MEANING, REFERENTIALITY AND DISTRIBUTION: A COMPUTATIONAL INVESTIGATION OF MARKERS IN GERMAN COMPOUNDING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ALI EREN ÇETINTAŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

APRIL 2024

MEANING, REFERENTIALITY AND DISTRIBUTION: A COMPUTATIONAL INVESTIGATION OF MARKERS IN GERMAN COMPOUNDING

submitted by **ALI EREN ÇETINTAŞ** in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics** ________________

Assoc. Prof. Dr. Barbaros Yet
Head of Department, **Cognitive Science** ________________

Prof. Dr. Cem Bozşahin
Supervisor, **Cognitive Science, METU** ________________

Examining Committee Members:

Assoc. Prof. Dr. Barbaros Yet
Cognitive Science, METU ________________

Prof. Dr. Cem Bozşahin
Cognitive Science, METU ________________

Assist. Prof. Dr. Murat Ulubay
Management, Ankara Yıldırım Beyazıt University ________________

Date: 22.04.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ali Eren Çetintaş

Signature :

ABSTRACT

MEANING, REFERENTIALITY AND DISTRIBUTION: A COMPUTATIONAL INVESTIGATION OF MARKERS IN GERMAN COMPOUNDING

Çetintaş, Ali Eren

M.S., Department of Cognitive Science

Supervisor: Prof. Dr. Cem Bozşahin

April 2024, 53 pages

Compounding is one of the known ways of word formation and it is available in many languages to some degree or another with crosslinguistic variations. It is also a productive way of word formation in German (Neef, 2009). Compounding in German makes use of some markers, mostly called linking elements, between the constituents, and this phenomenon is highly common. Whether these markers have any meaning or what primary functions they have are seemingly highly controversial. In this study, we suggest that the close relation between meaning and reference on the one hand and categorization on the other can be explored computationally in distributional properties of these markers, which are difficult to identify analytically.

Keywords: compounding, linking elements, referentiality, distribution, word embeddings, syntax, semantics

ÖZ

ANLAM, GÖNDERIMSELLIK VE DAĞIMSALLIK: ALMANCA TAMLAMALARDAKI EKLERIN HESAPLAMALI BIR INCELEMESI

Çetintaş, Ali Eren
Yüksek Lisans, Bilişsel Bilimler Bölümü
Tez Yöneticisi: Prof. Dr. Cem Bozşahin

Nisan 2024, 53 sayfa

Tamlama, kelime oluşturmanın bilinen yollarından biridir ve birçok dilde diller arası farklılıklarla bir dereceye kadar mevcuttur. Bu aynı zamanda Almancada kelime oluşturmanın da verimli bir yoludur (Neef, 2009). Almancada tamlama, bileşenler arasında çoğunlukla bağlayıcı birimler olarak adlandırılan bazı eklerden yararlanır ve bu durum oldukça yaygındır. Bu birimlerin herhangi bir anlamı olup olmadığı ya da hangi temel işlevlere sahip oldukları görünüşte oldukça tartışmalıdır. Bu çalışmada, bir yanda anlam ve gönderimsellik, diğer yanda kategorizasyon arasındaki yakın ilişkinin, analitik olarak tanımlanması zor olan bu birimlerin dağılım özellikleri ile hesaplamalı olarak incelenebileceğini öneriyoruz.

Anahtar Kelimeler: tamlama, bağlayıcı ekler, gönderimsellik, dağılımsallık, kelime vektörleri, sözdizim, anlambilim

To my family

ACKNOWLEDGMENTS

First of all, I am grateful to all my family members and friends who have supported me up to this point, as well as my advisor Prof. Dr. Cem Bozşahin and other professors and teachers who helped me throughout this journey.

I would like to express my gratitude to committee members, Assoc. Prof. Dr. Barbaros Yet, Prof. Dr. Cem Bozşahin and Assist. Prof. Dr. Murat Ulubay for being understanding and providing valuable their feedback.

I would like to thank TÜBİTAK (The Scientific and Technological Research Council of Turkey) for their financial support of this thesis via the BİDEB 2210-A scholarship I received during my master's studies.

I also appreciate the feedback that I received from Anıl Öğdül, İbrahim Ethem Deveci and Rojda Özcan. Thanks to their feedback, I could express the ideas in a better way.

I thank Cihan Başoğlu and Mehmet İlteriş Bozkurt for giving some feedback on the title and abstract of the thesis.

I also want to thank Burcu Alakuş Çınar and Ecenur Çağırıcı Doğan for helping me with the format of thesis. Extra thanks to Ecenur Çağırıcı Doğan for having answered my questions and helped me since before I came to the department.

I am also grateful to staff at Informatics Institute for doing their best to help us.

TABLE OF CONTENTS

List of Tables

List of Figures

LIST OF ABBREVIATIONS

LE	Linking Element
PL	Plural/Plural Marker/Plural Suffix
MASK/mask	Masculine
FEM/fem	Feminine
NEUT/neut	Neuter
SG/sg	Singular
NOM/nom	Nominative
GEN/gen	Genitive
V	Verb
N	Noun
A	Adjective
PCA	Principle Component Analysis
N1	Modifier of compound
N2	Head of compound
AVG/Avg	Average
$\emptyset$	Zero Linking Element/No Linking Element

CHAPTER 1

INTRODUCTION

Compounding is one of the known ways of word formation, which follows different rules and leads to different results in different languages (see Lieber & Štekauer, 2009), and it is a productive way of word formation in German Neef (2009).

Schlücker (2019) sheds some lights upon the basics of German compounding. She says that compounds in German are formed with the combination of two stems, and are mostly right-headed. She conveys that the modifier has the stress on it and it “cannot be specified” (p. 78).

Schlücker (2019) further adds that while adjectival and nominal compounds are highly productive and common, verbal compounds are much rarer.

In German, while some compounds have linking elements such as -*s*- in *Kapitän-s-mütze* (Neef, 2009), some other compounds do not have any linking elements such as *Rotwild* and *blutrot* (Schlücker, 2019). Barz (as cited in Neef, 2009) states that around 30 % of the all compounds contain linking elements, while in the CELEX lexical data base, 35 % of the noun-noun compounds have linking elements (Krott et al., 2007). This indicates that the existence of linking elements in German cannot be simply overlooked. The linking elements in German include -*s*-, -*e*-, -*n*-, -*en*-, -*ens*-, -*es*-, and -*er*- based on Becker (as cited in Neef, 2009, p. 392) and based on Krott et al. (2007)). The combination of the modifier with some linking elements may also lead to a vowel chance in the modifier through umlaut, reduction of the some elements in the modifier (Neef, 2009; Krott et al., 2007) and a replacement of an element with the linking element in the modifier (Neef, 2009). The examples of the linking elements alongside with the possible changes in the modifier are presented as follows based on Becker (as cited in Neef, 2009, p. 392) as can be seen in Figure 1:

a	*e*	Tag+Buch	Tagebuch	‘diary’
b	*er*	Kind+Garten	Kindergarten	‘kindergarten’
c	*s*	Wirt+Haus	Wirtshaus	‘tavern’
d	*es*	Jahr+Zeit	Jahreszeit	‘season’
e	*n*	Affe+Haus	Affenhaus	‘ape house’
f	*en*	Präsident+Wahl	Präsidentenwahl	‘presidential election’
g	*ns*	Name+Tag	Namenstag	‘name day’
h	*ens*	Herz+Wunsch	Herzenswunsch	‘dearest wish’
i	umlaut	Mutter+Heim	Mütterheim	‘mother house’
j	umlaut+*e*	Gans+Braten	Gänsebraten	‘roast goose’
k	umlaut+*er*	Buch+Regal	Bücherregal	‘bookshelf’
l	subtraction of *e*	Sprache+Insel	Sprachinsel	‘linguistic enclave’
m	replacement of *e* by *s*	Geschichte+Buch	Geschichtsbuch	‘history book’
n	∅	Büro+Tür	Bürotür	‘office door’

Figure 1: List of Linking Elements

Out of the linking elements potrayed above, -*e*-, -*er*- and -*(e)n*- are homophonous with plural suffixes, which leads some people to believe that these linking elements come from plural suffixes (Wegener, 2008). Wegener (2008) rejects this idea, and claims they are distinctive distinguishable morphs, being in some cases linking element and in some others plural suffix in the compound. Following the tradition of categorial grammar going back to Ajdukiewicz (1935) that referential differences and meaning can be related to the distribution and cause the existence of distinctive categories for the homophonous linguistic items, these two categories can be identified. With the help of word embeddings, which are basically vectorial representations of the word meanings, we expect that this is possible from a computational perspective as well.

1.1 Research Question

The meaning and referential properties of compound constituents, especially those of head, impacts categorization of the marker in compound, and this enables us to identify the marker in a compound as a linking element or a plural suffix following the categories Wegener (2008) offers for these. When the interpretation that compound constituents have are taken into consideration, it is possible to identify the marker correctly, and the failure to take the meaning of compound constituents into consideration may lead to difficulties with this distinction.

According to Wegener (2008), there is a parallel development, namely “grammaticalization”, of some linking elements and plural suffixes from stem suffixes, which has led to the existence of compound examples whose structures are ambiguous because they might be the type of either stem + plural suffix or stem + linking elements as in the following examples (p. 342):

(1) a. *Kind-er-chor*
child-PL-choir

'children's choir'

b. *Kind-er-star*
child-LE-star

'child star'

(2) a. *Frau-en-rechte*
woman-PL-rights

'women's rights'

b. *Frau-en-stimme*
woman-LE-voice

'woman's voice'

(3) a. *Hund-e-meute*
dog-PL-pack

'pack of dogs'

b. *Hund-e-hütte*
dog-LE-house

'dog-house'

Wegener (2008) suggests that in (1a), (2a) and (3a), the modifiers have plural interpretation, and thus the markers are plural suffixes giving us the structure of stem + plural suffix structure while in (1b), (2b), (3b), the modifiers have singular interpretation, and thus the markers are linking elements giving us the structure of stem + linking element. Therefore, it is possible to say that the meaning of the compound constituents makes the identification of the marker's category possible.

In this study, the aim is to prove that by making use of this knowledge of meaning, generalizing it to novel compounds as well as offering our own interpretations, one can identify the marker as a linking element or a plural suffix correctly. As the semantics of compound constituents are also reflected in words embeddings, one can expect the identification of compound markers to be possible from a computational perspective with the help of word embeddings.

1.2 Organization of the Thesis

In the next chapter, some background information regarding German compounding and word embeddings will be provided. Then, related work will be touched upon after which our approach and the data sets will be introduced. Thereafter, the experiments and the results will be presented and these will be discussed. Then, we will have a conclusion.

CHAPTER 2

BACKGROUND

2.1 Properties of German Compounding

Compounding is one of the known ways of word-formation, which seems to be available in many languages to some degree or other, and there does not seem to be a universal rule to define compounding in general (see Lieber & Štekauer, 2009). Therefore, for a better understanding, it is more appropriate to look at the features of German compounding rather than only try to figure out general rules regarding compounding.

