these have to be imported into the UBY database. You can see the variables in `de.wangtang.diplom.Config` and their corresponding folders on the DVD in Table D.1, for the data which is supplied.

To start the preprocessing, run `preparations.PreprocessDocuments`. This reads the DOCX files and annotates gold standard metaphors, tokenizes and tags the documents and conducts dependency parsing. It also creates the candidate annotations. At the end, the resulting CASes are written into the `Config.CAS_DIR` directory.

Now the resources have to be built. First build the Lucene index over deWaC by calling `preparations.DewacToLucene`. deWaC has to be parsed by running `preparations.ParseDewac`, which creates CASes in `Config.DEWAC_CAS`. The experiments were conducted using the first 60.000 parsed documents, so you might want to stop the parser before it finishes parsing all 1.75 million documents. Then you can build the selectional preference repositories by call-

---

<sup>1</sup> <http://uima.apache.org/>

<sup>2</sup> <http://eclipse.org/>

<sup>3</sup> <http://maven.apache.org/>

<sup>4</sup> <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>. You need the files under German, Version 20120701; the `total_counts` file and the 3-gram files for extracting novelty, the dependency files for selectional preferences

<sup>5</sup> <http://wacky.sslmit.unibo.it/doku.php?id=download>

<sup>6</sup> <http://code.google.com/p/uby/>

<sup>7</sup> <http://www.sfs.uni-tuebingen.de/lsd/licenses.shtml>

---

Config variable	folder on DVD containing the data
<code>DATA_DIR</code>	<code>DVD://data</code>
<code>DOCUMENTS_DIR</code>	<code>DVD://documents</code>
<code>CAS_DIR</code>	<code>DVD://data/cas</code>
<code>GOOGLE_BOOKS_NGRAMS</code>	download <sup>4</sup>
<code>DEWAC_DOCS</code>	download <sup>5</sup>
<code>DEWAC_CAS</code>	create
<code>DEWAC_LUCENE_LEMMA</code>	create

---

Table D.1: Configuration options and their corresponding folders on the supplied DVD.

ing `preparations.BuildVerbSelPrefRepo` and `preparations.BuildNounSelPrefRepo`. Using `preparations.CreateNoveltyTable`, extract the n-gram information needed for novelty calculation. At the end, build the feature vectors using `preparations.BuildVectors`.

To run the crossvalidation pipeline, use `lab.Evaluation`. Testing of a model also is situated there, uncomment the testing line and comment out the crossvalidation line. Reports are created in the directory specified by `Config.REPORTS_DIR`.

You can obtain the kappa scores by calling `lab.kappa.ComputeKappa`.

To build an example export file for importing into CSniper, run `lab.ClusterAndExportExample` after editing in the options you want to use.

The code for the new CSniper import page and new DatabaseEngine is included on the DVD in the folder `DVD://code/csniper-google`.

A patch file is supplied at `DVD://code/csniper-google/import_patch.txt`.

---

---

## List of Figures

---

2.1	Example extract showing a noun - genitive noun modifier construction . . . . .	14
2.2	Example extract showing a verb - direct object construction . . . . .	15
3.1	deWaC coverage of lemmas in dataset documents . . . . .	25
3.2	Excerpt from a Google Books Ngram Corpus dependency file (j) . . . . .	28
3.3	Excerpt from a Google Books Ngram Corpus dependency file (noun) . . . . .	30
4.1	Example extract showing two noun - genitive noun modifier constructions . . . . .	33
4.2	CSniper import page . . . . .	39
4.3	CSniper assessment page . . . . .	40