Before going into the details of German compounding, it is helpful to look at the very basics of German compounding. Neef (2009) emphasizes that compounding is considerably productive in German, and Schlücker (2019) sheds light on some general aspects of compounds. She states that compounds in German are formed with the combination of two stems or more. She says that while adjectival and nominal compounds are highly productive and common, verbal compounds are much rarer. She adds that the compounds in German are mostly right-headed, the modifier carries the stress on it and the modifier cannot be specified. For example, she states that **sehr Extremposition* (p. 78) is ungrammatical, while its syntactic counterpart *sehr extreme Position* (p. 78) is acceptable. However, it seems that the coordination of modifiers in compounds are possible as in *Kapitäns- und Admiralsmützen* (Fuhrhop, 1998, as cited in Neef, 2009, p. 390)

2.1.1 Grammatical Features of German Compounds

One of the essential features of German compounding is the right-headedness: the right constituent of the compound is the head (Neef, 2009; Schlücker, 2019) while the left constituent is modifier (Schlücker, 2019). Neef (2009) illuminates headedness in German compounding. He states that the head determines the grammatical features of the compound such as word-class, gender, plurality, and case.

As for how the word-class of the head impacts the compound, Neef (2009) states that German normally has three open word-classes, namely nouns, adjectives and verbs, and these classes also allow compounding. To start with, the following examples also illustrate how the noun head determines the word-class of the whole compound by making it also nominal (cf. Fleischer & Barz, 1992; Meibauer, 2001, as cited in Neef, 2009, p. 388):

(4)

a.	N → N+N	*Holz+haus* ‘wooden haus’	*Freiheits+kampf* ‘fight for freedom’
		Merkel+rede ‘Merkel-speech’	*Helfers+helfer* ‘accomplice’
b.	N → A+N	*Groß+segel* ‘mainsail’	*Dunkel+kammer* ‘dark-room’
c.	N → V+N	*Web+fehler* ‘weaving flaw’	*Radier+gummi* ‘rubber’

In these examples, it can be seen that nominal compounds can be derived through juxtaposition of two stems from word-classes such as verb, adjective and noun as long as the head is a noun. Neef (2009) also provides examples to illustrate the formation of adjectival and verbal compounds (pp. 388–389):

(5)

a.	A → N+A	*fleisch+farben* ‘flesh-coloured’	*herz+zerreißend* ‘heartbreaking’
b.	A → A+A	*alt+klug* ‘precocious’	*alt+eingeführt* ‘introduced long ago’
c.	A → V+A	*treff+sicher* ‘accurate’	*fahr+tüchtig* ‘able to drive’

(6)

a.	V → N+V	*rad+fahren* ‘to cycle’	*lob+preisen* ‘to praise’
b.	V → A+V	*trocken+legen* ‘to drain’	*froh+locken* ‘to rejoice’
c.	V → V+V	*kennen+lernen* ‘to get to know’	*schwing+schleifen* ‘to grind by swinging’

In (5) and (6) , one can also see that adjectival and verbal compounds can be formed through the juxtaposition of two stems from different word-classes, namely nouns, adjectives and verbs. Regardless of the word-class of the modifier, the word-class of the whole compound is that of the head. In other words, the examples in (5) and (6) as well as those in (4) demonstrate that the word-class of the head determines that of the whole compound. The situation of these verbal compounds seems to be controversial though because Schlücker (2019) claims that there are different analyses regarding N+V patterns.

Regarding the frequency and productivity of these compounds, adjectival and nominal compounds are highly productive and common, verbal compounds are much rarer (Neef, 2009; Schlücker, 2019).

In addition, Neef (2009) also informs us about how the head plays a role in gender, plurality and case of the whole compound and provides the following examples to demonstrate this (p. 389):

(7)

a. *gender*:

$(Dampf_{mask} + schiff_{neut})_{neut}$ 'steamship'
$(Schiff_{neut} + fahrt_{fem})_{fem}$ 'shipping'

b. *plural*:

$Dampf_{sg}$ –$Daempfe_{pl}$ 'steam'	$Schiff_{sg}$ –$Schiffe_{pl}$ 'ship'
$(Dampf + schiff)_{sg}$ –$(Dampf + schiffe)_{pl}$	$*(Daempfe + schiff)_{pl}$

c. *genitive*

$Haus_{nom}$ –$Hauses_{gen}$ 'house'	$Herr_{nom}$ –$Herren_{gen}$ 'master'
$(Haus{+}herr)_{nom}$ –$(Haus{+}herren)_{gen}$ 'host'	$*(Hauses + herr)_{gen}$

In (7a), one can see that the compounds headed by *Schiff* 'ship' and *Fahrt* 'ride' which are neutral and feminine nouns respectively inherit the gender of their heads. In (7b), it can be seen that the plurality of the compound is determined by whether the head, not the modifier, is in the plural or singular form: while the plural form of the head *Shiff* 'ship' makes the whole compound plural, the plural modifier *Daempfe* does impact the plurality of the whole compound. In (7c), one can see the same in the inheritance of case. The case of the head, not the modifier, determines the case of the whole compound: the nominative head *Herr* "master" makes the compound *Hausherr* 'host' nominative, and the genitive head *Herren* makes the whole compound genitive, but the genitive modifier *Hauses* does not make the whole compound genitive.

2.1.2 Stress in German Compounding

The place of stress in compounds is one of the defining properties of German compounds. While each word in a compound normally has one stress in isolation, their combination into a compound leads to the existence of one stress in the whole compound (Neef, 2009). The stress is on the modifier in a compound as in *Fríschluft* ‘fresh air’ in contrast to the syntactic counterpart *frische Lúft* ‘fresh air’, which has the stress on the head (Schlücker, 2019). Neef (2009), as for complex compounds, which have more than two stems, offers a stress rule: “In a compound, the second constituent is stressed if it is a compound by itself; otherwise, the first constituent is stressed” (p. 393). Neef (2009) claims that stress can distinguish the meaning and provides the examples *Lebens‘mittelpunkt* ‘center of life’ (p. 394) and *‘Lebensmittelpunkt* ‘marker on groceries’ (p. 394) where the former has a compound as its head and the latter has a compound as its modifier.

Neef (2009) goes on to inform us about the possible exceptions to this stress pattern and provides some examples regarding this. One significant exception among his examples is *‘Nordbahnhof* ‘North station’ (p. 394) as well as *‘Westbahnhof* ‘West station’ (p. 394), where the stress is on the modifier even though it is supposed to be on the head due to the head being a compound itself. He claims that this exception results from the contrastive function of the stress which distinguishes both compounds by the element being stressed.

2.1.3 Semantics of German Compounds

Another important thing about German compounding is its semantic and Neef (2009) informs us about it. He takes the compositionality as the base of semantics in German compounding, whereby he means that the meaning of the compound can be derived from the meaning of its elements and the way they are constructed into a compound. He sees headedness as a significant aspect in the structure of German compounds and adds that most compounds result in an interpretation, saying “grammatical and semantic head coincide” (p. 395). Neef (2009) also says that this enables us to interpret a compound AB as “B that has something to do with A” (p. 395), which is most often seen in N+N compounds and he calls this "determinative reading" (p. 395). Heringer (as cited in Neef, 2009) offers some possible interpretations for *Fish+frau* ‘fish+woman’, which is "a woman that has something to do with fish" (p. 395):

(8)

a. woman that sells fish

b. woman that has brought fish

c. woman standing close to fish

d. woman eating fish

e. woman looking like a fish

f. spouse of a fish

g. woman and fish at the same time (i.e. mermaid)

h. woman having Pisces as zodiac (German Fisch)

i. woman as cold as a fish

Neef (2009) elaborates on these possible interpretations and says that in such cases, the context helps one to infer the correct interpretation. He also adds that “a familiar compound” (p. 395) normally has an established meaning and that the case with *Fish-frau* ‘fish woman’ (Translation ours) having many interpretations is not commonly seen in contrast to *Haustür* ‘house door’ (Translation ours), which mostly refers to the door leading into an house.

In contrast to compositionality, non-compositionality can also be seen in German compounding due to the lexicalization, as in *Eselsbrücke* (p. 396), which normally consists of *Esel* ‘donkey’ (p. 396) and *Brücke* ‘bridge’ (p. 396), but actually means ‘mnemonic trick’ even though it is also possible to derive its meaning in a determinative way (Neef, 2009).

Neef (2009) mentions analogy as another significant tool for deriving the meaning of a compound and describes its role. He claims that the meaning of a compound can be derived based on another established compound with a known meaning. He implies that even though *Haus+mann* ‘house husband’ (p. 395) could have many possible interpretations regarding a man who is in some relation to a house, its meaning is mostly derived based on *Haus+frau* ‘housewife’ (p. 395), which shares the same modifier *Haus* (p. 395).

In addition, Neef (2009) also talks about copulative compounds, where the grammatical head and semantic head are different. He notes that these compounds are still right-headed, but the interpretation of such a compound AB is "A and B" (p. 396), which presents "a coordinative structure" (p. 396) between compound constituents. He also emphasizes that there are two classes of such compounds (p. 396):

(9)

a. *rot-grün* ‘red and green’

b. *schwarz-rot-gold* ‘black and red and golden’

c. *Dichter-Maler-Komponist* ‘poet and painter and composer’

(10)

a. *Hosen+rock* ‘pant+skirt’

b. *Kinder+pilot* ‘child pilot’

c. *Kino+café* ‘cinema and café’

d. *nass+kalt* ‘chilly and damp, lit. wet+cold’

While the examples in (9) are strictly copulative compounds (cf. Becker, 1992 as cited in Neef, 2009), Neef (2009) claims that the examples in (10) can be interpreted in a determinative reading as well as in a cordinative reading. On the other hand, what Neef (2009) calls copulative compounds are named "coordinate compounds" by Bisetto & Scalise (2009), who try to propose a new classification framework for compounds.

Schlücker (2018) informs us about the semantics of compounding further and talks about possible semantic relations between compound constituents. For some semantic relations between compound constituents that stand out among many other semantic relations, such as HAVE, LOC, ABOUT, MAKE and SIMILAR , she provides the following examples that include both common noun compounds and proper name compounds with proper names acting as modifiers (p. 287):

(11)

HAVE	*Regierungspläne* ‘government plans’
	Merkel-Pläne ‘Merkel plans’
LOC	*Waldwiese* ‘forest glade’
	Irak-Krieg ‘Iraq war’
ABOUT	*Tierschutzstudie* ‘study on animal welfare’
	Berlin-Studie ‘study on Berlin’
MAKE	*Milchkuh* ‘milk cow’
	Mozart-Sinfonie ‘Mozart symphony’
SIMILAR	*Baumdiagramm* ‘tree diagram’
	Einsteinhirn ‘Einstein brain’

Schlücker (2018) goes on to describe these semantic relations in these examples, where the first example is a common noun compound and the second one is a proper name compound in each relation. She says that some of these relations can be reversed such as HAVE(α, β) as in *government plans* or HAVE (β, α) as in *picture book* because in the former example, the modifier is the possessor while in the latter, the head is the possessor. The equivalent of these examples in German, where one could see the reversibility of HAVE relation would be *Regierungspläne* ‘government plans’ given above Schlücker (2018) and *Bilderbuch* ‘picture book’, where the possessor is realized by modifier and head respectively.

Schlücker (2018) also illuminates the SIMILAR relations and emphasizes that common noun compound and proper name compounds differ from each other regarding this relation. She notes that in common noun compounds, the similarity is between the concepts denoted by the constituents, while in proper name compounds, the similarity is between the concept denoted by the head and a referent of the same concept possessed by the entity denoted by the proper name modifier: *tree diagram* is a diagram that looks like a tree, but *Einsteingehirn* 'Einstein brain' is a brain that is similar to that of Einstein.

Besides the explanations by Schlücker (2018) regarding the semantic relations in (11), we can also add that the LOC relation is actually self-evident: it enables the modifier to denote a location regarding the concept or entity denoted by the head. In *Waldwiese* 'forest glade', the modifier *Wald* 'forest' specifies the kind of *Wiese* 'glade', and in *Irak-Krieg* 'Iraq war', the modifier denotes the location of *Krieg* 'war'. Similarly, the ABOUT relation is also self-evident: in *Tierschutzstudie* 'study on animal welfare' and *Berlin-Studie* 'study on Berlin', the modifiers basically denote the scope of *Studie* 'study'. However, the MAKE relation seems a little bit trickier. In *Milchkuh* 'milk cow', the modifier *Milch* 'milk' denotes a product and the head *Kuh* 'cow' denotes some kind of a maker or producer while in *Mozart-Sinfonie* 'Mozart symphony', the modifier *Mozart* 'Mozart' denotes the producer or maker and the head *Sinfonie* 'symphony' denotes a product. In addition to these examples, *Kuhmilch* 'cow's milk', where the modifier is *Kuh*, denotes some kind of maker or producer and the head *Milch* 'milk' denotes a product, is also attested in German. This shows us that as described for the SIMILAR relation by Schlücker (2018), common name compounds and proper name compounds can also differ from each other regarding the MAKE relation.

2.2 Word Embeddings

When one wants to input the representations of the words or texts, there is a need to transform the text data into some numeric data. This can normally be done in a few ways.

One of the traditional ways of numeric representation of raw text data could be one-hot encoding. If one wanted to represent a word with one-hot encoding, one would have a vector in the size of the whole vocabulary in the corpus and this vector would consist of all zeros and only the index that corresponds to the word in question would be one. Such sparse vectors would not be efficient in representing the meaning.

One reason for this inefficiency is the presence of many zeros in our vector. Normally, it is possible to have many words in our vocabulary (The size of the vocabulary is denoted as V.). Such a vector representing one word will have V-1 times 0 and only the index that corresponds to this word would be 1. That many zeros would increase the amount of calculations necessary with no significant contribution, which can be considered as highly unnecessary and costy.

Another issue with one-hot encoding is that the meaning may not be represented properly, which makes it impossible to get the semantic similarities between words correctly. Let's say there are words such as *cat*, *dog*, *tree*, *truck* in our vocabulary, whose vectors would be [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1] respectively. When one tries to take the cosine similarity between these words, one will always have the same cosine similarity, which is 0. For example, the similarity between cat and dog will be the same as the one between cat and truck, which is not the case.

To overcome such problems, it is better to make use of words embeddings, which are learned representations. They are based on distributional semantics. The main idea is that words with similar meanings will appear in similar contexts. More precisely, as Firth (1957) puts it: "You shall know a word by the company it keeps!" (p. 11). This should result in words with similar contexts having similar representations capturing their meanings. Accordingly, *dog* and *cat* should have similar contexts and thus similar vectors. In other words, the embeddings of words are learned with the consideration of neighbour words in their contexts.

Word embeddings can provide dense vectors of size 50, 100, 300 etc, where each value in the vector corresponds to a different feature/dimension of the word. Compared to one-hot encoding vectors, such vectors should be computationally less costy because with one hot encoding, one may get word vectors of high sizes owing to the number of vocabulary items.

In addition to being less costy, word embeddings also seem to be more informative. They represent semantic and syntactic regularities, and by simple arithmetic operations, it is possible to manipulate the meaning that the word vectors represent (Mikolov et al., 2013). A typical example provided by Mikolov et al. (2013) is the subtraction of "Man" from "King" and adding "Woman" to it and getting a word vector close to that of "Queen". The difference between word pairs like "Man" and "Woman", "King" and "Queen", "Aunt" and "Uncle", stemming from gender, is also more or less the same (Mikolov et al., 2013), which gives us some idea about how effective word embeddings are in representing meaning and differences between words numerically. Considering this, one can also expect to get some proper numerical representations of compound constituents as well as their referential and semantic differences.

CHAPTER 3

RELATED WORK

One of the significant studies that includes the categorization of homophonous markers in compounds is Wegener (2008) where the markers that are homophonous with plural suffixes like *-er-*, *-e-*, and *-(e)n-* -s-, are dealt with, and a marker is categorized as a linking element or a plural suffix depending on whether the modifier has a singular or plural interpretation. Wegener (2008) delves into the origins of markers used in compounds, illuminates their development into linking elements and plural suffixes and attempts to portray this as regrammaticalization.

As for the origins of linking elements and plural suffixes, Wegener (2008) provides us with some explanations. She says that one assumed origin of the linking element *-s-* is the genitive suffix *–(e)s-*, which became a genuine linking element after the lexicalization of prenominal genitive modifiers carrying this suffix. This marker is not the concern of our study though. Wegener (2008) also states that the assumed origin of the linking elements that are homophonous with plural suffixes such as *-er-*, *-e-*, *-(e)n-* and *-s-*, are plural suffixes themselves, which she thinks is actually not true and offers some reasons.

One reason Wegener (2008) provides against this is that there are some modifiers which carry these elements, but have never had plural meaning diachronically or synchronically such as *Kindergesicht* 'child's face' (p. 337), *Bilderrahmen* 'picture frame' (p. 337) and claims that if the plural suffixes had been the origin of these linking elements, the modifiers would bear some plural meaning at some point.

Another significant reasons that Wegener (2008) offers is that in Old High German, there already existed "compositionsvocal" (Grimm, 1978 as cited in Wegener, 2008, p. 338), which were placed between the constituents in compounds such as *tag-a-sterro* 'day star, morning star' (p. 338), *naht-i-gal* 'nightingale' (p. 338), and these composition vowels are the preceders of the today's linking elements and underwent some phonological changes in Middle High German as in *lemb-ir-bah* >*Lämm-er-bach* (p. 338), whereas "the homophonous plural markers developed later, in the Early New High German period" (p. 338), which means that the linking elements *-er-*, *-(e)n-* and *-e-* cannot have developed from the plural suffixes.

Wegener (2008) proposes that there is a common origin for both the linking elements and the homophonous plural suffixes: "the old Indo-European stem-building suffixes of the *ir-* class, the *n-* class and the *a-*, *o-* or *i*-class (fused to *-i*)" (p. 339), and there is a parallel development from the stem suffixes, which normally indicated the inflection class, to plural suffixes and linking elements over time, and this explains the formal identity between the two forms.

Wegener (2008) describes the development of linking elements from stem suffixes and starts by stating that one example of the stem suffixes, which normally followed the root, but preceded the inflectional suffix for case and number, was "Hühnerhofdeklination" (p. 339) with mostly "agricultural meanings" (p. 339) in Old High German, and the stem suffix of this class was, in Proto-Germanic, *-iz*, which occurred in all the forms of the paradigm. She explains that this suffix underwent some changes: it became *-ir* in Old High German, and with "the phonological reduction of the unstressed syllables" (p. 339) became *-er* in Middle High German. She adds that the suffix *-iz* disappeared at the end of the polysyllabic words as early as Pre-Old High German, and some forms in the paradigm (singular genitive and dative as well as all of the plural forms) had this suffix while some (singular nominative and accusative forms) did not such as singular genitive *hrind-ir-es* (p. 340) and singular nominative *hrind* (p. 340) respectively, as a result of which the suffix did not have any interpretation as a stem suffix or as a marker of a grammatical category and it became afunctional junk as described by Lass (1990, as cited in Wegener, 2008). Wegener (2008) goes on to say that since Old High German, there had been some compounds such as *lemb-ir-bah* that existed still with the stem suffix between the constituents, and such compounds preserved the stem suffix that was reanalysed as linking element, which then allowed the modifiers in the compounds to be a pure root or a stem with a stem suffix as in *Rind-fleisch* 'beef' (p. 340) and *Rind-er-braten* 'roast beef' (p. 340).

Wegener (2008) clarifies the development of plural suffixes from the stem suffixes, which occurred in parallel to the development of linking elements from stem suffixes, and starts by reminding us that after the disappearance of the suffix in some of the forms as early as the Pre-Old High German, the stem suffix existed in all of the plural forms and in two of the singular forms, which was later followed by the systemization of the paradigm with analogy to other classes. She explains that the stem suffix, with no more interpretation, is dropped completely in the singular forms, and because the stem suffix, then, remained only in the plural forms, it was reanalysed as a plural suffix.

According to Wegener (2008), the double development of the stem suffixes described above brought about "two formally identical but nevertheless distinguishable morphs" (p. 342), namely liking elements and plural suffixes, which explains why there are examples of compound modifiers whose structure is ambiguous because they might be the type of either "stem + plural suffix" (p. 342) or "stem + linking elements" (p. 342) as in follows (p. 342):

(12) a. *Kind-er-chor*
child-PL-choir

'children's choir'

b. *Kind-er-star*
child-LE-star

'child star'

(13) a. *Frau-en-rechte*
woman-PL-rights

'women's rights'

b. *Frau-en-stimme*
woman-LE-voice

'woman's voice'

(14) a. *Hund-e-meute*
dog-PL-pack

'pack of dogs'

b. *Hund-e-hütte*
dog-LE-house

'dog-house'

Wegener (2008) suggests that in (12a), (13a) and (14a), the modifiers have plural interpretation, and thus the markers are plural suffixes giving us the structure of stem + plural suffix structure while in (12b), (13b), (14b), the modifiers have singular interpretation, and thus the markers are linking elements giving us the structure of stem + linking element, In other words, homophonous markers are categorized differently depending on the interpretation of modifiers, which indicates that the categories of these markers have something to do with distribution and the categorization is impacted by the meaning and the referential properties of compound constituents.

Wegener (2008) also provides the following examples (p. 337), which could further improve our understanding of the meaning and referential properties of the constituents and their relation to distribution, which can be glossed as follows:

(15) a. *Kind-er-gesicht*
child-LE-face

'child's face'

b. *Bild-er-rahmen*
picture-LE-frame

'picture frame'

(15) c. *Geist-er-fahrer*
ghost-LE-driver

'wrong way driver'

d. *Kleid-er-bügel*
cloth-LE-hanger

'coat hanger'

Wegener (2008) states that the modifiers in (15) have never had plural interpretation and implies that they have singular interpretations, and thus the markers are linking elements. We can say that exploring more about such compounds and their constituents can allow us to categorize these distinctive but homophonous markers more efficiently and offer insights into the use of these markers. Wegener (2008) also notes that the plurality effect in modifier can also be achieved without such markers as in

Buchhändler ‘book seller’ (p. 336). This means that even though the plurality in modifier has something to do with these markers, it is not restricted to them.