---

## List of Tables

---

2.1	Statistics for the used dataset. Met. stands for novel metaphor, Amb. for ambiguous.	13
3.1	GermaNet semantic fields . . . . .	29
4.1	Kappa scores for Nietzsche: “Vom Nutzen und Nachteil der Historie ...” . . . . .	34
4.2	Kappa scores for Hegel: “Grundlinien der Philosophie des Rechts, Einleitung” . . . . .	34
4.3	Kappa scores for Hegel: “Grundlinien der Philosophie des Rechts, Vorrede” . . . . .	34
4.4	Kappa scores for Helmholtz: “Über die Erhaltung der Kraft” . . . . .	34
4.5	Confusion matrices for the Helmholtz text. . . . .	35
4.6	Baseline results for noun - genitive noun modifier candidates . . . . .	36
4.7	Baseline results for verb - object candidates . . . . .	36
4.8	Results of the classifier with the best feature set for noun - genitive noun modifier candidates, but without novelty filter . . . . .	37
4.9	Results of the classifier with the best feature set for verb - object candidates . . . . .	37
4.10	Results of the classifier with the best feature set for noun - genitive noun modifier candidates, but with novelty filter . . . . .	38
C.1	Confusion matrices for Nietzsche: “Vom Nutzen und Nachteil der Historie ...” . . . . .	55
C.2	Confusion matrices for Hegel: “Grundlinien der Philosophie des Rechts, Vorrede”. . . . .	55
C.3	Confusion matrices for Hegel: “Grundlinien der Philosophie des Rechts, Einleitung”. . . . .	55
D.1	Configuration options and their corresponding folders on the supplied DVD. . . . .	57

---

## Bibliography

---

- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98, pages 86–90.
- Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.
- Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, ICSI Berkeley.
- Black, M. (1955). XII - Metaphor. In *Proceedings of the Aristotelian Society*, volume 55, pages 273–294.
- Black, M. (2011). More about metaphor. In Orthony, A., editor, *Metaphor and Thought*, pages 19–41.
- Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97, Beijing, China.
- Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*, pages 597–620.
- Cilibrasi, R. and Vitanyi, P. M. B. (2004). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):15.
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.
- Eckart de Castilho, R., Bartsch, S., and Gurevych, I. (2012). CSniper - Annotation-by-query for non-canonical constructions in large corpora. In *Proceedings of the 50th Meeting of the Association for Computational Linguistics (ACL) 2012 (Demo section)*, pages 85–90.
- Eckart de Castilho, R. and Gurevych, I. (2011). A lightweight framework for reproducible parameter sweeping in information retrieval. In *Proceedings of the 2011 workshop on Data infrastructures for supporting information retrieval evaluation, DESIRE '11*, pages 7–10, New York, NY, USA. ACM.

- 
- Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, pages 1–21.
- Grabowski, S. and Swacha, J. (2012). Google Books Ngrams Recompressed and Searchable. *Foundations of Computing and Decision Sciences*, 37(4).
- Gurevych, I. and ECKLE-KOHLER, J. (2012). Uby: A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics EACL 2012*, pages 580–590.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.
- Hamp, B. and Feldweg, H. (1997). Germanet - a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications (1997)*, pages 9–15.
- Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Lakoff, G. and Johnson, M. (2003). *Metaphors we live by*. University of Chicago Press Chicago, Ill. ; London.
- Li, L. and Sporleder, C. (2010). Using gaussian mixture models to detect figurative language in context. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 297–300.
- Lin, Y., Michel, J.-B., Aiden Lieberman, E., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174.
- Manning, C. and Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 26. MIT Press, Cambridge, MA.
- Mason, Z. J. (2004). CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System. *Computational Linguistics*, 30(1):23–44.
- Michel, J.-B. et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*.
- Miller, G. a. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

- 
- 
- Olson, M. A., Bostic, K., and Seltzer, M. (1999). Berkeley DB. In *Proceedings of the FREENIX Track: 1999 USENIX Annual Technical Conference*.
- Resnik, P. (1993). Selection and information: a class-based approach to lexical relationships. *IRCS Technical Reports Series*.
- Richards, I. (1936). *The philosophy of rhetoric*. Oxford UP, New York.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*.
- Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis.
- Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley, New York.
- Shutova, E., Sun, L., and Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10*, pages 1002–1010.
- Wilks, Y. (1978). Making Preferences More Active. In *Artificial Intelligence, vol. 11*, pages 197–223.