In addition to Wegener (2008), some other also touch upon the plural vs. singular distinction in modifiers in German compounds even though they do not agree with Wegener (2008) in some issues. Nübling & Szczepaniak (2013) provide the following examples (pp. 80–81) where this distinction can be seen, and which we can gloss following Wegener (2008) as follows:

(16) a. *Kind-er-auge*
child-LE-eye
‘child eye’

b. *Männ-er-kopf*
man-LE-head
‘man’s head’

(17) a. *Ärzt-e-kongress*
doctor-PL-congress
‘doctors’ congress’

b. *Städt-e-planer*
city/town-PL-planner
‘town planner’

Some more examples from Nübling & Szczepaniak (2008) where the same distinction can be seen are as follows after we gloss them following Wegener (2008):

(18) a. *Kind-er-arzt*
child-PL-doctor
‘pediatrician’

b. *Kind-er-wagen*
child-LE-carriage
‘buggy’

Schäfer & Pankratz (2018b) also studied the relation between these markers homophonous to plural suffixes, which we are concerned with, which they call “pluralic linking elements” (p. 325), and their possible link to plurality in modifier, which they call “N1” (p. 325) while they call the head “N2” (p. 325). Their concern was whether there was a tendency for these markers to occur in a compound when there was plural interpretation in modifier triggered by one of the two effects, namely “external plural effect” (p. 335) and “internal plural effect” (p. 335) as they call them:

- External plural effect: “A plural on the entire compound (formally on the head constituent) might trigger the use of a pluralic linking element” (p. 335).
- Internal plural effect: “Certain semantic classes of N2s standing in an appropriate semantic relation with N1 might force N1 to have a plural interpretation and therefore lead to a preference for using the pluralic linking element” (p. 335).

Schäfer & Pankratz (2018b) provide two possible underlying reasons for external plural effect. One is that the plurality of the whole compound might lead modifier to refer to a set with more than one entity. To better illustrate this, Schäfer & Pankratz (2018b) state the following:

> For example, we might see a preference for a compound *Hund+herz* 'dog's heart' with the non-pluralic linking element (zero in this case) in the singular but *Hund-e+herzen* 'dogs' hearts' with the pluralic linking element (*-e* in this case) in the plural. (p. 335)

The other possible underlying reason for external plural effect offered by Schäfer & Pankratz (2018b) is "a purely formal one; there might simply be plural agreement within the compound" (p. 336). Schäfer & Pankratz (2018b) say "Banga et al. (2013a) and Banga et al. (2013b) have found that there are effects related to the mere presence of a formal plural of N1 in the context of the compound" (p. 336). On the other hand, Schäfer & Pankratz (2018b) also emphasise that "the plural on the whole compound" (p. 336) might lead modifier to take a pluralic linking element even when there is no plural semantic that could require it, and this creates a possibility that the marker might not be related to plurality even if there is evidence for external plural effect.

As for internal plural effect, Schäfer & Pankratz (2018b) attribute it to "the lexical meanings of the compound's constituents" (p. 336), and state that heads lead modifiers to have some plural meaning and that such heads are mostly those with collective meanings. Possible examples by Schäfer & Pankratz (2018b) are "true collectives like *Kindergruppe* 'group of children', metaphorical collectives as in *Zitateregen* 'rain of quotations', reciprocals such as *Räderwechsel* 'swapping of tyres', or relational N2s as in *Lochdistanz* 'distance between (the) holes'" (p. 336).

Schäfer & Pankratz (2018b) approached the issue from a probabilistic perspective such that the choice for pluralic linking element based on the plural semantic are not binary "binary but probabilistic, weighted, and better described as numerical rather than discrete" (p. 335).

Schäfer & Pankratz (2018b) investigated these two effects in a corpus study and "in an experimental paradigm (split-100 ratings)" (p. 336). For the corpus of the corpus study, they chose "the web-crawled DECOW16A corpus (Schäfer and Bildhauer 2012; Schäfer and Bildhauer, in preparation)" (p. 336) to extract the sample compounds from.

For the corpus study, Schäfer & Pankratz (2018b) annotated compound samples "for whether they are plurals" (p. 341) and whether plural meaning is evident in modifier as a result of "a plural-enforcing semantic relation" (p. 341) between compound constituents that provoke plural meaning in it. When annotating samples, Schäfer & Pankratz (2018b) figured out some features for the above-mentioned semantic relation that creates plural meaning:

- Collectives: "*Gruppe* 'group' as in *Kind-er+gruppe* or *Kind+gruppe* 'group of children'" (p. 343).

- Metaphorical collectives: "*Regen* 'rain' in compounds like *Zitat-e+regen* or *Zitat+regen* 'rain of quotations'" (p. 343).

- Some other words: "*Distanz* 'distance' as in *Löch=er+distanz* or *Loch+distanz* 'distance'" (p. 343), where the meaning depended on the context.

Schäfer & Pankratz (2018b) also had some cases where the picture was blurry with respect to the same semantic relation as it heavily required "context and world knowledge" (p. 344). To clarify this, Schäfer & Pankratz (2018b) provide some examples and explain:

> Examples include *Äpfel=+lager* or *Apfel+lager* 'storage for apples', *Brief-e+katalog* or *Brief+katalog* 'catalogue of letters', and *Lied-er+buch* or *Lied+buch* 'book of songs/songbook'. In theory, the N2s in these compounds could denote some sort of container which holds only one object (for instance, it is conceivable—if unlikely—that the storage space for apples could have only one apple in it), but both world knowledge and the particular context in which the compound appears render this sort of interpretation impossible. (p. 343)

As for the results of the corpus study conducted by Schäfer & Pankratz (2018b), there seems to be no significant evidence for the external plural effect while this is not the case for internal plural effect.

In the split-100 experiment conducted by Schäfer & Pankratz (2018b), the purpose was to test the tendency of native speakers of German to choose pluralic linking elements in two parts studying internal and external plural effects. The participants were provided with two options, a compound containing a pluralic linking element and one that did not, and they expressed their preference in a range from 0 to 100 (Schäfer & Pankratz, 2018b). As for the options to be offered to the participants, Schäfer & Pankratz (2018b) attempted to "find for each N1 some semantically appropriate N2s which clearly trigger internal plural semantics and ones which clearly do not" (p. 348).

The results of the split-100 experiment were similar to those of corpus study (Schäfer & Pankratz, 2018b). Schäfer & Pankratz (2018b) found no significant result for external plural effect as the participants showed tendencies to choose compounds without pluralic linking elements while the participants showed tendencies to choose compounds with pluralic linking elements when there was a semantic motivation for plurality in modifier due to the relation between compound constituents.

The overall results of study by Schäfer & Pankratz (2018b) show us that it is possible to see some connection between plural semantics and the homophonous markers in compounds. Note that Schäfer & Pankratz (2018b) do not exclude the possibility to have pluralic linking elements to be present in compounds without any relation related to plurality between the constituents and the internal plural effect, as Schäfer & Pankratz (2018b) name it, can also be seen without the use of these markers, which Wegener (2008) also points out. Considering all these indicates that investigating more into this question could be promising and show us some more results.

One study where the attempt at the classification of meaning with the help of word embeddings can be found, is Krotova et al. (2020), who were concerned with the detection of idiomatic German compounds before their splitting task so that idiomatic compounds could be detected and remain unsplit. They used dword2vec and fastext trained on Wikipedia of size 300.

Krotova et al. (2020) annotated samples based on GermaNet, and this data set had non-idiomatic compounds and idiomatic compounds varying in their idiomaticity. They also used two splitters: "either the source gold standard split or our own splitter, based on Char-GRU" (p. 4414) as they state it. Regarding the feature representation, they simply concatenated the word embeddings of the compound and the compound constituents having 900-dimensional vectors for each sample.

Krotova et al. (2020) used two separate classification models for this task: Logistic Regression from scikit-learn and Gradient Boosting (XGBoost). For their dummy model, they used a model that detected every compound as idiomatic compound, gave and F-1 score of 0.21.

As for the results of idiomatic detection, Krotova et al. (2020) had F-1 scores of ranging 0.54 to 0.58, which they considered an improvement compared to their dummy model. Logistic regression and gradient boosting had almost the same performance with gradient boosting performing slightly better (Krotova et al., 2020).

Considering what Krotova et al. (2020) did, it is possible to say that machine learning models can be utilized with word embeddings in order to classify meaning, and such ideas could set paves for the computational aspect of our study as well.

CHAPTER 4

DISCUSSION AND DATA SETS

4.1 Theoretical Approach and Early Discussion

It seems that the meaning and the referential differences of the compound constituents, especially that of head, can cause a distributional difference, and this distributional difference enables us to identify the markers *-er-*, *-(e)n-* and *-e-* as a linking element or as a plural suffix in the compound following the categorization of Wegener (2008). By studying the examples provided in the literature such as those of Wegener (2008), Nübling & Szczepaniak (2008), Nübling & Szczepaniak (2013), we can learn more about the referential properties of the compound constituents, and generalize this knowledge to other compounds, which will, in turn, allow us to see the distributional difference and identify the marker in a novel compound correctly.

When we consider the heads in (15a), (16a) and (16b) where the modifiers have singular interpretation, and thus the markers can be said to linking elements, it is clear that the heads, namely *Gesicht* 'face', *Auge* 'eye' and *Kopf* 'head', are body parts and we can say that body parts might yield reference to the singular in modifier which, in turn, allows us to identify the marker as linking element. As for (14b), (15b), (15d), it might be possible to say that the heads' reference to the singular in the modifiers may stem from the nature of the objects that the heads in these compounds denote: a 'dog-house' will house a single dog rather than a group of dogs, a 'frame' should contain a single picture, and a 'hanger' should have a single piece of cloth hanged onto it, which leads us to take the markers in these compounds as the linking element.

By drawing some analogies upon the interpretations of modifiers in such examples, we can detect the distributional difference resulting from the difference in meaning and referential properties of the constituents, and determine whether the marker in a novel compound is a linking element or a plural suffix in novel compounds. In other words, we can generalize our knowledge regarding the meaning the referential properties of the constituents and get some new interpretations for novel compounds. For example, considering the reference to the singular in body parts as heads in (15a), (16a), (16b), we can also say that the modifiers in *Frau-en-gesicht* 'woman's face', *Kind-er-kopf* 'child's head' and *Kind-er-haut* "child's skin' should also have singular interpretations too. By generalizing the referential properties in (15a), we can get a singular interpretation for the modifier of the former novel compound: if *Gesicht* 'face' yields reference to the modifier *Kind* 'child', it should also do the same for *Frau* ‘woman’. As for the latter example, we go one step further in our generalization and see that *Haut* ‘skin’, which does not exist in the previous examples, yields reference to the singular in the modifier as it is a body part. These interpretations prove the markers to be linking elements. Similarly, following the referential properties that we attribute to (14b), (15b), (15d), we can say that in a novel example such as *Kind-er-rucksack* ‘child’s backpack’, the head should yield reference to the singular in the modifier because of the nature of ‘backpack’: a ‘backpack’ concerns and tends to be used by a single child rather than a group of children, which brings us to the point that the marker is a linking element.

Only -*er*-, -*e*-, and -*(e)n*- are our concerns. We, for example, did not make distinction between -*(e)n*- and -*en*- as Nübling & Szczepaniak (2013) suggest.

We also did not make distinction between some plural suffixes and their equivalents which also induce an umlaut in the stem of the modifier unlike Schäfer & Pankratz (2018b). For example, the plural suffixes in *Kinder* ‘children’ and *Häuser* were both considered as -*er*-. This was also the case for -*e*- as well which was treated separately in the same way by Schäfer & Pankratz (2018b).

-*s*- or -*es*- are not our concerns either because they could be related to genitive case as well (Wegener, 2008).

We also did not discriminate against -*en*- in any way even though they could be related to genitive case (Wegener, 2008) or they could be some rules that forces its existence according to Nübling & Szczepaniak (2013).

4.2 Data Sets

4.2.1 Nominal Compounds in GermaNet

One of the data sets that are utilized in this study is the nominal compounds in GermaNet discussed in Henrich & Hinrichs (2011) while some other major works on GermaNet are Henrich & Hinrichs (2010) and Hamp & Feldweg (1997). Currently, GermaNet v18.0 (2023) contains 121655 nominal compounds and the compounds are provided with their modifiers and heads split.

As for the format of this data set, there is a compound, its modifier and its head in each line (Krotova et al., 2020).

4.2.2 Schäfer and Pankratz Data Set

Schäfer & Pankratz (2018a) is a data set of 9415 compounds where compound heads are annotated regarding what meaning heads provoke in modifier giving us four classes at hand: one class is individual-denoting while the others are not. In other words, we interpret the other three classes to be related to plurality in modifier or the plurality-relation between compound constituents.

The relevant columns that we mainly focused on were as follows:

- DocID: An ID assigned to the sample
- LC: The context preceding the given compound sample
- Match: The compound that is extracted
- LC: The context following the given compound sample
- N2Num: Whether the compound or the compound head is in plural form or not
- N2Typ: Whether the head is individual-denoting or not
- N1Lemma: The modifier itself
- N1Pl: The plural suffix of the modifier
- LE: Whether the compound contains a 'pluralic linking element' as Schäfer & Pankratz (2018b) name it

4.2.3 Trained Word2Vec - German Wikipedia

As for the word embeddings, we will be using Word2Vec word embdeddings trained on German Wikipedia and provided by deepset (2018). There are Word2Vec embeddings of 854776 words each one of them present in lowercase with an embedding of unknown tokens available.

4.2.4 DeReWo – Corpus-Based Lemma and Word Form Lists

This data set is a list of German word forms and their lemmas with their frequency scores based on the Mannheim German Reference Corpus DeReKo (Institut für Deutsche Sprache, 2014). It has 100000 samples. This will be used to get the lemmas of inflected linguistic items in Schäfer & Pankratz (2018a) where spaCy's German transformer model (Honnibal et al., 2020) is not enough.

4.3 Attempt at Creating Data Set for Classification and Annotation Process

To prepare the data, we started to annotate compounds that are from nominal compounds in GermaNet (Henrich & Hinrichs, 2011) by taking into account whether there is a plural or singular interpretation in modifier mostly resulting from the meaning of the head as discussed above. In order to do this, we also extracted linking elements in compounds. We basically got the string length (the number of characters) of the modifier and the heads, cut the compound from its start and its end giving us the remaining characters in the middle.

During the annotation, it can be seen that the kind of semantic relation between the relations can impact the interpretation of the modifier. In *Brett-er-wand* 'wooden/-timber/board wall', there is a MAKE relation as Schlücker (2018) names it while in *Bühne-n-wand* 'stage wall', there is a LOC relation as Schlücker (2018), again, names it. In the former example, the modifier *Brett* 'board/timber' has a plural interpretation, but in the latter, the modifier has a singular interpretation even though both compounds share the same head. This shows us that the semantic relation between the constituents as well as the lexical meaning of the head can impact the interpretation that the modifier has.

In addition, the direction of the semantic relation might be also important in determining whether the modifier has a plural interpretation or not. Both in *Bild-er-wand* 'wall with pictures/wall full of pictures' and in *Bühne-n-wand* 'stage wall', where the modifier is the same, there is a LOC relation as Schlücker (2018), again, defines it. However, in the former example where the modifier has a plural interpretation, the head denotes the location of the modifier while in the latter where the modifier has a singular interpretation, it is the other way around: the modifier denotes the location of the modifier. This might indicate the need to consider the modifier as well and try to capture such relations.

4.4 Further Data Cleaning and Preparation

Because annotating samples this way was taking too long, Schäfer & Pankratz (2018a) was a data set available to be used for our purpose, we turned to this data set to enlarge our data set faster.

Because the heads are not provided by Schäfer & Pankratz (2018a), we had to get the heads somehow. First, we utilized the German compound splitter provided by repodiac (2020). repodiac's German compound splitter is based on "the Aho–Corasick algorithm, pyahocorasick, for multi-pattern string search and retrieval" and works with a German dictionary (repodiac, 2020). It is also possible to get the singular forms by setting make_singular parameter to True in splitting (repodiac, 2020), which could normally be relevant for this study as it allows us to get rid of the markers homophonous to plural suffixes during splitting. In practice, this splitter returns all the available compound constituents at once in the order they appear in the compound rather than splitting the compound recursively and giving only the immediate constituents at each step as discussed in Henrich & Hinrichs (2011).

For an external German dictionary, we used the Free German Dictionary provided by Schreiber (2021), which has one word per line. We also manually added some words to this dictionary considering the failures in some splitting trials.

To split all the compound samples in Schäfer & Pankratz (2018a), we iterated through the data set and input the samples one by one into repodiac's German compound splitter. We set make_singular parameter to False as this made it easier to check the result.

To make sure that the given compound was split correctly, we considered a number of factors:

1. Whether the splitter returned two elements.
2. Whether the proper combination of modifier given in Schäfer & Pankratz (2018a) and the second constituent returned by the repodiac (2020) matches the compound itself as it is in Schäfer & Pankratz (2018a).

If these two conditions did hold, we considered the split to be a successful split and extracted the head, which corresponded to the last constituent returned. The compounds which were successfully split were labeled 1 for their Split_Type to indicate that the split was a result of repodiac's German compound splitter.

We attempted to get all the compounds during whose splitting some problems were encountered regarding four criteria so that necessary ones among those compounds could be handled separately in the later steps. These criteria can be stated as follows:

1. The compound is split into two wrong constituents.
2. The compound is not split at all.
3. The compound is split into more than one constituent.
4. The splitter returns an error because of some problem regarding dictionary.

After extracting the compounds that were problematic in splitting, we turned to some other methods to split these compounds, and we especially focused on the ones whose heads were annotated as non-individual denoting and aimed to extract all of those heads because they were lower in number in the data set.

The other significant method was to split the compound based on the string lengths of the modifier and the marker. We basically got the length of the modifier and the length of the marker if there was any. We cut the start of the string based on this/these length(s), which gave us the rest of the compound. Upon uppercasing the first character of this, we got what we believed to be a noun. We split around 300 compounds this way. Compounds split successfully this way were labeled 2 for their Split_Type.

Apart from these, we manually split some 9 compounds, which would be problematic in splitting because some of them were not split correctly by repodiac's German compound splitter, and the others had umlauted vowels replaced with their equivalents. This creates some mismatch between the compound itself and the modifier provided with an umlauted vowel, which leads to problems with the splitting based on string lengths of modifiers and linking element. These were labeled 3 for their Split_Type.

After being completely done with splitting the compounds whose heads were basically non-individual-denoting, we further attempted to split the remaining unsplit compounds which share some of these non-individual-denoting heads, but whose heads were labeled individual-denoting. In other words, all the compounds with non-individual-denoting heads were split while only some of the compounds with individual-denoting heads are split, and out of the remaining unsplit compounds with individual-denoting heads, we aimed to split those of them whose heads were also labeled non-individual-denoting in other compounds. The reason behind this is that when in later stages, extracting compound samples from GermaNet based on the heads of compound samples in Schäfer & Pankratz (2018a), it might be necessary to exclude those to avoid ambiguity. In order to achieve this, we just iterated through all the remaining unsplit compound samples whose heads were labeled individual-denoting, and checked whether the compounds contained any of the previously extracted non-individual-denoting heads. That is we checked whether the compound strings ended with any of the previously extracted heads which were labeled non-individual-denoting. If yes, we updated the head accordingly. These compound were labeled 4 for their Split_Type.

Then, we moved on to lemmatize the heads that were extracted in splitting. To this end, we used the spaCy's German transformer model (Honnibal et al., 2020). We lemmatized all the heads whether they were in plural or singular. While it is self-evidently clear that we need lemmatize the heads that are in plural form to get their stems, it was also better to lemmatize heads in singular in order to get their stem out of their forms some of which, for example, seemed to be genitive. We did not lemmatize the compounds that were manually split. These samples labeled 1 for their Lemmatization

In order to make sure that the lemmatization of heads went smoothly to a great degree, we iterated through all the samples and used DeReWo – Corpus-Based Lemma and Word Form Lists (Institut für Deutsche Sprache, 2014) to get their lemmas properly. We only used the lemmas that were nouns, foreign words and cardinal numbers in order to avoid getting undesired lemmas such as verbs etc. in DeReWo. We updated the lemmas returned by spaCy if it has a different lemma in DeReWo. These were labeled 2 for their Lemmatization. We ignored any possible changes that might have occurred such as *ss* becoming *ß*.

We also got around 50 samples that DeReWo had more than one options. Out of these, we attempted to manually lemmatize the heads that were labeled non-individual-denoting and the heads that were labeled individual-denoting, but labeled individual-denoting in other compounds. We considered whether these compounds/heads were labeled plural, what determiner they had in their contexts, whether they existed in GermaNet, and in which form they did exist in GermaNet if they ever did. These samples were labeled 3 for Lemmatization. As for the remaining ones, we simply removed their lemmas in order to avoid ambiguity in later stages.

Then, we also checked whether any heads started with lowercase meaning they were not lemmatized to nouns. We again focused only on those that were either labeled non-individual-denoting or that were labeled individual-denoting, but labeled individual-denoting in other compounds. We fixed these lemmas manually in the same way. These samples were also labeled 3 for Lemmatization. As for the remaining ones, we attempted to remove their lemmas in order to avoid ambiguity in later stages if there remained any.

Then, we moved on to check whether our annotated data set and Schäfer & Pankratz (2018a) in its modified form had any common compounds. We first got the lemmas of the samples in Schäfer & Pankratz (2018a) whose heads were labeled plural meaning the whole compound is in plural form, which carried any 'pluralic linking element' as Schäfer & Pankratz (2018b) name it. We simply made the given modifier plural and concatenated it with their lemmatized heads if they had any. Then, we went on to compare the compounds in both data sets and got the common compound that were in both data sets. Out of these common compounds which did not have any lemmatized heads, we assigned the heads in our annotated data set to them. These were labeled 4 for Lemmatization.

CHAPTER 5

EXPERIMENTS AND RESULTS

5.1 Approach

As discussed above, the classification of the markers is possible thanks to the meaning of compound constituents, and this classification should also be possible computationally with the help of word embeddings because there are attempts at meaning classification as Krotova et al. (2020) did with idiomatic compound detection using word embeddings.

For our classification task, we used scikit-learn logistic regression classifiers (Pedregosa et al., 2011) as Krotova et al. (2020) used them in idiomatic compound detection and almost got the same results as they did with gradient boosting. As for the hyperparameters, we set the number of iteration to 100000 and the rest were the default ones.

As for the features we used word2vec embeddings trained on German Wikipedia as provided by deepset (2018). When it came to the use of multiple compound constituents such as the use of modifier and the head at the same time, we used the concatenation of word embeddings of these separate linguistic items which is basically the approach taken by Krotova et al. (2020) in idiomatic compound detection.

We also reduced the number of dimensions or features in data and trained alternative models with them. To this end, we benefited from Principal Component Analysis (PCA), which be used to to reduce the number of dimensions or features in data set (Jolliffe, 2002). The main reason for reducing the number of dimensions or features is that we did not have so many samples in our data sets throughout the experiments considering the number of features. For the implementation of PCA, we mainly followed steps portrayed in Kavlakoglu (2024). We also trained some models with the scaled features, for which we also benefited from scikit-learn Pedregosa et al. (2011).

As for the naming of linguistic categories, we followed Wegener (2008) and any marker related with plurality in modifier was considered as plural suffix (PL) and any marker that was not related to plurality was considered as linking element (LE).

5.2 Experiments and Results

5.2.1 Experiment I - Working with Word Embeddings of Heads

In this part, we only used the unique heads in Schäfer & Pankratz (2018a) that were only labeled individual-denoting or non-individual-denoting and appeared with the the markers in question. We had every head only once in the data set in order to avoid any bias in the training. Because we had lower number of compound samples with non-individual-denoting heads, we tried to get only a limited number of heads that were individual denoting. After getting the available word embeddings of these, and balancing the classes in data set by removing random samples from dominant class. We had a data set of 446 samples.

We moved on to preprocess these embeddings. We implemented standard scaling and reduced dimension to 100 using PCA. We used 70% of the data for training and the remaining for testing.

We both trained separate models using the word embeddings as they were, their scaled equivalents and their reduced versions with Principal Component Analysis (PCA).

As for the results, they are given in the following tables (1, 2, 3):

Table 1: Experiment I Part I: With Original Word Embeddings

Class	Precision	Recall	F1-score	Support
Non-individual-denoting heads	0.62	0.69	0.65	67
Individual-denoting heads	0.65	0.58	0.61	67
Accuracy			0.63	134
Macro avg	0.64	0.63	0.63	134
Weighted avg	0.64	0.63	0.63	134

Table 2: Experiment I Part I: With Scaled Word Embeddings

Class	Precision	Recall	F1-score	Support
Non-individual-denoting heads	0.61	0.67	0.64	67
Individual-denoting heads	0.63	0.57	0.60	67
Accuracy			0.62	134
Macro avg	0.62	0.62	0.62	134
Weighted avg	0.62	0.62	0.62	134

Table 3: Experiment I Part I: With Word Embeddings of Size 100

Class	**Precision**	**Recall**	**F1-score**	**Support**
Non-individual-denoting heads	0.65	0.66	0.65	67
Individual-denoting heads	0.65	0.64	0.65	67
Accuracy			0.65	134
Macro avg	0.65	0.65	0.65	134
Weighted avg	0.65	0.65	0.65	134

As can be seen, we have an overall accuracy scores ranging from 0.62 to 0.65, which is not highly satisfactory. One explanation for this result is the number of samples we have: our data is not large enough. Another factor is the samples that we picked for individual-denoting heads that were, in number, greater than non-individual-denoting heads. When we picked different samples by shuffling, it was possible to get different results.

When we shuffled the whole data set again and do everything else again in the same way, it was possible to get different results as shown below (4, 5, 6):

Table 4: Experiment I Part II: With Original Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Non-individual-denoting heads	0.87	0.63	0.73	75
Individual-denoting heads	0.65	0.88	0.75	59
Accuracy			0.74	134
Macro avg	0.76	0.75	0.74	134
Weighted avg	0.77	0.74	0.74	134

Table 5: Experiment I Part II: With Scaled Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Non-individual-denoting heads	0.79	0.69	0.74	75
Individual-denoting heads	0.66	0.76	0.71	59
Accuracy			0.72	134
Macro avg	0.72	0.73	0.72	134
Weighted avg	0.73	0.72	0.72	134

Table 6: Experiment I Part II: With Word Embeddings of Size 100

Class	**Precision**	**Recall**	**F1-score**	**Support**
Non-individual-denoting heads	0.77	0.68	0.72	75
Individual-denoting heads	0.65	0.75	0.69	59
Accuracy			0.71	134
Macro avg	0.71	0.71	0.71	134
Weighted avg	0.72	0.71	0.71	134

With somewhat different samples for one class that we achieved by shuffling the data a few times, we changed the results drastically having overall accuracy scores ranging from 0.71 to 0.74. Especially, we got better results with original word embeddings, namely an increase of 0.9. Considering the low number of samples, it is possible to conclude that there is room for improvement and that it might be a good idea to enlarge the data set. In the next experiment, we mixed these data set with some head samples from our own annotated data set.

5.2.2 Experiment II - Working with Word Embeddings of Heads

In our annotated data set, we had only 454 samples. We got the unique heads of these samples, removed the contrasting ones among them. Then, we removed the samples that were in Schäfer & Pankratz (2018a) but did not belong to the same class. In other words, in case of contrasting samples, we assigned priority to the annotation in Schäfer & Pankratz (2018a). We also removed the compounds common to both data sets so that in addition to increasing the number of samples, the impact of samples unique to our own annotated data could be seen as well. We got the available word embeddings of these and removed the common samples. When we balanced the number of samples in both classes by removing samples from the dominant class and tried to mix these data set with the previous share of Schäfer & Pankratz (2018a), which yielded previous results, we had 596 samples.

After following the same steps as in Experiment I, we achieved some lower scores, as shown below tables (7, 8, 9):

Table 7: Experiment II Part I: With Original Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Non-individual-denoting heads	0.67	0.67	0.67	92
Individual-denoting heads	0.65	0.64	0.65	87
Accuracy			0.66	179
Macro avg	0.66	0.66	0.66	179
Weighted avg	0.66	0.66	0.66	179

Table 8: Experiment II Part I: With Scaled Word Embeddings

Class	Precision	Recall	F1-score	Support
Non-individual-denoting heads	0.61	0.54	0.57	92
Individual-denoting heads	0.57	0.63	0.60	87
Accuracy			0.59	179
Macro avg	0.59	0.59	0.59	179
Weighted avg	0.59	0.59	0.59	179

Table 9: Experiment II Part I: With Word Embeddings of Size 100

Class	Precision	Recall	F1-score	Support
Non-individual-denoting heads	0.70	0.70	0.70	92
Individual-denoting heads	0.68	0.69	0.69	87
Accuracy			0.69	179
Macro avg	0.69	0.69	0.69	179
Weighted avg	0.69	0.69	0.69	179

The overall accuracy scores range from 0.59 to 0.69, which could not be considered as an improvement compared to the results in Experiment I. One reason behind this lower scores despite the enlargement of the data could be the lack of annotator/inter-annotator agreement between two data sets. We annotated the compound samples from GermaNet mainly based on the samples provided by Wegener (2008), Nübling & Szczepaniak (2008), Nübling & Szczepaniak (2013) and somewhat by Neef (2009) without much consideration of context. On the other hand, samples in Schäfer & Pankratz (2018a) were annotated in their contexts. In order to see whether the decreased performance stems from annotator/inter-annotator agreement, we could follow every steps only using our data.

When we, to this end, got the word embeddings of only the samples in our annotated data set, we had 356 samples, out of which 201 were non-individual-denoting heads while 155 were individual-denoting heads.

We attempted to get an equal number of samples for both classes, which gave us 310 samples. We also included contrasting samples with the one in Schäfer & Pankratz (2018a). After a few trials (meaning shuffling the data giving different samples for the class that was greater in number), we got better results as shown below (10, 11, 12):

Table 10: Experiment II Part II: With Original Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Non-individual-denoting heads	0.71	0.83	0.76	47
Individual-denoting heads	0.79	0.65	0.71	46
Accuracy			0.74	93
Macro avg	0.75	0.74	0.74	93
Weighted avg	0.75	0.74	0.74	93

Table 11: Experiment II Part II: With Scaled Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Non-individual-denoting heads	0.78	0.74	0.76	47
Individual-denoting heads	0.75	0.78	0.77	46
Accuracy			0.76	93
Macro avg	0.76	0.76	0.76	93
Weighted avg	0.76	0.76	0.76	93

Table 12: Experiment II Part II: With Word Embeddings of Size 100

Class	**Precision**	**Recall**	**F1-score**	**Support**
Non-individual-denoting heads	0.74	0.72	0.73	47
Individual-denoting heads	0.72	0.74	0.73	46
Accuracy			0.73	93
Macro avg	0.73	0.73	0.73	93
Weighted avg	0.73	0.73	0.73	93

The overall accuracy scores range from 0.64 to 0.76. Even though the samples were less in number compared to the data set of heads from Schäfer & Pankratz (2018a) or that of mixed data (the combined data of our own annotated data and a part Schäfer & Pankratz (2018a) which yielded better results), our annotated data set excelled. However, Just like in the Experiment I, there could possibly also be cases where one could get better or worse accuracy scores depending on the shuffling of the data in balancing. Coming back to the idea that led to this experiment which is that the lack of annotator/inter-annotator agreement could be the reason why we got worse accuracy results in the case of merged data set, one could say that this idea had some truth.

So far, we were only concerned with the heads and got some results. From this point on, we tried to take modifiers into consideration as well. This could enable us to capture the relation between the compound constituents. It would also enlarge the data set as we would not have to stick to unique heads, but they could be available in many compounds with different modifiers.

5.2.3 Experiment III - Working with Word Embeddings of Modifiers and Heads

We first started with the samples in Schäfer & Pankratz (2018a). We first lemmatized the singular compounds using spaCy's German transformer model (Honnibal et al., 2020). It was to get rid of some genitive markers etc. in singular forms. In order to avoid undesired lemmatization, we checked whether the head that we had previously extracted matched the end of compound. If not, the compound was lemmatized. This was important in order to avoid getting duplicate samples that only differed in inflection.

We got rid of the duplicate samples that might have resulted from lemmatization of compounds or getting their base forms earlier. After getting the available word embeddings for modifiers and heads, and removing the duplicate compound samples among these, we had 2749 samples. After shuffling the data and removing the redundant samples which have individual-denoting-heads and linking elements to balance the number of samples in both classes, we had a total of 1456 samples. This time we reduced the dimensions to 200 instead of 100. The initial results can be seen in 13, 14, 15:

Table 13: Experiment III Part I: With Original Word Embeddings

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.74	0.73	0.73	214
Compounds with linking elements	0.74	0.75	0.75	223
Accuracy			0.74	437
Macro avg	0.74	0.74	0.74	437
Weighted avg	0.74	0.74	0.74	437

Table 14: Experiment III Part I: With Scaled Word Embeddings

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.76	0.78	0.77	214
Compounds with linking elements	0.78	0.76	0.77	223
Accuracy			0.77	437
Macro avg	0.77	0.77	0.77	437
Weighted avg	0.77	0.77	0.77	437

Table 15: Experiment III Part I: With Word Embeddings of Size 200

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.73	0.75	0.74	214
Compounds with linking elements	0.75	0.74	0.75	223
Accuracy			0.74	437
Macro avg	0.74	0.74	0.74	437
Weighted avg	0.74	0.74	0.74	437

Our accuracy scores now ranged from 0.74 to 0.77. These are greater than the previous results, which again shows that it is possible to get lower or higher scores with concatenated embeddings similar to what we saw in Experiment I.

As for the results in general, these could be considered as an improvement compared to the results that we got only using heads in Schäfer & Pankratz (2018a), especially considering that the dimension size is 600 now.

However, trying to get different results by having different samples in the data as result of shuffling before balancing the classes in the data set was still possible as it did Experiment I while it did not seem to create the same amount of effect as seen in 16, 17, 18:

Table 16: Experiment III Part II: With Original Word Embeddings

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.79	0.75	0.77	210
Compounds with linking elements	0.78	0.81	0.79	227
Accuracy			0.78	437
Macro avg	0.78	0.78	0.78	437
Weighted avg	0.78	0.78	0.78	437

Table 17: Experiment III Part II: With Scaled Word Embeddings

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.78	0.82	0.80	210
Compounds with linking elements	0.82	0.78	0.80	227
Accuracy			0.80	437
Macro avg	0.80	0.80	0.80	437
Weighted avg	0.80	0.80	0.80	437

Table 18: Experiment III Part II: With Word Embeddings of Size 200

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.78	0.78	0.78	210
Compounds with linking elements	0.80	0.80	0.80	227
Accuracy			0.79	437
Macro avg	0.79	0.79	0.79	437
Weighted avg	0.79	0.79	0.79	437

We now had overall accuracy scores ranging from 0.78 to 0.80.

In order to see the impact of addition of our annotated data, we trained another model with some samples from Schäfer & Pankratz (2018a) after removing the contrasting samples in it.

5.2.4 Experiment IV - Working with Word Embeddings of Modifiers and Heads Together

After removing the contrasting samples from our annotated data (just like in Experiment II, in case of contrasting samples, we gave priority the samples in Schäfer & Pankratz (2018a)) and adding the samples that do not have any equivalents in Schäfer & Pankratz (2018a) to the part of Schäfer & Pankratz (2018a) which gave the results in 16, 17, 18, we had 1748 samples in total. After following every step in Experiment III, we got the following results seen in 19, 20, 21:

Table 19: Experiment IV Part I: With Original Word Embeddings

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.77	0.71	0.74	253
Compounds with linking elements	0.75	0.80	0.77	272
Accuracy			0.76	525
Macro avg	0.76	0.76	0.76	525
Weighted avg	0.76	0.76	0.76	525

Table 20: Experiment IV Part I: With Scaled Word Embeddings

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.75	0.74	0.74	253
Compounds with linking elements	0.76	0.78	0.77	272
Accuracy			0.76	525
Macro avg	0.76	0.76	0.76	525
Weighted avg	0.76	0.76	0.76	525

Table 21: Experiment IV Part I: With Word Embeddings of Size 200

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.75	0.74	0.74	253
Compounds with linking elements	0.76	0.76	0.76	272
Accuracy			0.75	525
Macro avg	0.75	0.75	0.75	525
Weighted avg	0.75	0.75	0.75	525

The accuracy scores, ranging from 0.75 to 0.76, shows again that the addition of our annotated data did not improve the results seen in 16 17, 18 similar to the condition in Experiment II, but only worsened them. The reason still could be the lack of annotator/inter-annotator agreement. To see if this was possible, we trained models following the every same step only with our own annotated data set. After getting the available embeddings and removing duplicate compounds, we had 352 samples in total, which gave us these results in 22, 23, 24:

Table 22: Experiment IV Part II: With Original Word Embeddings

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.84	0.79	0.81	61
Compounds with linking elements	0.73	0.80	0.77	45
Accuracy			0.79	106
Macro avg	0.79	0.79	0.79	106
Weighted avg	0.80	0.79	0.79	106

Table 23: Experiment IV Part II: With Scaled Word Embeddings

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.87	0.87	0.87	61
Compounds with linking elements	0.82	0.82	0.82	45
Accuracy			0.85	106
Macro avg	0.85	0.85	0.85	106
Weighted avg	0.85	0.85	0.85	106

Table 24: Experiment IV Part II: With Word Embeddings of Size 200

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.87	0.85	0.86	61
Compounds with linking elements	0.80	0.82	0.81	45
Accuracy			0.84	106
Macro avg	0.84	0.84	0.84	106
Weighted avg	0.84	0.84	0.84	106

Our accuracy scores ranging from 0.79 to 0.85 were now better even though the number of samples were lower. In addition to these, just like in Experiment III, it was also possible to get different results by shuffling the data again and having different samples in data set in balancing as can be seen in the following tables seen 25, 26, 27:

Table 25: Experiment IV Part III: With Original Word Embeddings

Class	Precision	Recall	F1-score	Support
Compounds with plural suffixes	0.92	0.80	0.85	55
Compounds with linking elements	0.81	0.92	0.86	51
Accuracy			0.86	106
Macro avg	0.86	0.86	0.86	106
Weighted avg	0.87	0.86	0.86	106

Table 26: Experiment IV Part III: With Scaled Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.86	0.91	0.88	55
Compounds with linking elements	0.90	0.84	0.87	51
Accuracy			0.88	106
Macro avg	0.88	0.88	0.88	106
Weighted avg	0.88	0.88	0.88	106

Table 27: Experiment IV Part III: With Word Embeddings of Size 200

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.86	0.91	0.88	55
Compounds with linking elements	0.90	0.84	0.87	51
Accuracy			0.88	106
Macro avg	0.88	0.88	0.88	106
Weighted avg	0.88	0.88	0.88	106

Considering what we saw in Experiment III and IV, we can say that using the concatenation of word embeddings, we could work with more samples rather than restricting us to some heads that were lower in number. One contribution of this could be the higher number of samples in the data. One could say that the increased number of unique heads that we theoretically believe to be more significant in creating the interpretation in modifier could be the reason to improve the performance. However, we also had better results using the samples in our own annotated data, which even outperformed the model giving results in 16, 17, 18 even though the number of samples were greatly lower which would prevent the increased number of unique heads among the samples.

Another significant takeaway could be the significance of annotator/inter-annotator agreement. Having some reasonable result only with the part of Schäfer & Pankratz (2018a) or only with our annotated data did not ensure a good performance with the combination of these data. In contrast, it also yielded worse results especially considering the good performance of our own annotated data set on its own.

Before we move on to extract more samples from GermaNet using head samples from Schäfer & Pankratz (2018a), we will follow the same steps in Experiment III and Experiment IV to work with word embeddings of compounds as well as modifiers and heads.

5.2.5 Experiment V - Working with Word Embeddings of Compounds, Modifiers and Heads Together

In this part, we concatenated embeddings of compounds, modifiers and heads. We started with the part of Schäfer & Pankratz (2018a) which we worked on, giving us vectors of size 900. After getting the available embeddings, which decreased significantly due to the productive nature of compounding in our opinion, and balancing the number of samples in both classes, we had 226 samples in total. We trained the same models following the same steps, but only decreased the size of dimension to 120 with PCA and got the following results in 28, 29, 30:

Table 28: Experiment V Part I: With Original Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.68	0.87	0.76	30
Compounds with linking elements	0.87	0.68	0.76	38
Accuracy			0.76	68
Macro avg	0.78	0.78	0.76	68
Weighted avg	0.79	0.76	0.76	68

Table 29: Experiment V Part I: With Scaled Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.69	0.80	0.74	30
Compounds with linking elements	0.82	0.71	0.76	38
Accuracy			0.75	68
Macro avg	0.75	0.76	0.75	68
Weighted avg	0.76	0.75	0.75	68

Table 30: Experiment V Part I: With Word Embeddings of Size 120

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.66	0.77	0.71	30
Compounds with linking elements	0.79	0.68	0.73	38
Accuracy			0.72	68
Macro avg	0.72	0.73	0.72	68
Weighted avg	0.73	0.72	0.72	68

We had accuracy scores ranging from 0.72 to 0.76. Even though the results were not better than those in 13, 14, 15, where we used the embeddings of only modifiers and heads and had a greater amount of samples, they were still comparable. These can still be considered a success because it seems that concatenating compound embeddings might have compensated the loss of many samples in the data set.

From this point on, we assumed that it was always possible to get different results at each shuffle in balancing the data and did not attempt to get any alternative result to these.

When we used the mixed data set again which now gave us 366 samples after balancing, we got the results in 31, 32, 33:

Table 31: Experiment V Part II: With Original Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.75	0.79	0.77	53
Compounds with linking elements	0.80	0.75	0.77	57
Accuracy			0.77	110
Macro avg	0.77	0.77	0.77	110
Weighted avg	0.77	0.77	0.77	110

Table 32: Experiment V Part II: With Scaled Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.68	0.72	0.70	53
Compounds with linking elements	0.72	0.68	0.70	57
Accuracy			0.70	110
Macro avg	0.70	0.70	0.70	110
Weighted avg	0.70	0.70	0.70	110

Table 33: Experiment V Part II: With Word Embeddings of Size 120

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.66	0.79	0.72	53
Compounds with linking elements	0.76	0.61	0.68	57
Accuracy			0.70	110
Macro avg	0.71	0.70	0.70	110
Weighted avg	0.71	0.70	0.70	110

We still had overall accuracy scores ranging from 0.70 to 0.77, which again did not improve the previous results achieved with a part of Schäfer & Pankratz (2018a).

Then, we moved on to train the same models using only our annotated data. After following the same steps before and balancing the classes, we had 180 samples in total and got the results in 34, 35, 16:

Table 34: Experiment V Part III: With Original Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.81	0.85	0.83	26
Compounds with linking elements	0.85	0.82	0.84	28
Accuracy			0.83	54
Macro avg	0.83	0.83	0.83	54
Weighted avg	0.83	0.83	0.83	54

Table 35: Experiment V Part III: With Scaled Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.81	0.81	0.81	26
Compounds with linking elements	0.82	0.82	0.82	28
Accuracy			0.81	54
Macro avg	0.81	0.81	0.81	54
Weighted avg	0.81	0.81	0.81	54

Table 36: Experiment V Part III: With Word Embeddings of Size 120

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.81	0.81	0.81	26
Compounds with linking elements	0.82	0.82	0.82	28
Accuracy			0.81	54
Macro avg	0.81	0.81	0.81	54
Weighted avg	0.81	0.81	0.81	54

The overall accuracy scores ranging from 0.80 to 0.82 outperforming the models trained with mixed data showed us that we got reasonable results only with our annotated data, which could still be attributed to the significance of annotator/inter-annotator agreement.

In the next two parts, we trained models with the data set that were created with samples extracted from GermaNet based on heads unique to one class in Schäfer & Pankratz (2018a).

5.2.6 Experiment VI - Working with Word Embeddings of Modifiers and Heads Together

We used the heads unique to one class in Schäfer & Pankratz (2018a) and extracted more compound samples from compounds in GermaNet to increase the number of samples data. After getting the available word embeddings of modifiers and heads and removing duplicate ones, we had 2602 samples in total. By following the same steps as in Experiment III and IV where the word embeddings of only modifiers and heads were used, we got the results in as in 37, 38, 39.

Table 37: Experiment VI: With Original Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.84	0.82	0.83	391
Compounds with linking elements	0.83	0.85	0.84	390
Accuracy			0.83	781
Macro avg	0.84	0.83	0.83	781
Weighted avg	0.84	0.83	0.83	781

Table 38: Experiment VI: With Scaled Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.82	0.83	0.83	391
Compounds with linking elements	0.83	0.81	0.82	390
Accuracy			0.82	781
Macro avg	0.82	0.82	0.82	781
Weighted avg	0.82	0.82	0.82	781

Table 39: Experiment VI: With Word Embeddings of Size 200

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.81	0.82	0.82	391
Compounds with linking elements	0.82	0.81	0.82	390
Accuracy			0.82	781
Macro avg	0.82	0.82	0.82	781
Weighted avg	0.82	0.82	0.82	781

We had accuracy scores ranging from 0.82 to 0.83, which can be considered satisfactory. One obvious reason behind this could be the higher number of samples in general. Having the same unique heads in many samples could also give some more weights to certain features.

5.2.7 Experiment VII - Working with Word Embeddings of Compounds, Modifiers and Heads Together

In this part, we used the concatenation of word embeddings of compounds, modifiers and heads. After getting the available word embeddings of modifiers and heads and removing duplicate ones and balancing the classes, we had 770 samples in total. Following the steps in Experiment V and Experiment VI and training the models we got the results in 40, 41, 42:

Table 40: Experiment VII: With Original Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.89	0.77	0.82	128
Compounds with linking elements	0.75	0.88	0.81	103
Accuracy			0.82	231
Macro avg	0.82	0.82	0.82	231
Weighted avg	0.83	0.82	0.82	231

Table 41: Experiment VII: With Scaled Word Embeddings

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.90	0.80	0.85	128
Compounds with linking elements	0.78	0.88	0.83	103
Accuracy			0.84	231
Macro avg	0.84	0.84	0.84	231
Weighted avg	0.85	0.84	0.84	231

Table 42: Experiment VII: With Word Embeddings of Size 120

Class	**Precision**	**Recall**	**F1-score**	**Support**
Compounds with plural suffixes	0.90	0.77	0.83	128
Compounds with linking elements	0.76	0.89	0.82	103
Accuracy			0.83	231
Macro avg	0.83	0.83	0.83	231
Weighted avg	0.84	0.83	0.83	231

We had accuracy scores ranging from 0.82 to 0.83, which are slightly better than those in Experiment VI and thus can be considered satisfactory considering the highly lower amount of samples (almost a third of it).

We could say that losing some more samples for the sake of adding the embeddings of compounds are actually compensated of its informativeness.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion and Implications of the Study

Overall, as Ajdukiewicz (1935) points out, the categories of linguistic items is determined by its meaning and the variations in the meanings. In the case of markers in German compounding which are homophonous to plural suffixes, it is possible to say that the categorization of these markers can be achieved by the consideration of compound constituents, especially the head, in its environments. When we have plural interpretation in modifier due to the lexical meaning of the head or the semantic relation between the constituents, we can categorize the markers a plural suffix (PL) while the singular interpretation emerging in the modifier out of the same reasons give us the category linking element (LE) for the marker. As we could make this classification with word embeddings as well, it is possible to say that this categorical distinction can also be seen in distribution considering the results of the experiments.

Due to this relation between the categorization and the meaning of the linguistic items in their environments, this categorization can also be achieved with word embeddings which are numerical representation of meaning of text data based on their distribution.

One main thing to consider is: there are clear cases with respect to the some compound heads which can create plural interpretation in modifier such as *Bündel* 'cluster' (Translation ours), *Ansammlung* 'accumulation' (Translation ours) which exist both in our annotated data set and in Schäfer & Pankratz (2018a), which makes the classification of markers possible. On the other hand, there are also contrasting examples between these two data sets. In other words, there are heads that can create both singular and plural interpretation in modifier such as *Mannschaft* 'team' (Translation ours) creating plural interpretation in modifier *Frauenmannschaft* 'women's team' (Translation ours), but singular interpretation in *Nationmannschaft* 'national team' (Translation ours) as given in Schäfer & Pankratz (2018a). The existence of such contrasting samples might indicate the insufficiency of only-head-based classification and the need for the meaning of the modifier as well.

Another finding of this study is that it enables us to question the situation of the modifier in German compounds in terms of referentiality. Normally, modifiers in German compounds are considered non-referential unless they are not proper nouns that are not lexicalized (Schlücker, 2018). However, with the possibility to have plural interpretation in modifier, we might also have some referentiality in it because if we are dealing with linguistic items having plurality rather than or in addition to denoting simply concepts despite this possibility, it means that something is being referenced. For example, in *Kindergruppe* 'children's group', where there is plural interpretation triggered by the collective meaning of head Schäfer & Pankratz (2018b), we can sense some referentiality due to this plurality even though this might not be as intense as the referentiality of the head, and there could still be concept-denotation in modifier.

One thing to keep in mind is that while it is possible to achieve some increased performance with the concatenation of the word embeddings of the linguistic items, some might think that this performance might be somewhat related to the increased number of the heads in different samples as well, and this might have been instrumental in creating some bias giving some more weights to the features relevant for the embedding of the head rather than capturing the semantic relation between the constituents. However, this seems to be refuted by the fact that it was still possible to get better results using the concatenation of compounds, modifiers and heads despite the lower number of samples due to the unavailability of word embeddings for many compounds.

Another finding of this study is the importance of inter-annotator/annotator agreement. As seen in Experiment II, enlarging the part of Schäfer & Pankratz (2018a)that previously yielded some results by adding some samples from our own annotated data set based on GermaNet (Henrich & Hinrichs, 2011) to it did not improve the model even though adding samples from our annotated data set increased the number of samples in total, and it performed well on its own with a lower number of samples. The reason behind this could be that there was no annotator agreement between us and Schäfer & Pankratz (2018b) or Schäfer & Pankratz (2018a) and data sets were prepared for different purposes in different ways. One difference include that Schäfer & Pankratz (2018b) approach this question from a probabilistic approach while our approach was not like this. Our approach was much binary classifying the compound or head in a much more strict way and putting it in one class or the other. Another difference is that during the annotation, we only used the general knowledge of compound and compound constituents while this might not be the case for Schäfer & Pankratz (2018b) because there is also context provided for samples in Schäfer & Pankratz (2018a) and Schäfer & Pankratz (2018b) also talk about context as well with respect to this.

The findings of this study hold for the written language as we used word embeddings trained on Wikipedia. If we attempt to generalize this to spoken language, we might not get the same results.

This study also might also have some implications for the field of Computational Linguistics/NLP. German compound splitting is significant for NLP tasks such as machine translation because of the lack of equivalents of some compounds in the target language (Krotova et al., 2020). Commonly, German compounds are split in preprocessing before training and translation (Stymne, 2008 as cited in Krotova et al., 2020). With compounds splitting, it is also possible to translate compounds that do not exist in the target language by translating the constituents separately (Weller et al., 2014). To this end, some models like ours presented here can be integrated in machine translation models in larger scales to help to determine how to translate split constituents. If the model, for example, identifies the marker as a plural suffix, this can be taken into consideration, and the split constituent can be translated in plural form.

6.2 Future Work

One interesting thing to consider for future work is the increased performance with the concatenation of the word embeddings of the linguistic items discussed above. Because this performance might stem from the increased number of the heads in different samples which might rather than capturing the semantic relations between the constituents, it could be convenient to do something about this in future studies.

To remove such a possibility from our minds in future studies, one thing to do could be to check whether weights are mathematically affected by this.

Another thing to do could be to see how the concatenation of embeddings of separate linguistic constituents fare in capturing the semantic relation between them. To this end, we can annotate some compounds with respect to semantic relations between constituents and train a classification model in the same way.

Another alternative to concatenation of word embeddings could be the arithmetic operations. Instead of concatenating the word embeddings of separate linguistic items, it might also be a good idea to add the vectors and see how it fares.

It might also be a good idea to stick to one data set and use it completely because we had tendency to have lower performance when we added samples from our own annotated data set based on GermaNet (Henrich & Hinrichs, 2011).

In order to increase performance, we might also try out different machine learning models. So far, we used logistic regression following the steps taken in idiomatic compound detection in Krotova et al. (2020), who also used gradient boosting. We might also try out different classification models to see how the performance is affected.

Because the findings of this study is only relevant for written language due to the use of word embeddings trained on Wikipedia, we might try to see if the same holds for the spoken language using word embeddings trained on spoken corpus.

REFERENCES

Ajdukiewicz, K. (1935). Die syntaktische konnexitat. *Studia Philosophica*, *1*, 1–27. (English translation in S. McCall (ed): *Polish Logic*, pp. 207–231, Oxford University Press, 1967)

Bisetto, A., & Scalise, S. (2009). The Classification of Compounds. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Compounding* (p. 34-53). Oxford: Oxford University Press.

deepset. (2018). *German Word Embeddings.* `https://www.deepset.ai/german-word-embeddings`.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in linguistic analysis* (pp. 1–32). Oxford: Basil Blackwell. (Reprinted in 1962)

Hamp, B., & Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In *Automatic information extraction and building of lexical semantic resources for NLP applications.* Retrieved from `https://aclanthology.org/W97-0802`

Henrich, V., & Hinrichs, E. (2010, May). GernEdiT - the GermaNet editing tool. In N. Calzolari et al. (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA). Retrieved from `http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf`

Henrich, V., & Hinrichs, E. (2011, September). Determining immediate constituents of compounds in GermaNet. In *Proceedings of the international conference recent advances in natural language processing 2011* (pp. 420–426). Hissar, Bulgaria: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/R11-1058`

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. doi: 10.5281/zenodo.1212303

Institut für Deutsche Sprache. (2014). *Korpusbasierte wortformenliste DeReWo, DeReKo-2014-II-MainArchive-STT.100000.* Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland. Retrieved from `http://www.ids-mannheim.de/derewo`

Jolliffe, I. T. (2002). *Principal component analysis.* New York: Springer-Verlag.

Kavlakoglu, E. (2024, March 6). *Reducing dimensionality with principal component analysis.* `https://developer.ibm.com/tutorials/awb-reducing`

-dimensionality-with-principal-component-analysis/. (Accessed: March 2024 on IBM Developer)

Krotova, I., Aksenov, S., & Artemova, E. (2020, May). A joint approach to compound splitting and idiomatic compound detection. In N. Calzolari et al. (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 4410–4417). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.543

Krott, A., Schreuder, R., Harald Baayen, R., & Dressler, W. U. (2007). Analogical effects on linking elements in German compound words. *Language and cognitive processes*, *22*(1), 25–57.

Lieber, R., & Štekauer, P. (2009). Introduction: Status and Definition of Compounding. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Compounding* (p. 3-18). Oxford: Oxford University Press.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746–751).

Neef, M. (2009). IE, Germanic: German. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Compounding* (p. 386-399). Oxford: Oxford University Press.

Nübling, D., & Szczepaniak, R. (2008). On the way from morphology to phonology: German linking elements and the role of the phonological word. *Morphology*, *18*, 1–25.

Nübling, D., & Szczepaniak, R. (2013). Linking elements in German origin, change, functionalization. *Morphology*, *23*, 67–89.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

repodiac. (2020). *german_compound_splitter.* (Copyright 2020 by repodiac. Available at https://github.com/repodiac for updates and further information.)

Schäfer, R., & Pankratz, E. (2018a, July). *Dataset: The plural interpretability of German linking elements ("Morphology").* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.1323211 doi: 10.5281/zenodo.1323211

Schäfer, R., & Pankratz, E. (2018b). The plural interpretability of German linking elements. *Morphology*, *28*(4), 325–358. doi: 10.1007/s11525-018-9331-5

Schlücker, B. (2018). Genitives and proper name compounds in German. In T. Ackermann, H. J. Simon, & C. Zimmer (Eds.), *Germanic genitives* (pp. 275–299). John Benjamins.

Schlücker, B. (2019). Compounds and multi-word expressions in German. In B. Schlücker (Ed.), *Complex lexical units: Compounds and multi-word expressions* (pp. 69–94). De Gruyter.

Schreiber, J. (2021). *Free German Dictionary.* Online. Retrieved from `https://sourceforge.net/projects/germandict/` (Available from SourceForge)

Wegener, H. (2008). The regrammaticalization of linking elements in German. In E. Seoane & M. J. LópezCouso (Eds.), *Theoretical and empirical issues in grammaticalization* (pp. 333–355). Amsterdam: Benjamins.

Weller, M., Cap, F., Müller, S., im Walde, S. S., & Fraser, A. (2014). Distinguishing degrees of compositionality in compound splitting for statistical machine translation. In *Proceedings of the first workshop on computational approaches to compound analysis (comacoma 2014)* (pp. 81–90).